

# Grounded Agreement Games: Emphasizing Conversational Grounding in Visual Dialogue Settings

David Schlangen

Computational Linguistics / Department of Linguistics

University of Potsdam, Germany

david.schlangen@uni-potsdam.de

## Abstract

Where early work on dialogue in Computational Linguistics put much emphasis on dialogue structure and its relation to the mental states of the dialogue participants (e.g., Allen 1979, Grosz & Sidner 1986), current work mostly reduces dialogue to the task of producing at any one time a next utterance; e.g. in *neural chatbot* or *Visual Dialogue* settings. As a methodological decision, this is sound: Even the longest journey is a sequence of steps. It becomes detrimental, however, when the tasks and datasets from which dialogue behaviour is to be learned are tailored too much to this framing of the problem. In this short note, we describe a family of settings which still allow to keep dialogues simple, but add a constraint that makes participants care about reaching mutual understanding. In such *agreement games*, there is a secondary, but explicit goal besides the task level goal, and that is to reach mutual understanding about whether the task level goal has been reached. As we argue, this naturally triggers meta-semantic interaction and mutual engagement, and hence leads to richer data from which to induce models.

## 1 Introduction

If you're good at replying to a single request, are you also likely to be good at doing dialogue? Much current work seems to assume that the answer to this question is yes, in that it attempts a scaling up from single pairs of utterance plus response to longer dialogues: See, e.g., the work on *neural chatbots* following on from (Serban et al., 2016), where the main evaluation metric is “next utterance retrieval”; and on *visual dialogue* (Das et al., 2017), which views itself as a natural extension of *visual question answering* (Antol et al., 2015).

If you assume, however, that dialogue crucially is a *joint project* between its participants in a way

that single exchanges are not, you're likely to put more focus on *coordination phenomena* (Clark, 1996), but may end up with settings that combine multiple language capabilities in ways that current methods cannot yet seem to capture. (See, for example, the dialogues collected in (Zarri   et al., 2016).) In this short paper, we contribute a type of setting that introduces such coordination phenomena, while still allowing for control of the complexity of the resulting interaction.<sup>1</sup>

## 2 Visual Dialogue as Example of the Scaling Up Approach

Figure 1 shows an example interaction with the original Visual Dialogue system (Das et al., 2017). The competence of the system is impressive from a multimodal grounding perspective — it gets right several questions aiming at different aspects of the image. It is also clear, however, that this is a modest step beyond single-shot visual question answering (Antol et al., 2015). It seems that here the (human) questioner is doing all the work of keeping the dialogue alive, and there is little that suggest that the answerer is keeping any state about the dialogue. Later work by Kottur et al. (2018) on the “visual dialogue” dataset (Das et al., 2017) indeed identified *co-reference* in the questions as the main issue that distinguishes this setting from one-shot question answering.

One shortcoming of this setting—that the questioner was not provided with a good reason for *why* they are asking questions in the first place—was addressed in some related work: In the *Guess-What?* setting introduced by De Vries et al. (2017) at around the same time as Visual Dialogue, a questioner is asking polar questions about an image, with the goal of identifying an object known

<sup>1</sup>The datasets discussed in this paper can be viewed in a common format using the code at <https://github.com/clp-research/semplx>, (Schlangen, 2019).

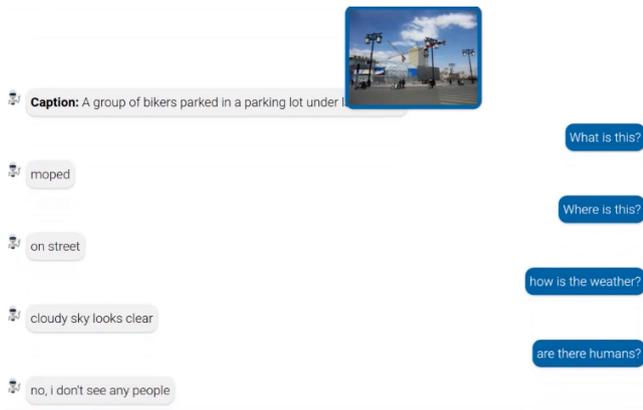


Figure 1: Original example of Visual Dialogue System, from official demo video (<https://vimeo.com/193092429>)

only to the questioner. In the ALICE variant of the visual dialogue setting, the questioner is asking questions with the later goal in mind of identifying the image from a set also containing distractor images (Chattopadhyay et al., 2017).<sup>2</sup> These variants of the general setting provide purpose to the questioner, but not to the answerer, which is the target of the modelling effort; and, crucially, it does not give the dialogue a *joint purpose*, a shared sense of semantic ownership of the interaction, which is a central feature of most genres of human interaction (Clark, 1996).

