Pragmatics in Language Grounding:
Phenomena, Tasks, and Modeling Approaches

Daniel Fried[1*] **Nicholas Tomlin**[2*]   **Jennifer Hu**[3]
Roma Patel[4,5]   **Aida Nematzadeh**[5]
Carnegie Mellon[1]   UC Berkeley[2]   MIT[3]   Brown University[4]   DeepMind[5]
dfried@cs.cmu.edu   nicholas_tomlin@berkeley.edu
jennhu@mit.edu   {romapatel,nematzadeh}@deepmind.com

Abstract

People rely heavily on context to enrich meaning beyond what is literally said, enabling concise but effective communication. To interact successfully and naturally with people, user-facing artificial intelligence systems will require similar skills in *pragmatics*: relying on various types of context—from shared linguistic goals and conventions, to the visual and embodied world—to use language effectively.

We survey existing grounded settings and pragmatic modeling approaches and analyze how the task goals, environmental contexts, and communicative affordances in each work enrich linguistic meaning. We present recommendations for future grounded task design to naturally elicit pragmatic phenomena, and suggest directions that focus on a broader range of communicative contexts and affordances.

1 Using Language in Context

People often use language to achieve communicative goals (Wittgenstein, 1953; Austin, 1975; Clark, 1996; Frank and Goodman, 2012), producing effects on other people and the world. To achieve their goals efficiently, people rely on various types of context to allow their conversational partners to enrich meaning beyond what the speaker has literally said, making language highly context-dependent; a simple sentence such as "it's nice out today" can invoke different meanings depending on the situation—it can be an implicit invitation, a statement contrasting the weather with a previous day, or even ironic if the weather is poor. This broad ability to use language in context to achieve goals is known as *pragmatics*.

User-facing artificial intelligence systems require similar pragmatic skills to interact successfully and efficiently with people. Recent work has focused on general-purpose models (Tan and Bansal, 2019; Brown et al., 2020; Radford et al.,

2021; Bommasani et al., 2021, *inter alia*) that have achieved remarkable performance on a variety of task benchmarks intended to measure literal, semantic meaning. We believe the time is right to move beyond these popular benchmark tasks, to tests that require communicating with people in rich contexts: ones that elicit pragmatic phenomena from people, and benefit from pragmatic abilities in systems. We focus on work that incorporates multimodal context in pragmatic reasoning, motivated by the rich contexts of typical language use—our understanding of language is shaped by the environments that we use it in (Baldwin, 1995; Bloom, 2002; Tomasello, 2005; Vigliocco et al., 2014; Bender and Koller, 2020; Bisk et al., 2020).

We first survey how various communicative and environmental context types elicit pragmatic phenomena. Using these context types and phenomena, we then survey representative tasks and datasets which have been used both to study pragmatic communication in people, and to build goal-oriented multimodal systems. We present tasks along a spectrum of complexity, ranging from constrained reference games to goal-oriented embodied dialogue. We discuss how choices in task design—including environment properties, context types, and communicative affordances—shape the pragmatic phenomena that arise in tasks, and provide suggestions for future task and dataset designers. To model these tasks and phenomena, we give an overview of a range of computational pragmatic approaches that view communication as goal-directed actions by agents in context.

Finally, we suggest further integrations of computational pragmatics with NLP: using pragmatic modeling to allow systems to communicate and interact more successfully and efficiently with people; focusing on developing tasks with rich contexts or contextualizing existing tasks; and tackling challenges of human evaluation and data sparsity.

---

*Equal contribution.

2 Pragmatic Phenomena

In linguistics and cognitive science, *pragmatics* is often defined in contrast to *semantics*. Broadly speaking, semantics characterises the literal meanings of linguistic expressions, whereas pragmatics captures the context-dependent components of meaning, which may contain the bulk of actual communication (Clark, 1997; Casasanto and Lupyan, 2015). Pragmatic communication draws upon many different sources of information, ranging from environmental factors to inferences about other agents' unspoken information and goals. This makes pragmatics both a critical and challenging component for designing NLP systems that interact with people. In this section, we discuss the types of context in which language can be situated and the non-literal inferences that arise as a result of these contextual pressures.

2.1 Types of Context

Many aspects of pragmatics involve the maintenance of *common ground*, a set of contextual information shared between communicative partners (e.g., Lewis, 1969; Clark and Brennan, 1991; Traum, 1994; Stalnaker, 2002; Clark, 2015). Key elements of common ground include (1) social and communicative norms, (2) task goals and collaborative actions, (3) common knowledge, and (4) discourse context. In addition to common ground, we focus on pragmatic reasoning that also requires multimodal context, such as (5) visual information or (6) embodied interaction. See Appendix A for definitions and examples for each of these types of context; we also point readers to Levinson (1983) and Birner (2012) for more discussion.

2.2 Roles of Pragmatics

We adopt a use-oriented view of pragmatics (Clark, 1996), highlighting various ways that pragmatic reasoning may be used in grounded language tasks. Our taxonomy is not intended to be fully exhaustive, and we caveat that some categories may partially overlap with one another. For a focus on classic linguistic pragmatic phenomena, like *deixis* and *presupposition*, we refer readers to existing surveys such as Levinson (1983) and Birner (2012).

Reasoning About Alternatives. Much of linguistic meaning comes not just from what we say, but from what we do not say. The utterances that speakers choose *not* to say i.e., the set of alternative utterances which are likely in a context, can reveal

their intended meanings and mental states (Horn, 1984; Fox and Katzir, 2011; Degen, 2013; Buccola et al., 2022), e.g., *some of the apples are red* likely conveys that some are not, since the speaker did not say all were red. Many of the following roles of pragmatics also often involve reasoning over alternatives.

Understanding Ambiguity. Language is frequently ambiguous for many reasons (Piantadosi et al., 2012): ambiguity may be used strategically to achieve communicative efficiency or to remove information that is unhelpful to the task at hand. Moreover, ambiguous instructions often require listeners to reason pragmatically about alternative intentions that speakers might have. For example, when asked to *pass the knife* in a cooking scenario, a pragmatic agent might have to reason about the context to determine whether to provide a butter, bread, or steak knife. By relying on contextual information to resolve ambiguities in situations such as these, pragmatic interlocutors can communicate more efficiently (Solé and Seoane, 2015; Fortuny and Corominas-Murtra, 2013).

Collaborative Planning. Many grounded dialogue tasks require agents to coordinate to carry out joint activities, e.g., collaboratively agree to a goal before executing it. To succeed at tasks like these, participants often must reason about each other's possible goals, for example in a collaborative building setting, inferring that *four planks* can be either a command or a description depending on what effect the speaker is trying to produce on the listener. In environments with many world states, there are a combinatorial number of goals to reason about and actions to take, but a participant can usually only communicate with their partner for a limited time. Therefore, participants must trade-off between communicating efficiently and acting.

Convention Formation and Abstraction. Conventions, as characterised by Lewis (1969), are arbitrary but stable solutions to recurring coordination problems that typically form out of the maxims of rational communication (Grice, 1975). For example, a team of workers who communicate with one another daily might initially have lengthy descriptions to refer to certain items, but after a while, might start to develop a common ground of simpler words to refer to them. These abstractions or conventions are hypothesised to emerge as a result of repeated interactions (Garrod and Doherty, 1994).

One theory is that conventions form to help resolve ambiguity, yielding more efficient communication at the levels of individuals (Hawkins et al., 2017) or populations (Hawkins et al., 2022).

Efficiency and Mutual Exclusivity. For many grounded tasks where the goal is to learn a correspondence between meanings and utterances, pragmatic reasoning can be used to avoid learning degenerate mappings. For example, on learning that a certain label (e.g., *cat*) refers to an object, an agent might use mutual exclusivity to rule out the possibility of another label (e.g., *dog*) also referring to the object (Markman and Wachtel, 1988; Clark, 1988). Models of pragmatic reasoning often induce biases toward mutual exclusivity that can lead to more efficient learning (Wang et al., 2016; McDowell and Goodman, 2019). More broadly, pragmatic reasoning may be used to manage the dual pressures of informativity and conciseness (Zipf, 1949; Horn, 1984; Blutner, 1998), which are explicitly factored into pragmatic models such as RSA (cf. Section 4.2). As a result, pragmatics may lead to communicative efficiency both during language learning and language use.

