Communication Emergence in a Goal-Oriented Environment: Towards Situated Communication in Multi-Step Interactions

Aleksandra Kalinowska
Northwestern University, Evanston, IL
& DeepMind, Edmonton, AB
ola@u.northwestern.edu

Elnaz Davoodi
DeepMind
Montreal, QC
elnazd@deepmind.com

Kory W. Mathewson
DeepMind
Montreal, QC
korymath@deepmind.com

Todd D. Murphey
Northwestern University
Evanston, IL
t-murphey@northwestern.edu

Patrick M. Pilarski
DeepMind & University of Alberta
Edmonton, AB
ppilarski@deepmind.com

Abstract

Effective communication enables agents to collaborate to achieve a goal. Understanding the process of communication emergence allows us to create optimal learning environments for multi-agent settings. Thus far, most of the research in the field explores *unsituated* communication in one-step referential tasks. These tasks are not temporally interactive and lack time pressures typically present in natural communication and language learning. In these settings, reinforcement learning (RL) agents can successfully learn *what* to communicate but not *when* or *whether* to communicate. Convergence is slow and agents tend to develop non-efficient codes, contrary to patterns observed in natural languages. Here, we extend the literature by assessing emergence of communication between RL agents in a temporally interactive, cooperative task of navigating a gridworld environment. Moreover, we *situate* the communication in the task—we allow the acting agent to actively choose between (i) taking an environmental action and (ii) soliciting information from the speaker, imposing an opportunity cost on communication. We find that, with situated communication, agents converge on a shared communication protocol more quickly. The acting agent learns to solicit information sparingly, in line with the Gricean maxim of quantity. In the same multi-step navigation task, we compare real-time to upfront messaging. We find that real-time messaging significantly improves communication emergence , suggesting that it is easier for agents to learn to communicate if they can exchange information when it is immediately actionable. Our findings point towards the importance of studying language emergence through situated communication in multi-step interactions.

Keywords: emergent communication; multi-agent reinforcement learning; cooperative AI

Acknowledgements

Work conducted while AK was an intern with DeepMind. We greatly appreciate conversations with Florian Strub, Ivana Kajic, and Michael Bowling at DeepMind that helped shape the research.

1 Introduction

Communication is a key skill for collaboration and hence largely beneficial in multi-agent settings. As humans, we share well-established communication protocols that have evolved over thousands of generations—shaped by functional pressures, such as time and articulation effort—to suit the needs of our daily tasks and to take advantage of our cognitive and physical capabilities. As an example, natural languages are known to be compositional, making them easier to learn and use [Kirby and Hurford, 2002]. Similarly, when we communicate, we are known to follow Grice's maxim of quantity—we try to be as informative as possible, giving only as much information as is needed [Grice, 1975]. If future artificial systems are to cooperate with humans, it will be beneficial for their communication protocols to follow these patterns. Understanding communication emergence among artificial agents will allow us to create optimal learning environments for multi-agent settings and supports the design of machines that will work well with each other and with people [Crandall et al., 2018, Steels, 2003].

With a recent increase in available computational power, the field has seen a lot of progress [Wagner et al., 2003, Lazaridou and Baroni, 2020]. Thus far, emergent communication has largely been studied in one-step referential games, such as the Lewis signalling task [Chaabouni et al., 2019, Li and Bowling, 2019, Lazaridou et al., 2018]. This type of learning environment is known to successfully enable language development [Kirby and Hurford, 2002] but does not allow agents to accelerate the learning process through back-and-forth interaction. In line with prior work [Evtimova et al., 2018] , we show that multi-step interactions can be beneficial for communication emergence, both in terms of agents' ability to converge to a collaborative solution and the time needed for convergence.

In most studies, the emerged language structures are analyzed for shared commonalities with natural languages, such as compositionality or encoding efficiency. Although desired, it is nontrivial for such properties to emerge spontaneously between artificial agents [Kottur et al., 2017]. For instance, artificial agents tend towards an anti-efficient encoding [Chaabouni et al., 2019]. This likely happens because in the Lewis signalling task, as well as in other simulated environments [Cao et al., 2018], agents have no incentive to be concise. In our approach, we show it is possible to obtain sparse communication by providing the agent with an action-communication trade-off, in line with the idea that *reward is enough to shape language* [Silver et al., 2021].

In our work, we explore the emergence of communication in a cooperative multi-step navigation task. Importantly, we *situate* the communication in the environment—we allow the acting agent to actively choose between (i) taking an action to move through the maze and (ii) soliciting information from the speaker. Our contributions are two-fold: (1) we study the emergence of *situated* communication and how it affects the communication protocol, and (2) we explore the effect of multi-step interactions on communication emergence.