Coming back to the visual dialogue setting, it can be assumed that the crowd workers that created the original data did try to orient themselves to the usual maxims that govern conversational behaviour. However, being constrained by the rigid roles of questioner and answerer, and with the perceptual task being so easy for them, a need for dealing with miscommunication never arose for them and hence no such strategies can be learned from that data. That this is missing from the resulting agents can easily be shown in cases where something goes wrong, but normal repair mechanisms (see e.g. Hayashi et al., 2013) are not available, as in the example interaction we created shown in Figure 2.<sup>3</sup>

<sup>2</sup>Interestingly, this setup was created to evaluate dialogue agents trained on the visual dialogue dataset, not to collect another type of data.

<sup>3</sup>As the authors have made the laudable effort of opening their system to the general public, one can find similar examples of interactions that laypeople had with the agent and posted on social media; e.g. [https://twitter.com/r\\_speer/status/1037358574735904768](https://twitter.com/r_speer/status/1037358574735904768).

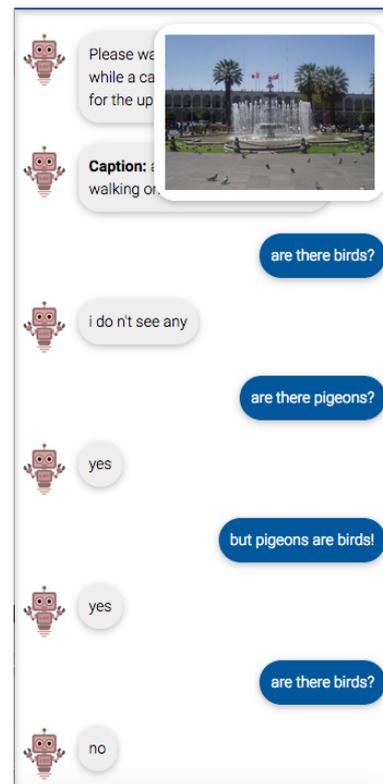


Figure 2: Interacting with a visual chat bot, see <http://demo.visualdialog.org/>

### 3 Agreement Games

In Herbert Clark’s (1996) model of dialogue,<sup>4</sup> the mutual need for ensuring understanding—“sufficient to current purposes”—is the main structuring force in dialogue. As a metaphor for this interaction management process, Clark uses the notion of a “secondary track”, on which the constant negotiation of this understanding happens. This can be done through quite subtle methods, such as simply producing a continuation that displays through its fit an understanding of the previous utterance, as well as through specially designed markers (such as feedback utterances like “uhu”). This model has been influential in the design of spoken dialogue systems (Traum, 1999; Skantze, 2005), where it has been incorporated in the design of dialogue state update rules. In the data-driven era, however, it seems to have become less well known, and many datasets almost appear as to be designed in such a way as to limit the possibility of grounding interactions.

<sup>4</sup>Bringing together ideas from Conversation Analysis (Sacks et al., 1974; Schegloff, 2007) and also the design of Human/Computer interaction (Norman, 1988).

The idea behind our setting of “agreement games” is to make this secondary track more prominent and hence more easy to pick up from the data, by making reaching mutual understanding on the answer to the game question an explicit goal. Or, in Clark’s term, the *grounding criterion* for answering the question is raised so as to make reaching mutual understanding on it an explicit, rather than as normally implicit, goal.

The representational challenge is that it is an abstract object—the understanding of the discourse—that is jointly constructed, purely through verbal actions.

**More formally** An *Agreement Game* is a dialogue game with two regular participants,  $\mathcal{P} = \{P_1, P_2\}$ , and a disinterested third participant,  $N$  (for *Nature*).  $N$  poses a question  $Q$  to the players  $\mathcal{P}$ , and provides them with information  $I$  required to answer the question; possibly split up over the players. If  $I$  contains visual information, we call the game a *Grounded Agreement Game*.