3 Existing Tasks and Environments

In this section, we critically evaluate several well-studied grounded language tasks. We focus on tasks in multimodal domains that make use of natural language data.[1]

3.1 Types of Tasks

Grounded, task-oriented dialogue provides a general setting to study pragmatics. Dialogue tasks provide rich and varied contexts (e.g., different types of common ground, goals, and environments) as well as communicative affordances (e.g., the ability to ask questions, provide information in installments, and adapt to a partner's conventions). These contexts and affordances interact to produce a diverse range of pragmatic behavior (Clark, 1996). However, many of these contexts, affordances, and behaviors are also present in more restricted and controlled tasks for which data collection, analysis, modeling, and evaluation are often more tractable. For example, image captioning tasks simplify data

collection and modeling by limiting the number of conversational turns to one; instruction interpretation tasks additionally simplify evaluation (so long as it is possible to carry out and validate actions in the world).

We focus on reference games, image captioning, instruction following, and grounded dialogue tasks that give us a broad characterization of the different properties that tasks might have, as summarized in Table 1.[2] For each task, we specify what type of context is needed, how pragmatic behavior is typically exhibited, and several important elements of task design: partial observability, symmetry, and iterated interaction (see Section 3.2). We present these domains in order of increasing complexity, finding that the most complex grounded dialogue tasks are more likely to involve features like partial observability or symmetry which induce additional pragmatic phenomena.

Reference Games. Reference games typically involve two players, a listener and a speaker agent. Both players are presented with a shared set of referents, e.g., images, objects, or abstract illustrations, and the speaker is tasked with describing a target referent to the listener, who must then guess the target (Clark and Wilkes-Gibbs, 1986; Gorniak and Roy, 2004; Steels and Belpaeme, 2005; Golland et al., 2010; Frank and Goodman, 2012; Kennington and Schlangen, 2015). An example reference game is the Colors in Context (Monroe et al., 2017) task, in which players are presented with three color swatches and asked to describe one of them. Even simple phrases like *plain blue* may have different meanings depending on visual context in this task.

Image Captioning. A broad class of image captioning tasks require producing text to describe an image (Barnard et al., 2003; Farhadi et al., 2010; Mitchell et al., 2012; Kulkarni et al., 2013). Most captioning work has only been implicitly goal-oriented: corpora have been constructed by asking annotators to determine and describe the important parts of an image (Hodosh et al., 2013; Young et al., 2014; Chen et al., 2015). Systems are evaluated on how closely their descriptions match these human-

---

[1]We omit large bodies of work on unimodal pragmatics (Degen, 2013; Jeretic et al., 2020; Choi et al., 2021, *inter alia*) or language that might be grounded, but is synthetically generated (Johnson et al., 2017; Bastings et al., 2018; Zhong et al., 2020).

[2]This is not an exhaustive taxonomy of grounded language learning tasks. For example, VQA (Antol et al., 2015), NLVR2 (Suhr et al., 2019b), the Hateful Memes Challenge (Kirk et al., 2021), and Winoground (Thrush et al., 2022) do not fit perfectly into any of the above categories, although most bear some similarities to image captioning.

Task (Dataset)	Types of Context	Role of Pragmatics	Par.Ob.	Sym.	Iter.
Reference Game *Colors in Context* Monroe et al. (2017)	Visual	Reasoning about alternatives, understanding ambiguity	✗	✗	✓
Image Captioning *Abstract Scenes* Andreas and Klein (2016)	Visual, common knowledge	Reasoning about alternatives	✗	✗	✗
Conceptual Captions Alikhani et al. (2020)	Visual, common knowledge, joint goals, norms of interaction	Efficiency considerations	✗	✗	✗
Instruction Following *SHRDLURN* Wang et al. (2016)	Visual	Mutual exclusivity, convention formation, efficiency considerations	✗	✗	✓
CerealBar Suhr et al. (2019a)	Visual, embodied	Collaborative planning, understanding abstractions and conventions	✓	✗	✓
Hexagons Lachmy et al. (2022)	Visual, norms of interaction	Understanding abstractions and ambiguity, efficiency considerations	✗	✗	✓
Grounded Dialogue *Cards Corpus* Potts (2012)	Visual, embodied, joint goals, norms of interaction, discourse	Collaborative planning, understanding ambiguity, efficiency	✓	✓	✗
OneCommon Udagawa and Aizawa (2019)	Visual, joint goals, norms of interaction, discourse	Collaborative planning, understanding ambiguity, efficiency	✓	✓	✗
PhotoBook Haber et al. (2019)	Visual, common knowledge, joint goals, norms of interaction, discourse	Convention formation, understanding ambiguity, efficiency	✓	✓	✓

Table 1: Example grounded language learning datasets that involve pragmatic reasoning, organized by task type. The task attributes refer to: partially observable, symmetric, and iterated (multi-turn) interactions. We observe that grounded dialogue and instruction following tasks often involve a broader range of pragmatic reasoning behaviors.

written references, which poses challenges given considerable variation in what annotators chose to describe and how they wrote the descriptions (Anderson et al., 2016).

Other work, particularly in the computational pragmatics literature, has formulated captioning as a contrastive task (Andreas and Klein, 2016; Vedantam et al., 2017; Cohn-Gordon et al., 2018), where a target image must be described to contrast it from other similar, *distractor* images. This setting can be viewed as a scaled-up reference game involving complex visual inputs, and many such pragmatically-motivated variations on standard image captioning have appeared in recent years: Nie et al. (2020) define *issue-sensitive image captioning*, in which models implicitly caption several target images at a time, while Alikhani et al. (2020) train *coherence-aware* captioning models which may vary in the degree of subjectivity or the extent to which inferences about target images are made.

Of the task categories we discuss, image captioning has the most immediate real-world applicability, especially for accessibility e.g., to provide descriptions that could substitute for images for visually-impaired users on the web (Pont-Tuset et al., 2020). Additionally, practical considerations in this domain often require pragmatic reasoning

e.g., specifically describing salient characteristics of an image (e.g., *a man* versus *Barack Obama*), being concise, or describing the relevance of the image to document context. We refer the reader to MacLeod et al. (2017) and Kreiss et al. (2021) for further information on this topic.

Instruction Following. Instruction following tasks require a listener to take instructions from a speaker, predicting *trajectories* in an environment (Branavan et al., 2009; Vogel and Jurafsky, 2010; Chen and Mooney, 2011; Tellex et al., 2011; Anderson et al., 2018). Trajectories can be grammar-based actions (e.g., ADD(LEFTMOST(WITH(BROWN)), ORANGE), to specify *add an orange block to the left-most brown block* in the block-stacking setting of Wang et al. 2016), sequences of discrete movements (e.g., between nodes in a navigation graph in Chen et al. 2019; Ku et al. 2020), or continuous sequences (e.g., of orientations in Ku et al. 2020).

A speaker must describe a target trajectory in a way that allows the listener to correctly carry it out in the presence of (often exponentially many) alternative trajectories (e.g., *left* versus *sharp left*). These environments often involve visually-grounded observations (Anderson et al., 2018; Chen et al., 2019; Ku et al., 2020), ac-

tion hierarchies (Shridhar et al., 2020) or programmatic abstractions (Lachmy et al., 2022) and some parts of the environment may be unobserved to the speaker, the listener, or both (see Section 3.2), causing language to be more ambiguous and context-dependent.

Grounded Goal-Oriented Dialogue. We focus on grounded dialogue tasks that involve two-way communication between partners to achieve a shared goal (e.g., Chai et al., 2004; Rieser and Lemon, 2008; Das et al., 2017; De Vries et al., 2017; Kim et al., 2019; Narayan-Chen et al., 2019; Ilinykh et al., 2019).[3] These tasks generalize the one-way communication settings above; however, two-way communication provides additional affordances—allowing players to ask clarification questions, acknowledge understanding, and coordinate actions. For example, in the Cards task (Potts, 2012), players collaboratively collect a set of cards in a grid world environment by communicating with other players while moving around to pick up cards. Observability is limited to parts of the environment close to the players, requiring them to pool information, and they must collaboratively plan to agree on one of the multiple possible sets of cards they can collect.