2 Experimental Setup

The environment. We define a cooperative navigation task as a Markov Decision Process (MDP) with two reinforcement learning (RL) agents. The environment is set up as a pixel-based gridworld (7 by 7 cells). As illustrated in Figure 1, the maze includes 3 T-junctions, each allowing a right and left turn. Features of the world are represented with colors: walls are black, the maze is white, the agent is green, and the target is blue. The features are encoded with binary vectors.

The agents. There are two agents, a speaker and a listener (i.e. acting agent). The listener is embedded inside the gridworld and can take actions to move between cells. The action space of the listener spans 5 actions [move up, move down, move right, move left, stay in place]. The listener's observation consists of the environmental view (if any) concatenated with the message from the speaker. We test the listener under two conditions: (1) with no visibility, where the listener's observation consists solely of the speaker's message, and (2) with partial visibility, where the listener can see the 3 pixels directly in front of them. The second variant gives the listener environmental context to take actions without needing to rely solely on communication. The speaker does not reside within the gridworld and cannot take environmental actions (i.e. navigate the maze) but instead can communicate information to the listener. The message space of the speaker spans 5 symbols $[0, 1, ..., 4]$. At each timestep, the speaker can see the entire gridworld, including the location of the agent and the location of the goal. The speaker's view of the world map is rotated to align with the direction that the listener is facing. In our experiments, we test agents with and without memory. Agents without memory have to rely only on their current observations to generate messages or pick actions. Agents with memory have an internal representation of the history of an episode—they can use accumulated knowledge from prior timesteps to make decisions in the current timestep.

Agent architectures. The speaker and the listener share the same architecture without sharing weights or gradient values. They both have a 2-layer Convolutional Neural Network (CNN) that generates an 8 to 32 bit representation of the environment. In the case of the listener, this representation of the environment gets concatenated with the message received from the speaker. In both cases, the vector gets passed into a fully connected layer that generates the agent's action (a move or a message). Agents with memory have an additional single-layer LSTM [Hochreiter and Schmidhuber, 1997] after their fully connected layer. We train the agents using neural fitted Q learning [Riedmiller, 2005], with an Adam optimizer [Kingma and Ba, 2015] and $Q_t(\lambda)$ where $\lambda = 0.9$ and $\gamma = 0.99$. During training, agents use an ϵ-greedy policy with the exploration rate set as $\epsilon = 0.01$.

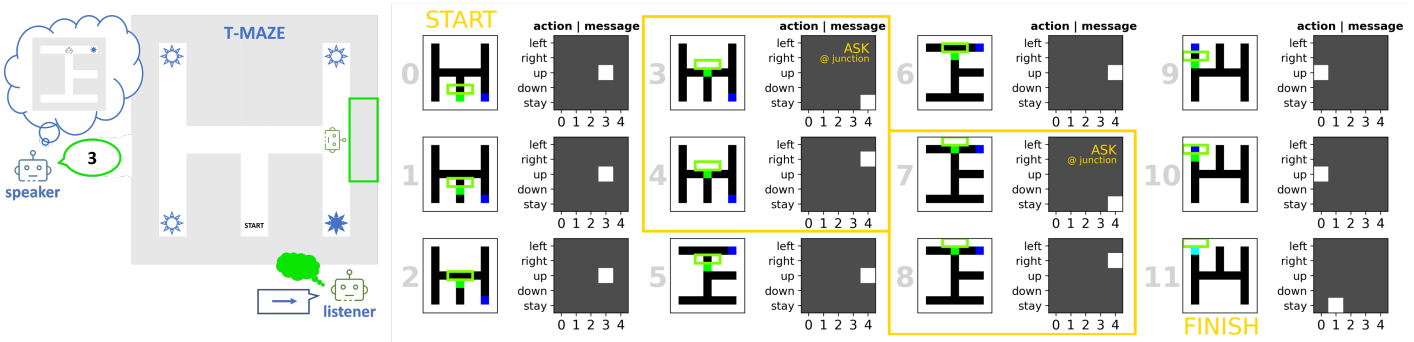

Figure 1: **Experimental setup and a walk-through of an example episode with *situated* communication.** In the maze on the left, stars indicate possible goal locations. To the right, we visualize an example episode of an active listener with partial visibility. The listener learns to solve the task optimally, deciding to stay and ask for information when at a junction (twice during the episode).