The players can exchange messages in an unrestricted way. The game ends when one of the players explicitly proposes an answer  $A$  and the other player explicitly agrees with the proposal. As the answer  $A$  will be based on a construal of  $I$ , the agreement on  $A$  is also an agreement on that construal. Optionally, a reward can be given to the players after they have provided their joint answer, tied to some measure of quality of  $A$ .

We illustrate the concept by discussing some instantiations that we have recently experimented with.<sup>5</sup>

## 4 Some Examples

### 4.1 The MeetUp Game

In the MeetUp game (Schlangen et al., 2018; Ilinykh et al., 2019),<sup>6</sup> the two participants are presented with an environment through which they

<sup>5</sup>Concurrently, two other datasets have recently been published using settings that can also be viewed as grounded agreement games: In the PhotoBook setting (Haber et al., 2019), two players must agree on which images from sets of (natural) images that they individually are presented with are identical. They do this repeatedly, so that occasion for re-references occurs. In the “One in Common” setting (Udagawa and Aizawa, 2019), players are presented with images of overlapping, but non-congruent views on a synthetic scene, and their goal is to determine which objects belong to the overlapping part.

<sup>6</sup><https://github.com/clp-research/meetup>

can (separately) navigate and which is represented to static photographs or real indoor scenes (e.g., a picture of a bedroom, a kitchen, etc.) Their goal is to meet up in the same room, of a type previously told to them. (E.g., they might be told at the start of the game: “your goal is to meet up in a room of type *kitchen*.”) As the positions of the players are not represented graphically, the only way they can be sure of whether they have reached that goal is by conversing (via chat messages). Once they have come to the conclusion that they are in the same room, they can end the game by each sending a special signal. If they are indeed in such a winning constellation, they will receive a bonus.

Unlike the Visual Dialogue setting discussed above, this setting ensures informational symmetry between the participants (both have access to the same *type* of information; but not the same information, as they can’t “see” each other). More importantly, however, the constraint that the game only ends if they both agree ensures a “commitment symmetry”, where the success of the game must be ensured by both participants. The design also provides for a clear “relevance place” at which an opportunity arises for *semantic negotiation*, namely, before the final decision is made.<sup>7</sup> An example of this is shown in the example below. (The number in the parentheses indicate the time, relative to the beginning of the interaction, when the utterance was made.)

- (1) B (00:00:34): okay I think I’m there if I understand utility room  
B (00:00:42): It has a washer and dryer  
A (00:00:46): I was wondering too. This is sorta like a laundry room.  
A (00:00:55): This has pet bowl on the floor below a window.  
B (00:01:00): ok... let us keep looking  
A (00:01:22): And a small kids looking suit hanging on the wall. And a big banner above the window.  
B (00:01:33): Are you saying a utility room is like a laundry room?  
B (00:02:00): let me find you  
A (00:02:07): Google says, a room equipped with appliances for washing and other domestic work.  
A (00:02:09): So I think so.

<sup>7</sup>“Relevance place” in analogy to the *transition relevance places* studied for turn taking in conversational analysis by Sacks et al. (1974).

## 4.2 The MatchIt Game

The MatchIt Game (Ilinykh et al., forthcoming) is a yet further simplified visual game. Here, the goal simply is to decide whether you and your partner are both looking at the same image (of the same genre as in MeetUp). In that sense, it is a reduction of the MeetUP game to the final stage, taking out the navigation aspect. As example (2) shows, this can similarly lead to meta-semantic interaction, where classifications are revised. As (3) shows, even in cases where a decision can be reached quickly, there can be an explicit mutual confirmation step, before the (silent) decision signal is sent.