The multi-turn nature of dialogue also necessitates reasoning about past actions and interactions (perform *inference*) and likely outcomes in the future (*planning*). These are particularly evidenced in collaborative reference tasks such as OneCommon (Udagawa and Aizawa, 2019), where players must infer which items they share with their partners, aggregating information over the course of a dialogue. Finally, repeated interactions in dialogue can allow linguistic adaptation. For example, in PhotoBook task (Haber et al., 2019)—a collaborative reference task where players have repeated conversations about photographs—players adapt their language over time to match each other, becoming more efficient over time (e.g., reducing *the strange bike with three wheels* to *strange bike*).

3.2 Elements of Task Design

We now outline three especially pragmatically-relevant dimensions to consider when designing tasks and describe how they induce various types of pragmatic phenomena.

Observability. In *partially observable* tasks, participants can only see a limited portion of the environment, for example seeing only the parts of the grid closest to them in the Cards task (Potts, 2012). This can make language more context-dependent, in particular creating a dependence on *when* or *where* the language was produced. The most complex partially-observable settings, including all of the collaborative dialogue tasks in Table 1, involve participants observing different views of the environment — requiring them to collaboratively plan to pool their information. Different views can also lead to false agreements where participants believe they have coordinated but actually disagree (Chai et al., 2014; Udagawa and Aizawa, 2019), requiring more explicit pragmatic modeling of the partner's perspective to avoid and resolve ambiguity.

Symmetry. Tasks differ in the types of roles performed by the communicating agents, which in turn shapes the type of language produced and actions taken. We distinguish between *asymmetric* and *symmetric* roles. In an asymmetric setting — e.g., speaker and listener, or teacher and follower — pragmatics may be helpful for production and comprehension of language utterances. Symmetric settings (Vogel et al., 2013a,b) may be more naturalistic and are often used in coordination tasks, although designing such settings is often more complicated. Asymmetric settings (Monroe et al., 2017; Andreas and Klein, 2016) are often the simplest way to introduce pragmatic phenomena, since asymmetry occurs when one agent is missing information.

Interaction. The nature of interaction(s) between communicating agents affects the language that is produced. In a *one-turn* interaction, all usable information must be expressed in a single utterance, forcing speakers to balance informativity and conciseness. In *iterated one-sided* interactions, the speaker has the opportunity to respond to the listener's actions before planning each new utterance. Finally, in *dialogue*, agents can freely coordinate and participate in speech acts—they can jointly build common ground, ask clarification questions, and share useful information. These repeated interactions between agents require attention to conversation history, and may give rise to the formation of conventions (e.g., Hawkins et al., 2017).

---

[3] Our focus is on task-oriented dialogue, given that communicative goals are less explicit in chit-chat settings (but see Kim et al. (2020) for a recent pragmatic treatment).

3.3 Evaluating Pragmatic Models

The ultimate goal for user-facing, situated agents is to communicate (1) successfully and (2) efficiently with people. Human evaluations, where agents are paired with people at test-time, are an ideal way to measure this (Walker et al., 1997; Koller et al., 2010; Parent and Eskenazi, 2010; Suhr et al., 2019a, *inter alia*), but are not always feasible to carry out since they complicate controlling and replicating experimental setups. Thus, evaluation often resorts either to static, human-produced corpora or automated model-based evaluations.

Task success. Interpretation tasks are typically amenable to corpora-based evaluation. For example, listener agents in reference games can be easily evaluated based on the accuracy of referent selection. In contrast, evaluating language generation tasks for speaker agents is more challenging, given that many classical reference-based automated NLG metrics are unable to measure whether or not generated language will be understood correctly by human listeners (Krahmer and Theune, 2010; Fried et al., 2018a; Zhao et al., 2021; Gehrmann et al., 2022). Automated proxies for human listeners are models of how people interpret and respond to a system's language, known as *user simulation* or *self-play* (Georgila et al., 2006; Rieser and Lemon, 2011; Lewis et al., 2017; Kim et al., 2019) in dialogue and *communication-based evaluation* (Newman et al., 2020) in reference games, where speaker generations are fed to a listener model and evaluated on task success. Automated models can only give rough indicators of how humans might interpret the system's language. For this reason, we stress the importance of making the evaluation model dissimilar from the system and using human evaluations whenever possible.

Communicative efficiency. Beyond task success, a secondary criterion for situated agents is efficient communication. For example, if the language generated by a speaker, although correct, is difficult to understand, this calls for unnecessary interpretation effort from the other agent. To measure whether pragmatic agents enable efficient communication, evaluations can use metrics of communicative cost (Walker et al., 1997) such as time to task completion, utterance length and complexity (Effenberger et al., 2021), measures such as lexical entrainment (Clark and Wilkes-Gibbs, 1986; Parent and Eskenazi, 2010; Hawkins et al., 2020), and

quality ratings (Kojima et al., 2021).

4 Modeling Pragmatics

In this section, we discuss frameworks that have been proposed to characterize *how* listeners can derive pragmatic meaning, providing a starting point for modeling the phenomena and tasks above.

4.1 Gricean Maxims

In his seminal proposal, Grice (1975) argues that speakers and listeners are guided by an underlying *cooperative principle*: taking action to jointly achieve communicative goals, and assuming that other agents are acting similarly. Grice divides this principle up into a set of maxims. However, attempts to directly implement the Gricean maxims computationally (e.g., Hirschberg, 1985a) have had to grapple with substantial underspecification and overlap in Grice's proposal. Later *neo-Gricean* work in linguistics has streamlined the maxims considerably (Horn, 1984; Levinson, 2000) and characterizes many pragmatic effects in terms of the trade-off between speaker and listener effort in achieving cooperative goals. These approaches have had few direct computational implementations; however, a line of computational work, which we outline in Sections 4.2 and 4.3, *derives* maxim-like behavior through multi-agent modeling rather than by prescriptively implementing the maxims.

4.2 Multi-Agent Reasoning

A number of computational frameworks view utterance generation and interpretation using a multi-agent or game-theoretic lens (Rosenberg and Cohen, 1964; Cohen and Levesque, 1990; Golland et al., 2010; Jäger, 2012; Franke, 2013). Many of these frameworks derive pragmatic behavior by modeling communication partners as rational agents who jointly try to optimize a communicative utility function using theory-of-mind reasoning about other agents. We focus on one representative of these, the Rational Speech Acts (RSA) framework (Frank and Goodman, 2012; Goodman and Frank, 2016), as it has been successfully applied across a range of grounded language settings.

RSA defines a recursive reasoning process where speakers and listeners model each other's goals and interpretations. A *rational speaker* chooses utterances using an embedded model of how the listener will likely interpret utterances A *rational listener*, in turn, reasons counterfactually about a

rational speaker generating language in this way—reasoning about why the speaker choose an observed utterance rather than alternatives—which can resolve ambiguity in the speaker's utterances.

Theoretical work on the RSA framework has used it to model interpretation and generation of a wide range of pragmatic phenomena (Goodman and Frank, 2016; Bergen et al., 2016; Kao et al., 2014a,b). In practice, most RSA implementations involve speaker agents which sample candidate generations from a conditional language model, and listener agents which execute these generations in the environment (Andreas and Klein, 2016). A variety of work has also applied RSA to improve performance of NLP systems on a range of tasks involving complex natural language utterances, including reference games (Monroe et al., 2017), instruction following and generation (Fried et al., 2018a,b), image captioning (Andreas and Klein, 2016; Cohn-Gordon et al., 2018), summarization (Shen et al., 2019), MT (Cohn-Gordon and Goodman, 2019), and dialogue (Kim et al., 2020; Fried et al., 2021). A number of rational communication frameworks also include noteworthy variations on the core RSA setup, include varying the utility function (Zaslavsky et al., 2021), modeling mis-aligned objectives (Asher and Lascarides, 2013), using deeper levels of recursive reasoning between agents (Wang et al., 2020), and non-linguistic communication (Hadfield-Menell et al., 2017; Jeon et al., 2020; Pu et al., 2020).