The task. The goal of the agents is to cooperate so that the listener reaches the target. In each experimental episode, both agents receive a reward $R = 1$ if the listener reaches the target before the episode terminates. Episode timeout is set to 100 steps. The goal locations are randomly assigned to one of 4 corners in the T-maze, as indicated with stars in Figure 1. In each episode, the listener agent starts from the bottom middle cell. We evaluate agent performance using 3 metrics: (1) task success (via a mean return per episode), (2) optimality of task solution (via a normalized reward per step), and (3) communication sparsity (via the number of asks per episode).

Communication modes. We compare three modes of communication: (1) real-time messaging with a passive listener, (2) real-time messaging with an active listener, and (3) upfront messaging with a passive listener. In mode 1, the speaker generates a 1-token message at every timestep and the message gets broadcasted to the listener before they choose an action. The speaker has to reason about both the content and timing of their message, deciding both *what* and *when* to communicate. In mode 2, we implement real-time messaging with an active listener. Here, the message is only broadcasted to the listener after they ask for information. The active listener can solicit to receive information in the next timestep by choosing to stay in place at the current timestep. The active listener has to learn *whether* to communicate at all. In mode 3, the speaker generates a 1-, 2-, or 3-token message at the beginning of each episode and that message gets broadcasted to the listener at each timestep throughout the episode.

We define the communication in mode 1 and 3 as *unsituated*—it is free and guaranteed to the agent at every timestep. There is no opportunity cost to communication. The communication in mode 2 is *situated*—we allow the acting agent to actively choose between (i) taking an environmental action and (ii) soliciting information from the speaker. As a result, the active listener experiences an opportunity cost to communication. They have to forego a move in the environment in order to obtain information from the speaker and make an informed decision.

Experimental parameters. For each experiment, we run a hyperparameter sweep over learning rates of the speaker and listener $\alpha = [10^{-5}, 10^{-6}, 10^{-7}]$ and over the size of the environmental representation $s = [4, 8, 16, 32]$. We run the simulation with each hyperparameter setting with 10 different random seeds. In the figures, we present the best performing agent pair from our hyperparameter sweep and/or the mean over the 10 replicas with the same hyperparameters as the best performing pair. When we plot metric means, we include the standard error of the mean.

3 Results

We start by generating a baseline for the task. Experiments confirm that without communication agents are unable to reliably solve the task. Under partial visibility, agents without communication can succeed in the task with a mean return of ≈ 0.25 per episode. With memory, baseline performance improves. However, due to the random location of the target, the listener cannot consistently solve the task in an optimal number of steps, converging to a normalized reward per step of ≈ 0.45. When allowed to communicate, all agents in the T-maze environment learn to solve the task and best agent pairs find an optimal solution, as visualized with the grey line in Figure 2.

The pressure of time in a multi-step interaction can incentivise sparse communication. In the first set of experiments, we evaluate the impact of situated communication on language emergence. Figure 1 shows a step-by-step example episode for an active listener with partial visibility. Under the partial visibility condition, information solicitation takes place mostly at the junctions, where the acting agent has a choice between two viable environmental actions. The active listener can learn to near optimally solicit information, asking ≈ 9.76 and ≈ 2.06 times per episode under the two visibility conditions, respectively.

In Figure 2 on the left, we illustrate the learning curves of the best performing agent pairs. Note that the active listeners ask for information frequently at the beginning of the interaction and gradually less over time. This suggests that agents initially have opportunities to align on a protocol. Over time, listeners learn *when* and *whether* to solicit information as