- (2) B (00:00:25): white kitchen?  
A (00:00:25): im in a bathroom  
B (00:00:28): ah  
B (00:00:32): well wait  
B (00:00:38): there is something that looks like a big bath  
B (00:00:44): is it all white?  
A (00:00:54): yes its white and I see a bit of a kitchen  
A (00:01:11): yes  
B (00:01:11): are you sure it's a bathroom lol  
A (00:01:16): no its not a bathroom haha
- (3) A (00:00:24): i see stairs  
B (00:00:25): I see a staircase with a bike with wicker basket at the bottom of the staircase  
B (00:00:31): do you have a bike?  
A (00:00:39): no bike  
B: (00:00:46): okay..it is different  
A (00:00:54): yes

## 4.3 The Concept Learning Game

A third setting that we have explored (Attari et al., 2019) brings conceptual negotiation more clearly into the foreground. In that game, the players are presented with images of birds of particular species and are tasked with coming up with a description of common properties. Again, the final answer has to be approved by both participants. As (4) shows, this can lead to an explicit negotiation of conceptual content.

- (4) A: Looks like the birds under 2 have red-orange feet.  
B: The difference that I notice is that the birds in Section 1 are light feathered vs. the dark feathered birds of Section 2.  
A: Ah, I like your answer better than mine.  
B: /answer The birds in section 1 do not have red-orange feet like the birds in section 2. Also, the feathers of the birds in Section 1 are light-colored vs. the dark-colored feathers of the birds in Section 2.  
A: /agree

## 5 Conclusions

We have argued that some prominent current dialogue settings lack room for the occurrence of *coordination phenomena* prevalent in natural dialogue. We have shown a simple condition that brings out the need for coordination in an explicit way, but still can easily be added to controlled (and controllable) dialogue settings.

## References

- James F. Allen. 1979. *Plans, inference, and indirect speech acts*. In *Proceedings of the 17th Annual Meeting of the Association for Computational Linguistics*, pages 85–87, La Jolla, California, USA. 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 *International Conference on Computer Vision (ICCV)*.
- Nazia Attari, Martin Heckmann, and David Schlangen. 2019. From explainability to explanation: Using a dialogue setting to elicit annotations with justifications. In *Proceedings of SIGdial 2019*, Stockholm, Sweden.
- Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. 2017. *Evaluating Visual Conversational Agents via Cooperative Human-AI Games*. *ArXiv*.
- Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

- Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. GuessWhat?! Visual object discovery through multi-modal dialogue. In *CVPR 2017*.
- Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational 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 2019 meeting of the Association for Computational Linguistics*, Florence, Italy.
- Makoto Hayashi, Geoffrey Raymond, and Jack Sidnell. 2013. *Conversational Repair and Human Understanding*. OAPEN Library. Cambridge University Press.
- Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meetup! a corpus of joint activity dialogues in a visual environment. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2019 / LondonLogue)*, London, UK.
- Satwik Kottur, José M F Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. [Visual Coreference Resolution in Visual Dialog using Neural Module Networks](#). *ArXiv*.
- Donald A. Norman. 1988. *The Design of Everyday Things*. Basic Books, Inc., New York, NY, USA.
- Harvey Sacks, Emanuel A. Schegloff, and Gail A. Jefferson. 1974. A simplest systematic for the organization of turn-taking in conversation. *Language*, 50:735–996.
- Emanuel A. Schegloff. 2007. *Sequence Organization in Interaction*. Cambridge University Press, Cambridge, UK.
- David Schlangen. 2019. Natural language semantics with pictures: Some language & vision datasets and potential uses for computational semantics. In *Proceedings of the International Conference on Computational Semantics (IWCS)*, Gothenburg.
- David Schlangen, Nikolai Ilinykh, and Sina Zarrie. 2018. MeetUp! A Task For Modelling Visual Dialogue. In *Short Paper Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue (AixDial / semdial 2018)*.
- Iulian V. Serban, Alessandro Sordani, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proceedings of the Meeting of the Association for the Advancement of Artificial Intelligence (AAAI)*.
- Gabriel Skantze. 2005. Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication*, 45(3):325–341.
- David R. Traum. 1999. Computational models of grounding in collaborative systems. In *Proceedings of the AAI Fall Symposium on psychological models of communication*, pages 124–131.
- Takuma Udagawa and Akiko Aizawa. 2019. [A Natural Language Corpus of Common Grounding under Continuous and Partially-Observable Context](#). In *Proceedings of AAI*.
- Sina Zarrieß, Julian Hough, Casey Kennington, Rames Manuvinakurike, David DeVault, Raquel Fernández, and David Schlangen. 2016. Pentoref: A corpus of spoken references in task-oriented dialogues. In *Proceedings of LREC 2016*, Portoroz, Slovenia.