One key limitation of RSA is that it models speakers as choosing their utterances from a known and fixed set of candidate utterances. Because using the full space of possible natural language utterances is intractable for most tasks, some recent work has proposed incremental models which process one word or character at a time, leading to promising results on constrastive image captioning (Cohn-Gordon et al., 2018; Nie et al., 2020). A second notable limitation is that, with a few exceptions (e.g., Khani et al., 2018), applications of full recursive reasoning frameworks have been limited to single-turn interactions. However, the multi-turn approaches that we outline in Section 4.3 allow modeling repeated interactions by making the framework simpler along certain axes (e.g., removing higher-order theory-of-mind).

4.3 Multi-Turn Approaches

A variety of approaches to multi-turn pragmatics have arisen in work on task-oriented dialogue. Many of these treat communication as goal-directed decision-making under uncertainty (Rieser and Lemon, 2011; Young et al., 2013), and can be broadly viewed as generalizing the single-turn frameworks of Section 4.2. For generation, a variety of dialogue systems explicitly *plan* utterances or speech acts to convey information to their partners (Cohen and Perrault, 1979; Traum, 1994; Walker et al., 2004; Rieser and Lemon, 2009; Kim et al., 2020, *inter alia*). For interpretation, many systems *infer* the latent intent or state of the user (Allen and Perrault, 1980; Paek and Horvitz, 2000; Williams and Young, 2007; Schlangen et al., 2009; Young et al., 2013, *inter alia*).

Planning and inference are classic AI tasks with broad applicability, and most of the works above are closely related to general machinery developed for *decentralized POMDPs* (Bernstein et al., 2002; Oliehoek and Amato, 2016). However, given computational challenges, past work on algorithmic applications of POMDP algorithms to communication have focused on domain-specific formalisms (the works above) or restricted language settings (Zettlemoyer et al., 2008; Vogel et al., 2013a; Hadfield-Menell et al., 2016; Foerster et al., 2019; Jaques et al., 2019). To enable pragmatic modeling and interaction with people in naturalistic grounded dialogue settings, future work might draw on further progress that the multi-agent reinforcement learning and planning communities make on these underlying algorithmic challenges.

5 Discussion

5.1 What Can Pragmatics Give Us?

We predict that developing pragmatic capabilities will advance the frontier of NLP, and vice versa. Many of the shortcomings of current large-scale neural models are pragmatic limitations, such as processing implicatures (Jeretic et al., 2020; Ruis et al., 2022), predicting mental states (Sap et al., 2022), and building common ground with people (Benotti and Blackburn, 2021). More generally, many of the roles of pragmatics that we outline in Section 2.2 are linguistic manifestations of classical problems in machine learning and AI: inference and model calibration underlie reasoning about alternatives and ambiguity; multi-agent search underlies collaborative planning; adaptation and abstraction learning underlie convention formation. We expect pragmatics to be a challenging and productive driver of progress on these problems and on

developing more useful NLP models and agents.

Practical benefits of pragmatic inference include more efficient communication and learning. Reasoning about how context enriches meaning (Section 2.1) allows people to say less while still conveying their intentions, and to learn from what could have been said but was not. A long line of work in computational pragmatics explicitly examines this tradeoff between conciseness and communicative success (Section 4.1); we encourage future work in NLP to similarly evaluate models on communicative efficiency as well as task success (Section 3.3) and consider approaches that aim to achieve both desiderata (e.g., Section 4.2).

Pragmatics becomes most essential — and most challenging — in grounded and interactive settings, which have the richest contexts of language use. Pragmatic phenomena that have been understudied in NLP, such as convention formation (Section 2.2), emerge when people can rely on shared dialogue, environmental, and task contexts to convey and decode meanings. While pretraining on multimodal (Lu et al., 2019; Sun et al., 2019; Tan and Bansal, 2019; Radford et al., 2021) and interactive (Stiennon et al., 2020; Shuster et al., 2022; Bai et al., 2022) data has driven recent progress in their respective domains, training data and evaluation setups for settings that are both grounded and interactive remain limited in scope, size, and ecological validity (De Vries et al., 2020). This sparsity poses both a challenge and an opportunity for pragmatic language use.

5.2 Building Pragmatically-Informed Tasks

We encourage future work to focus on complex interactive scenarios which contain a wide range of pragmatic phenomena. As shown in Table 1, grounded, multi-turn dialogue tasks with partial observability and symmetry often encompass the widest variety of pragmatic language behavior. To date, relatively few tasks have all of these properties, although there are a few notable exceptions (e.g., Potts, 2012; Udagawa and Aizawa, 2019). We argue that such tasks provide useful testbeds for building models of convention formation and collaborative planning, which are currently understudied in the computational pragmatics literature.

In tandem, we also encourage future work to *contextualize* existing NLP tasks in order to bring them closer to real-world applicability. For example, as discussed in Section 3.1, MacLeod et al.

(2017) and Kreiss et al. (2021) argue that current image captioning systems fail to meet the needs of visually-impaired users because they do not provide descriptions in the context of the article or of the user's goals. Incorporating and modeling additional context—such as user intent in captioning (Alikhani et al., 2020), interaction history in instruction following (Kojima et al., 2021; Lin et al., 2022), or emotion (Kim et al., 2021) and personality (Wang et al., 2019) in dialogue—may help close the gap between current NLP benchmarks and useful real-world systems.

5.3 Challenges and Open Questions

Going forward, accurate evaluation of pragmatics is one key challenge in grounded language learning. Some recent benchmarks have been proposed to evaluate pragmatic language behavior in large language models (Sap et al., 2022; Hu et al., 2022; Ruis et al., 2022), but evaluation in interactive, multimodal scenarios often requires humans or human proxies. As discussed in Section 3.3, some models may be evaluated in self-play, and future work might extend evaluation practices from multi-agent reinforcement learning, where methods like population play, imitation learning from human data, and fictitious co-play (Strouse et al., 2021) provide proxies for human interaction. However, human evaluation remains the gold standard in most interactive settings, and we encourage future work to improve human evaluations and focus on tasks that are intrinsically-motivating (e.g., useful or fun) for people (e.g., Wang et al. 2017; FAIR et al. 2022).

Another key challenge is handling data sparsity. Although recent advances in large-scale pretraining have led to few-shot capabilities on many language tasks, interactive and grounded tasks pose additional challenges in sparsity. First, the addition of multimodal context widens the conditioning space. Second, many grounded settings involve domain-specific knowledge; for example, instruction following and grounded dialogue settings often have unique action spaces, as well as domain-specific abstractions or conventions in language use. Finally, many settings require adaptation to individual users, for which there will always be limited data.

As NLP expands to an ever-wider range of contexts, we encourage work to include pragmatics as a central component, with the goal of communicating successfully, efficiently, and naturally with people in challenging and useful settings.

Limitations

Although we aim to describe a representative sample of tasks in Table 1, our coverage is necessarily incomplete, especially in domains such as image captioning, instruction-following, and collaborative dialogue, so we refer readers to other surveys on these issues (e.g., Luketina et al., 2019). As noted in Section 3, we focus exclusively on task-oriented grounded domains involving natural language data. Our survey therefore includes limited discussion of pragmatic phenomena in unimodal text domains such as chitchat dialogue, purely textual task-oriented dialogue, and language classification tasks (although c.f. Section 4.3 and Appendix B), and omits much work on analyzing the abilities of models to perform classic pragmatic tasks such as implicature and presupposition (e.g., Ross and Pavlick, 2019; Jeretic et al., 2020). We also do not discuss tasks involving synthetic or emergent language, but see Lazaridou and Baroni (2020) for a survey of the latter.

Our discussion of modeling frameworks for pragmatics in Section 4 focuses on approaches that distinguish between semantics and pragmatics through social reasoning about other agents' beliefs and goals. Due to space limitations, we did not discuss alternate theories proposing that pragmatically enriched meanings are derived within the grammar of a language, without recourse to probabilistic social reasoning (e.g., Fox, 2007; Chierchia et al., 2012; Asherov et al., 2021). These theories remain difficult to implement at scale, but we encourage future work to explore them as candidate hypotheses alongside the frameworks discussed in Section 4.

There is also rich body of work on formalizing and modeling discourse context beyond the approaches we cover here, including conversational analysis (Schegloff, 1968; Sacks et al., 1974) and discourse coherence and structure (Hobbs, 1979; Grosz and Sidner, 1986; Webber, 1991; Kamp and Reyle, 1993; Grosz et al., 1995; Webber et al., 2003; Asher and Lascarides, 2003; Barzilay and Lapata, 2008). We refer to Cohen et al. (1990), Clark (1996), Jurafsky and Martin (2014), and Alikhani and Stone (2020) for entry points.

References

Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535, Online. Association for Computational Linguistics.

Malihe Alikhani and Matthew Stone. 2020. Achieving common ground in multi-modal dialogue. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 10–15, Online. Association for Computational Linguistics.

James F Allen and C Raymond Perrault. 1980. Analyzing intention in utterances. *Artificial Intelligence*, 15(3):143–178.

Prithviraj Ammanabrolu and Mark O. Riedl. 2021. Situated language learning via interactive narratives. *Patterns*, 2(9).

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 382–398. Springer International Publishing.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683.

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Nicholas Asher and Alex Lascarides. 2013. Strategic conversation. *Semantics and Pragmatics*, 6.

Daniel Asherov, Danny Fox, and Roni Katzir. 2021. On the Irrelevance of contextually given states for the computation of Scalar Implicatures.

John Langshaw Austin. 1975. *How to Do Things with Words*. Clarendon Press.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Dare A Baldwin. 1995. Understanding the link between joint attention and language. *Joint attention: Its origins and role in development*, 131:158.

Kobus Barnard, David Forsyth, David M Blei, Michael I Jordan, Jaz Kandola, Thomas Hofmann, Tomaso Poggio, and John Shawe-Taylor. 2003. Matching words and pictures. *Journal of Machine Learning Research*.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Jasmijn Bastings, Marco Baroni, Jason Weston, Kyunghyun Cho, and Douwe Kiela. 2018. Jump to better conclusions: SCAN both left and right. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 47–55, Brussels, Belgium. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Luciana Benotti and Patrick Blackburn. 2021. 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.

Leon Bergen, Roger Levy, and Noah Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 9.

Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. 2002. The complexity of decentralized control of Markov decision processes. *Mathematics of operations research*, 27(4):819–840.

Betty J Birner. 2012. *Introduction to Pragmatics*, volume 38. John Wiley & Sons.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Paul Bloom. 2002. *How children learn the meanings of words*. MIT press.

Reinhard Blutner. 1998. Lexical Pragmatics. *Journal of Semantics*, 15(2):115–162.

Dan Bohus and Eric Horvitz. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, number Article 5 in ICMI-MLMI '10, pages 1–8, New York, NY, USA. Association for Computing Machinery.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

S.R.K. Branavan, Harr Chen, Luke Zettlemoyer, and Regina Barzilay. 2009. Reinforcement learning for mapping instructions to actions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 82–90, Suntec, Singapore. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are Few-Shot learners. In *Neural Information Processing Systems*.

Brian Buccola, Manuel Križ, and Emmanuel Chemla. 2022. Conceptual alternatives. *Linguistics and Philosophy*, 45(2):265–291.

Daniel Casasanto and Gary Lupyan. 2015. All concepts are ad hoc concepts. In E Margolis and S Laurence, editors, *The Conceptual Mind: New directions in the study of concepts*. MIT Press.

Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, SIGGRAPH '94, pages 413–420, New York, NY, USA. Association for Computing Machinery.

Joyce Chai, Maya Cakmak, and Candace Sidner. 2019. Teaching robots new tasks through natural interaction. In *Interactive Task Learning*. The MIT Press.

Joyce Y Chai, Pengyu Hong, and Michelle X Zhou. 2004. A probabilistic approach to reference resolution in multimodal user interfaces. In *Proceedings of the 9th International Conference on Intelligent User Interfaces*, New York, New York, USA. ACM Press.

Joyce Y Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. 2014. Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, HRI '14, pages 33–40,

New York, NY, USA. Association for Computing Machinery.

David L. Chen and Raymond J. Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011*. AAAI Press.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *CVPR*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server.

Gennaro Chierchia, Danny Fox, and Benjamin Spector. 2012. The Grammatical View of Scalar Implicatures and the Relationship between Semantics and Pragmatics. In *Semantics: An International Handbook of Natural Language Meaning*, volume 3, pages 2297–2332. Mouton de Gruyter.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Eve V Clark. 1988. On the logic of contrast. *J. Child Lang.*, 15(2):317–335.

Eve V. Clark. 2015. Common Ground. In *The Handbook of Language Emergence*, pages 328–353. John Wiley & Sons, Ltd.

Eve V. Clark and Josie Bernicot. 2008. Repetition as ratification: How parents and children place information in common ground. *Journal of Child Language*, 35(2):349–371.

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

Herbert H Clark. 1997. Dogmas of understanding. *Discourse Processes*, 23(3):567–598.

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

Herbert H Clark and Meredyth A Krych. 2004. Speaking while monitoring addressees for understanding. *J. Mem. Lang.*, 50(1):62–81.

Herbert H Clark and Catherine R Marshall. 1981. Definite knowledge and mutual knowledge. In A. K. Joshi, B. Webber, & I. Sag, editor, *Elements of discourse understanding*, pages 10–63. Cambridge University Press.

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

Philip R Cohen and Hector J Levesque. 1990. Rational interaction as the basis for communication. In Philip R Cohen, Jerry Morgan, and Martha E Pollack, editors, *Intentions in Communication*.

Philip R Cohen, Jerry L Morgan, Morgan Jerry, and Martha E Pollack. 1990. *Intentions in Communication*. MIT Press.

Philip R. Cohen and C. Raymond Perrault. 1979. Elements of a plan-based theory of speech acts. *Cognitive Science*.

Reuben Cohn-Gordon and Noah Goodman. 2019. Lost in machine translation: A method to reduce meaning loss.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443, New Orleans, Louisiana. Association for Computational Linguistics.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Harm De Vries, Dzmitry Bahdanau, and Christopher Manning. 2020. Towards ecologically valid research on language user interfaces. *arXiv preprint arXiv:2007.14435*.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. GuessWhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5503–5512.

Judith Degen. 2013. *Alternatives in pragmatic reasoning*. University of Rochester.

Anca D Dragan, Kenton C T Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE.

Anna Effenberger, Rhia Singh, Eva Yan, Alane Suhr, and Yoav Artzi. 2021. Analysis of language change in collaborative instruction following. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2803–2811, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Meta Fundamental AI Research Diplomacy Team FAIR, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 15–29. Springer Berlin Heidelberg.

Jakob Foerster, Francis Song, Edward Hughes, Neil Burch, Iain Dunning, Shimon Whiteson, Matthew Botvinick, and Michael Bowling. 2019. Bayesian action decoder for deep multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1942–1951.

Jordi Fortuny and Bernat Corominas-Murtra. 2013. On the Origin of Ambiguity in Efficient Communication. *Journal of Logic, Language, and Information*, 22(3):249–267. Publisher: Springer.

Danny Fox. 2007. Free Choice Disjunction and the Theory of Scalar Implicatures. In Uli Sauerland and Penka Stateva, editors, *Presupposition and Implicature in Compositional Semantics*, pages 71–120. Palgrave Macmillan.

Danny Fox and Roni Katzir. 2011. On the characterization of alternatives. *Natural language semantics*, 19(1):87–107.

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

Michael Franke. 2013. Game theoretic pragmatics. *Philosophy Compass*, 8(3):269–284.

Daniel Fried, Jacob Andreas, and Dan Klein. 2018a. Unified pragmatic models for generating and following instructions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1951–1963, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Fried, Justin Chiu, and Dan Klein. 2021. Reference-centric models for grounded collaborative dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2130–2147, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018b. Speaker-follower models for vision-and-language navigation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3318–3329.

Simon Garrod and Gwyneth Doherty. 1994. Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*, 53(3):181–215.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text.

Kallirroi Georgila, James Henderson, and Oliver Lemon. 2006. User simulation for spoken dialogue systems: learning and evaluation. In *Interspeech*, pages 1065–1068.

Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 410–419, Cambridge, MA. Association for Computational Linguistics.

Noah D Goodman and Michael C Frank. 2016. Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, 20(11):818–829.

P Gorniak and D Roy. 2004. Grounded semantic composition for visual scenes.

Herbert P Grice. 1975. Logic and Conversation. In *Speech Acts*, pages 41–58. Brill.

Barbara J Grosz, Aravind K Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computional Linguistics*, 12(3):175–204.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook Dataset: Building Common Ground through Visually-Grounded Dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2016. Cooperative inverse reinforcement learning.

Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan. 2017. Inverse reward design. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. 2020. Interactive fiction games: A colossal adventure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7903–7910.

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.

Robert D. Hawkins, Michael Franke, Michael C. Frank, Adele E. Goldberg, Kenny Smith, Thomas L. Griffiths, and Noah D. Goodman. 2021. From partners to populations: A hierarchical Bayesian account of coordination and convention. *Psychological Review*.

Robert D Hawkins, Michael Franke, Michael C Frank, Adele E Goldberg, Kenny Smith, Thomas L Griffiths, and Noah D Goodman. 2022. From partners to populations: A hierarchical bayesian account of coordination and convention. *Psychological Review*.

Robert XD Hawkins, Michael C Frank, and Noah D Goodman. 2017. Convention-Formation in Iterated Reference Games. In *CogSci*.

Caitlin Hilliard and Susan Wagner Cook. 2016. Bridging gaps in common ground: Speakers design their gestures for their listeners. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(1):91–103.

Julia Bell Hirschberg. 1985a. *A Theory of Scalar Implicature (Natural Languages, Pragmatics, Inference)*. Ph.D. thesis, University of Pennsylvania, Ann Arbor, United States.

Julia Bell Hirschberg. 1985b. *A Theory of Scalar Implicature (Natural Languages, Pragmatics, Inference)*. PhD Thesis, University of Pennsylvania.

Jerry R Hobbs. 1979. Coherence and coreference. *Cognitive Science*, 3(1):67–90.

M Hodosh, P Young, and J Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Laurence Horn. 1984. Toward a new taxonomy for pragmatic inference: Q-based and r-based implicature. *Meaning, form, and use in context: Linguistic applications*, 11:42.

W. S. Horton and B. Keysar. 1996. When do speakers take into account common ground? *Cognition*, 59(1):91–117.

Julian Hough and David Schlangen. 2017. It's not what you do, it's how you do it: Grounding uncertainty for a simple robot. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, pages 274–282, New York, NY, USA. Association for Computing Machinery.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2022. A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meetup! a corpus of joint activity dialogues in a visual environment. In *23rd Workshop on the Semantics and Pragmatics of Dialogue*.

Gerhard Jäger. 2012. Game theory in semantics and pragmatics. *Semantics: An international handbook of natural language meaning*, 3:2487–2516.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3040–3049. PMLR.

Hong Jun Jeon, Smitha Milli, and Anca D Dragan. 2020. Reward-rational (implicit) choice: A unifying formalism for reward learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Dan Jurafsky and James H Martin. 2014. *Speech and Language Processing*. Pearson.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*, volume 42 of *Studies in Linguistics and Philosophy*. Springer.

Justine Kao, Leon Bergen, and Noah Goodman. 2014a. Formalizing the pragmatics of metaphor understanding. In *Proceedings of the annual meeting of the Cognitive Science Society*, volume 36.

Justine T Kao, Jean Y Wu, Leon Bergen, and Noah D Goodman. 2014b. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33):12002–12007.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the 53rd Annual*

Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 292–301, Beijing, China. Association for Computational Linguistics.

Fereshte Khani, Noah D. Goodman, and Percy Liang. 2018. Planning, inference and pragmatics in sequential language games. *Transactions of the Association for Computational Linguistics*, 6:543–555.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will I sound like me? improving persona consistency in dialogues through pragmatic Self-Consciousness.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240, Online and Punta Cana, Dominican Republic. 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.

Hannah Kirk, Yennie Jun, Paulius Rauba, Gal Wachtel, Ruining Li, Xingjian Bai, Noah Broestl, Martin Doff-Sotta, Aleksandar Shtedritski, and Yuki M Asano. 2021. Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 26–35, Online. Association for Computational Linguistics.

Noriyuki Kojima, Alane Suhr, and Yoav Artzi. 2021. Continual learning for grounded instruction generation by observing human following behavior. *Transactions of the Association for Computational Linguistics (TACL)*.

Y. Alex Kolchinski and Christopher Potts. 2018. Representing social media users for sarcasm detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1115–1121, Brussels, Belgium. Association for Computational Linguistics.

Alexander Koller, Konstantina Garoufi, Maria Staudte, and Matthew Crocker. 2012. Enhancing referential success by tracking hearer gaze. In *Proceedings*

of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 30–39, Seoul, South Korea. Association for Computational Linguistics.

Alexander Koller, Kristina Striegnitz, Andrew Gargett, Donna Byron, Justine Cassell, Robert Dale, Johanna D Moore, and Jon Oberlander. 2010. Report on the second NLG challenge on generating instructions in virtual environments (GIVE-2). In *Proceedings of the 6th International Natural Language Generation Conference*. The Association for Computer Linguistics.

Emiel Krahmer and Mariet Theune. 2010. *Empirical Methods in Natural Language Generation: Data-oriented Methods and Empirical Evaluation*. Springer.

R. M. Krauss and S. Weinheimer. 1966. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4(3):343–346.

Elisa Kreiss, Noah D Goodman, and Christopher Potts. 2021. Concadia: Tackling image accessibility with context. *arXiv preprint arXiv:2104.08376*.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 2020. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4392–4412, Online. Association for Computational Linguistics.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. Babytalk: understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903.

Royi Lachmy, Valentina Pyatkin, Avshalom Manevich, and Reut Tsarfaty. 2022. Draw me a flower: Processing and grounding abstraction in natural language. *Transactions of the Association for Computational Linguistics (TACL)*.

Angeliki Lazaridou and Marco Baroni. 2020. Emergent multi-agent communication in the deep learning era. *arXiv e-prints*, pages arXiv–2006.

Stephen Levinson. 2000. *Presumptive meaning: The theory of generalized conversational implicature*. MIT Press.

Stephen C Levinson. 1983. *Pragmatics*. Cambridge University Press.

David Lewis. 1969. *Convention: A Philosophical Study*. Harvard University Press, Cambridge, MA.

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 2017 Conference on Empirical Methods*

in Natural Language Processing, pages 2443–2453, Copenhagen, Denmark. Association for Computational Linguistics.

Elissa Li, Sebastian Schuster, and Judith Degen. 2021. Predicting Scalar Inferences From "Or" to "Not Both" Using Neural Sentence Encoders. In *Proceedings of the Society for Computation in Linguistics*, volume 4.

Jessy Lin, Daniel Fried, Dan Klein, and Anca Dragan. 2022. Inferring rewards from language in context. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob N Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, and Tim Rocktäschel. 2019. A survey of reinforcement learning informed by natural language. In *IJCAI*.

Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5988–5999.

Ellen M Markman and Gwyn F Wachtel. 1988. Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, 20(2):121–157.

Bill McDowell and Noah Goodman. 2019. Learning from omission. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 619–628, Florence, Italy. Association for Computational Linguistics.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daumé, III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 747–756, Avignon, France. Association for Computational Linguistics.

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Aparna S. Nadig and Julie C. Sedivy. 2002. Evidence of perspective-taking constraints in children's online reference resolution. *Psychological Science*, 13(4):329–336.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy. Association for Computational Linguistics.

Benjamin Newman, Reuben Cohn-Gordon, and Christopher Potts. 2020. Communication-based evaluation for natural language generation. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 116–126, New York, New York. Association for Computational Linguistics.

Allen Nie, Reuben Cohn-Gordon, and Christopher Potts. 2020. Pragmatic issue-sensitive image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1924–1938, Online. Association for Computational Linguistics.

Frans A Oliehoek and Christopher Amato. 2016. *A concise introduction to decentralized POMDPs*. Springer.

Tim Paek and Eric J Horvitz. 2000. Conversation as action under uncertainty. In *Uncertainty in Artificial Inteliigence*.

Gabriel Parent and Maxine Eskenazi. 2010. Lexical entrainment of real users in the let's go spoken dialog system. *Interspeech*.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. NOPE: A corpus of naturally-occurring presuppositions in English. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 349–366, Online. Association for Computational Linguistics.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *Computer Vision – ECCV 2020*, pages 647–664. Springer International Publishing.

Christopher Potts. 2012. Goal-Driven Answers in the Cards Dialogue Corpus. In *Proceedings of the 30th West Coast Conference on Formal Linguistics*, pages 1–20. Cascadilla Proceedings Project.

Zahar Prasov and Joyce Y Chai. 2008. What's in a gaze? the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*, IUI '08, pages 20–29, New York, NY, USA. Association for Computing Machinery.

Yewen Pu, Kevin Ellis, Marta Kryven, Josh Tenenbaum, and Armando Solar-Lezama. 2020. Program synthesis with pragmatic communication. *Neural Information Processing Systems*, 33:13249–13259.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Verena Rieser and Oliver Lemon. 2008. Learning effective multimodal dialogue strategies from Wizard-of-Oz data: Bootstrapping and evaluation. In *Proceedings of ACL-08: HLT*, pages 638–646, Columbus, Ohio. Association for Computational Linguistics.

Verena Rieser and Oliver Lemon. 2009. Natural language generation as planning under uncertainty for spoken dialogue systems. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 683–691, Athens, Greece. 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.

Seymour Rosenberg and Bertram D. Cohen. 1964. Speakers' and listeners' processes in a word-communication task. *Science*, 145(3637):1201–1203.

Alexis Ross and Ellie Pavlick. 2019. How well do NLI models capture verb veridicality? 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 2230–2240, Hong Kong, China. Association for Computational Linguistics.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2022. Large language models are not zero-shot communicators.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of Turn-Taking for conversation.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.

Emanuel A Schegloff. 1968. Sequencing in conversational openings. *American Anthropologist*, 70(6):1075–1095.

David Schlangen, Timo Baumann, and Michaela Atterer. 2009. Incremental reference resolution: The task, metrics for evaluation, and a bayesian filtering model that is sensitive to disfluencies. In *Proceedings of the SIGDIAL 2009 Conference*, pages 30–37, London, UK. Association for Computational Linguistics.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5387–5403, Online. Association for Computational Linguistics.

John R Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Sheng Shen, Daniel Fried, Jacob Andreas, and Dan Klein. 2019. Pragmatically informative text generation.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10740–10749.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In *International Conference on Learning Representations*.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Candace L Sidner, Christopher Lee, Cory D Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1):140–164.

Ricard V. Solé and Luís F. Seoane. 2015. Ambiguity in language networks. *The Linguistic Review*, 32(1):5–35.

Robert Stalnaker. 2002. Common Ground. *Linguistics and Philosophy*, 25(5):701–721.

Robert C Stalnaker. 1978. Assertion. In *Pragmatics*, pages 315–332. Brill.

Luc Steels and Tony Belpaeme. 2005. Coordinating perceptually grounded categories through language: a case study for colour. *Behavioral and Brain Sciences*, 28(4):469–89; discussion 489–529.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

DJ 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. Curran Associates, Inc.

Alane Suhr, Claudia Yan, Jack Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. 2019a. Executing instructions in situated collaborative interactions. 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 2119–2130, Hong Kong, China. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019b. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality encoder representations from transformers.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth J. Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011*. AAAI Press.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. Vision-and-Dialog navigation. In *Proceedings of the Conference on Robot Learning*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. *arXiv preprint arXiv:2204.03162*.

Michael Tomasello. 2005. *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.

David Traum and Jeff Rickel. 2002. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*, AAMAS '02, pages 766–773, New York, NY, USA. Association for Computing Machinery.

David R Traum. 1994. A computational theory of grounding in natural language conversation. Technical report, Rochester University Department of Computer Science.

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

Takuma Udagawa, Takato Yamazaki, and Akiko Aizawa. 2020. A linguistic analysis of visually grounded dialogues based on spatial expressions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 750–765, Online. Association for Computational Linguistics.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to Speak and Act in a Fantasy Text Adventure Game. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 673–683, Hong Kong, China. Association for Computational Linguistics.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Gabriella Vigliocco, Pamela Perniss, and David Vinson. 2014. Language as a multimodal phenomenon: implications for language learning, processing and evolution.

Adam Vogel, Max Bodoia, Christopher Potts, and Daniel Jurafsky. 2013a. Emergence of Gricean maxims from multi-agent decision theory. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1072–1081, Atlanta, Georgia. Association for Computational Linguistics.

Adam Vogel and Dan Jurafsky. 2010. Learning to follow navigational directions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 806–814.

Adam Vogel, Christopher Potts, and Dan Jurafsky. 2013b. Implicatures and nested beliefs in approximate decentralized-pomdps. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 74–80.

M A Walker, S J Whittaker, A Stent, P Maloor, J Moore, M Johnston, and G Vasireddy. 2004. Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, 28(5):811–840.

Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain. Association for Computational Linguistics.

Pei Wang, Junqi Wang, Pushpi Paranamana, and Patrick Shafto. 2020. A mathematical theory of cooperative communication. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:17582–17593.

Sida I. Wang, Samuel Ginn, Percy Liang, and Christopher D. Manning. 2017. Naturalizing a programming language via interactive learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 929–938, Vancouver, Canada. Association for Computational Linguistics.

Sida I. Wang, Percy Liang, and Christopher D. Manning. 2016. Learning language games through interaction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2368–2378, Berlin, Germany. Association for Computational Linguistics.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Bonnie Webber, Matthew Stone, Aravind Joshi, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.

Bonnie Lynn Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.

Jason D Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21(2):393–422.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Basil Blackwell.

Si On Yoon and Sarah Brown-Schmidt. 2019. Audience Design in Multiparty Conversation. *Cognitive Science*, 43(8):e12774.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, 2:67–78.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Zhou Yu, Dan Bohus, and Eric Horvitz. 2015. Incremental coordination: Attention-Centric speech production in a physically situated conversational agent. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 402–406, Prague, Czech Republic. Association for Computational Linguistics.

George Yule. 1996. *Pragmatics*, 1 edition. Oxford Introduction to Language Study. Oxford University Press.

Noga Zaslavsky, Jennifer Hu, and Roger P. Levy. 2021. A Rate–Distortion View of Human Pragmatic Reasoning. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 347–348, Online. Association for Computational Linguistics.

Luke Zettlemoyer, Brian Milch, and Leslie Kaelbling. 2008. Multi-agent filtering with infinitely nested beliefs. *Advances in Neural Information Processing Systems (NeurIPS)*, 21.

Ming Zhao, Peter Anderson, Vihan Jain, Su Wang, Alexander Ku, Jason Baldridge, and Eugene Ie. 2021. On the evaluation of Vision-and-Language navigation instructions. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Victor Zhong, Tim Rocktäschel, and Edward Grefenstette. 2020. Rtfm: Generalising to new environment dynamics via reading. In *ICLR*, pages 1–17. ICLR.

George K. Zipf. 1949. *Human behavior and the principle of least effort*. Addison-Wesley Press.

A Types of Context

In this section, we outline the broad types of context that lead to pragmatic enrichment of language, and point readers to Levinson (2000), Birner (2012), or Yule (1996) for a more comprehensive discussion. In this paper we focus mainly on visual and embodied contexts, for several reasons. First, human communication is typically situated in settings with modalities beyond language, which makes it important to capture in order to build NLP models that interact naturally with humans in the world. Indeed, recent work has argued that grounding is an essential component of language understanding (e.g., Bisk et al., 2020; Bender and Koller, 2020). Second, visual and embodied settings introduce enough complexity to elicit interesting linguistic behaviors and serve as a challenge for models, while still allowing researchers to control experimental aspects of the tasks. Finally, there has been a rapid increase in research on multimodal language learning, which makes studying pragmatics in these models and tasks timely and relevant.

A.1 Common Ground

To communicate successfully, speakers and listeners need to maintain a shared set of information, taken collectively to be *common ground* (e.g., Lewis, 1969; Stalnaker, 1978; Clark and Brennan, 1991; Traum, 1994; Stalnaker, 2002; Clark, 2015). A large body of work has demonstrated that humans produce and comprehend language in ways that depend on assumptions about the knowledge of their communicative partners (e.g., Krauss and Weinheimer, 1966; Horton and Keysar, 1996; Nadig and Sedivy, 2002; Clark and Bernicot, 2008; Hilliard and Cook, 2016; Yoon and Brown-Schmidt, 2019; Hawkins et al., 2021). Even in one-shot encounters where there is minimal partner-specific knowledge, the success of computational models of pragmatics (Frank and Goodman, 2012; Goodman and Frank, 2016) suggests that humans leverage a rich set of shared assumptions in pragmatic communication – from broad expectations that their partners abide by cooperative principles (e.g., Grice, 1975; Horn, 1984) to fine-grained knowledge of the potential utterances, meanings, and utterance-meaning mappings under joint consideration. Below, we discuss some key elements of common ground that give rise to pragmatically enriched meanings in naturalistic communication.

Norms of Interaction. As language is a social behavior, speakers and listeners typically abide by a set of norms. For example, Grice (1975) argues that it is generally understood that conversational partners act cooperatively and rationally. Grice also proposes a set of maxims that govern communication—rational speakers should be concise, informative, and relevant. These norms in turn give rise to a variety of nonliteral inferences known as *conversational implicatures*. Suppose, for example, Alice says to Bob: "Carl ate some of the cookies that we baked for the party". Bob likely draws the inference that Carl did not eat *all* of the cookies, even though the literal meaning of the utterance – that Carl ate at least one of the cookies – is logically compatible with such a scenario. This inference can be explained in the following way: if Alice knows that Carl ate all the cookies, and if she wants to be informative, then she would have said "Carl ate all of the cookies" instead.

Goals and Joint Actions. In addition to general norms of interaction, the particular social or task-related goals that elicit a linguistic expression can affect its meaning. The theory of *speech acts* (Searle, 1969; Austin, 1975) frames utterances (e.g., "please stand up") as actions on several levels: *locutionary*, the utterance itself; *illocutionary*, the intention (e.g., asking the listener to stand up); and *perlocutionary*, the actual effect that the action has in the world (e.g., the listener stands up). Context can have strong effects on the illocutionary and perlocutionary levels. This is particularly true for formal speech acts which can only take effect under *felicity conditions*, e.g. making a promise, or performing a marriage, but also occurs in commonplace situations e.g., asking "Did you get my email?" might be an indirect request to reply, or a direct question while debugging an internet connection. More generally, interlocutors typically recognize that they are undertaking *joint activities* together with their partners (Clark, 1996) and try to collaboratively plan and act to coordinate on and realize the relevant goals. These shared goals provide a source of context that enriches language.

Common Knowledge. Interpretation is aided by prior information that interlocutors bring to an interaction. For example, suppose Alice asks "What color was the woman's scarf?" and Bob answers "green". If Bob is a fashion designer with a keen eye for color palettes, this might implicate that

the scarf was a rather prototypical shade of green, and not olive green or chartreuse. On the other hand, if Bob doesn't know many specific color terms, Alice doesn't have grounds to infer that Bob meant to refer to a specific subspace of green. The world knowledge and commonsense relationships shared by conversational partners can also give rise to scalar implicatures formed by ad-hoc ordering relationships (Hirschberg, 1985b) and lead to pedagogic behavior (Chai et al., 2019).

Discourse Context. Communication is most often not a one-shot utterance, but instead unfolds over time. As a document or a conversation proceeds, the common ground can be updated with new information from the discourse context. At a basic level, discourse context includes previously-established information which can be referred to later on, whether explicitly (e.g., *a dog bounded into the room...* it *barked*) or implicitly (e.g., *a dog bounded into the room... Sam was surprised*). Information can also be introduced implicitly, for example through presupposition and accommodation (e.g., *Alex stopped smoking* presupposes that Alex smoked). Implicitly-introduced information can in some cases (implicature) also be reinforced or denied, e.g., *Carl ate some of the cookies; indeed, he ate all of them!*.

A.2 Multimodal Context

So far, we have discussed aspects of context given by social or linguistic factors. While all of the above types of context also arise in grounded and multimodal settings, the physical context in which communication is situated also plays an additional component in deriving linguistic meaning. As mentioned above, we focus on visual and embodied contexts in this paper, as these contexts reflect naturalistic communication while also allowing for fine-grained experimental control.

Visual. Visual context serves to disambiguate and enrich the language of meaning on multiple levels. On a level close to semantics, visual context can disambiguate word senses: e.g., "bank" likely has a different meaning in the caption of a photo of a river than in a photo of a city street. Referring expressions (e.g., *the red one*) often can only be resolved in a visual context, and deictic expressions, like English *here, there, this* and *that*, are frequently used in language to individuate referents in their immediate context, relying on mutual knowledge of what the speaker and listener can see (Clark

and Marshall, 1981). Reference intepretation can also be affected by the location of the speaker and hearer in the world (Birner, 2012), and can involve physical analogues of implicature (e.g., *the black one* might be a good description for a dark grey object if all other visible objects are lighter) (Golland et al., 2010; Udagawa et al., 2020).

Embodied. Facial expressions, gaze, and gestures (Cassell et al., 1994; Traum and Rickel, 2002; Sidner et al., 2005; Prasov and Chai, 2008; Bohus and Horvitz, 2010; Koller et al., 2012; Yu et al., 2015) can aid interpretation if they are available, e.g., a speaker first making eye contact with a listener, then looking at an intended object. Speakers can issue corrections if they are able to observe a listener carrying out actions (Clark and Krych, 2004; Koller et al., 2010; Thomason et al., 2019; Suhr et al., 2019a), and the physical movements of the listener can intentionally convey uncertainty (Hough and Schlangen, 2017) and intent (Dragan et al., 2013). Physical properties of the environment and tasks (Chai et al., 2019) and the capabilities of the speaker and listener (Chai et al., 2014), also affect the interpretation and generation of commands and requests — e.g., the classic pragmatic example *Can you pass the salt?*, which typically is an indirect request when spoken to a person, may have a literal interpretation when spoken to a robot with a faulty gripper.

B Unimodal Pragmatics

Although we primarily focus on the role of pragmatics in grounded environments, several text-only tasks that emphasize specific pragmatic phenomena also exist. For example, IMPPRES (Jeretic et al., 2020) and NOPE (Parrish et al., 2021) are benchmark datasets designed to test whether large language models can reliably predict implicatures and presuppositions, respectively. Similarly, Schuster et al. (2020) and Li et al. (2021) evaluate the ability of sentence encoding models to predict the rate at which humans draw scalar implicatures. Other datasets like the Self-Annotated Reddit Corpus (SARC) for sarcasm detection (Khodak et al., 2018) may also be viewed as pragmatic in nature (Kolchinski and Potts, 2018). While these datasets are limited to unimodal text, they have two main advantages over many multimodal tasks: (1) many unimodal pragmatic datasets are naturally-occurring, resulting in larger datasets with more realistic language, and (2) all of these datasets fo-

cus on specific pragmatic phenomena, such as presupposition. We suggest that future work on multimodal pragmatics should take inspiration from these properties and build larger and more targeted datasets.

A separate body of work has investigated situated language understanding through interactive fiction (IF) games (e.g., Ammanabrolu and Riedl, 2021; Hausknecht et al., 2020; Urbanek et al., 2019). IF games offer a framework for investigating goal-driven linguistic behaviors in a dynamic, richly structured world. Players observe natural-language descriptions of the simulated world, take actions via natural language, and receive scores based on their actions. The simulations are also partially observable, in that players must reason about the unerlying world state through incomplete textual descriptions of immediate surroundings. In this way, IF games avoid some of the practical issues of grounding in visual environments, while still requiring actions to be situated in rich, dynamic contexts. Furthermore, Shridhar et al. (2021) demonstrate that commonsense priors learned through IF games can be leveraged for better generalization in visually grounded environments, suggesting that text-only games induce representations that can be adapted to multimodal settings.