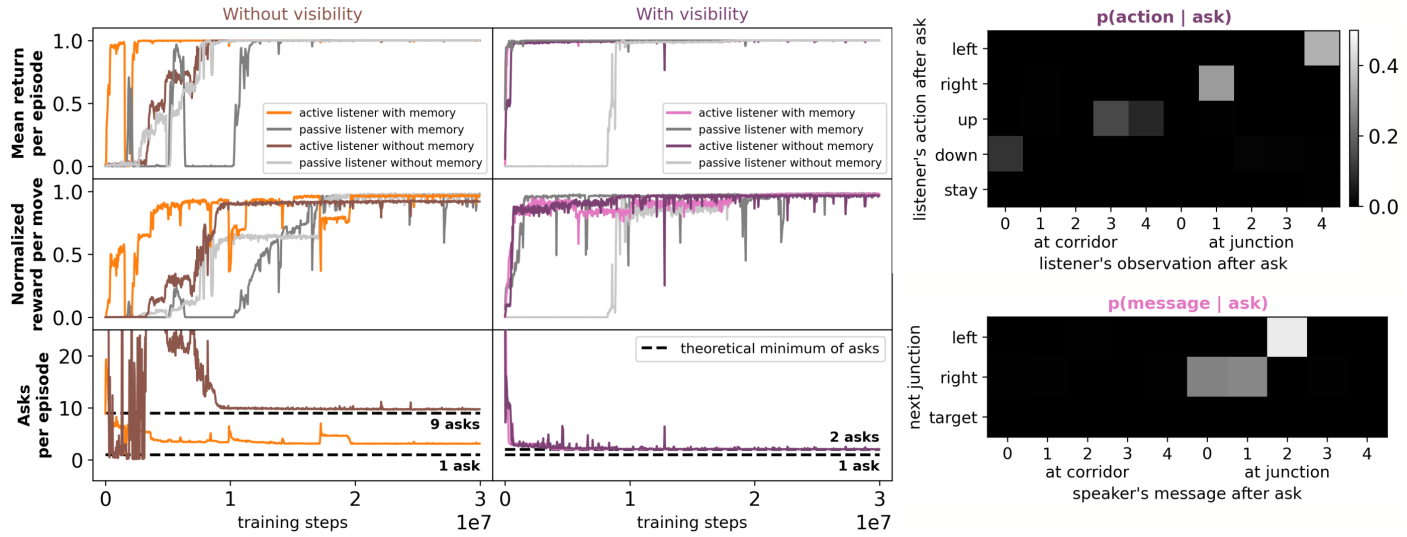

Figure 2: **Best performing pairs of agents with an active listener.** Under all conditions (with/without memory and with/without visibility) agents learn to solve the task via the shortest path. Listeners without memory learn to query the speakers in the optimal number of asks (once per step when the listener has no visibility and once per junction when the listener sees environmental context). Listeners with memory persist to ask for information when it is immediately actionable (instead of once at the beginning of an episode).

communication comes with a cost. We also observe that the best performing agent pairs with an active listener converge to an optimal solution faster than the best performing agent pairs with a passive listener. The results suggest that situated communication not only allows agents to learn a sparse communication protocol, in line with the Gricean maxim of quantity, but also has a positive impact on convergence speed.

The active listener exhibits a preference for just-in-time communication. Interestingly, when we test situated communication between agents with memory, agents continue to ask for information at the junctions (note the bottom heatmap in Figure 2). This is non-obvious—given memory, the active listener could ask for information at any point in the maze. In fact, if the agent were to be optimally sparse, they could (1) ask for information only once at the beginning of an episode, (2) receive a message encoding the address of the target, and (3) follow the relevant policy from memory. Instead, the active listener with memory learns sparse communication relative to a passive listener but they do not achieve the theoretically maximal sparsity, continuing to ask for information at the junctions when it is immediately actionable. This result suggests that it may be easier for agents to succeed at the task when they can control the timing of communication.

Real-time communication improves language emergence compared to upfront messaging. In our final experiment, we compare the real-time communication protocol (mode 1) with upfront messaging (mode 3). In both scenarios, the theoretical capacity of the communication channel allows the agents to communicate the necessary information, whether the agents choose to communicate directions, e.g., 'turn right', or a goal address, e.g., 'top left corner'. With upfront messages of length 1, 2, and 3, the speaker has 5, 25, or 125 unique messages available for communication, respectively.

With both real-time and upfront messaging, agents succeed in establishing a successful communication protocol when the listener has partial visibility—they converge to a mean return of 1 per episode. With no visibility for the acting agent, agent pairs with upfront messaging do not succeed at solving the task. Moreover, the real-time agents are more likely to converge to an optimal solution, being able to solve the T-maze task in 9 moves. With 1 upfront token, even the best agents learn to at-best solve the task in 12 steps. These agents seem to reliably learn unique messages to encode the action required at the first turn or the right/left part of the address, but they do not establish a unique encoding for the top/bottom portion of the address, as visible in the top heatmap in Figure 3. With 3 upfront tokens, the best agent pair agrees on 4 distinct symbols to encode the 4 possible goal locations. However, convergence is slow and on average agent pairs perform less optimally than under the real-time communication paradigm. We hypothesize that there are benefits to allowing communication to emerge from multi-step interactions. Our findings suggest that it is easier for agents to learn to communicate if they can exchange information when it is immediately actionable.

4 Conclusion & Discussion

Our results point towards the importance of studying emergent communication in multi-step interactions. The interactive aspect of communicating over time enables agents to learn both *what* and *when* to communicate. It improves overall task performance and speeds up convergence to an optimal solution. Secondly, we find that there is value in situating the communication in the task and giving the listener agency to choose *whether* to communicate at all. In this way, we allow the reward to shape the emergent communication protocol to exhibit properties of natural languages, such as sparsity.

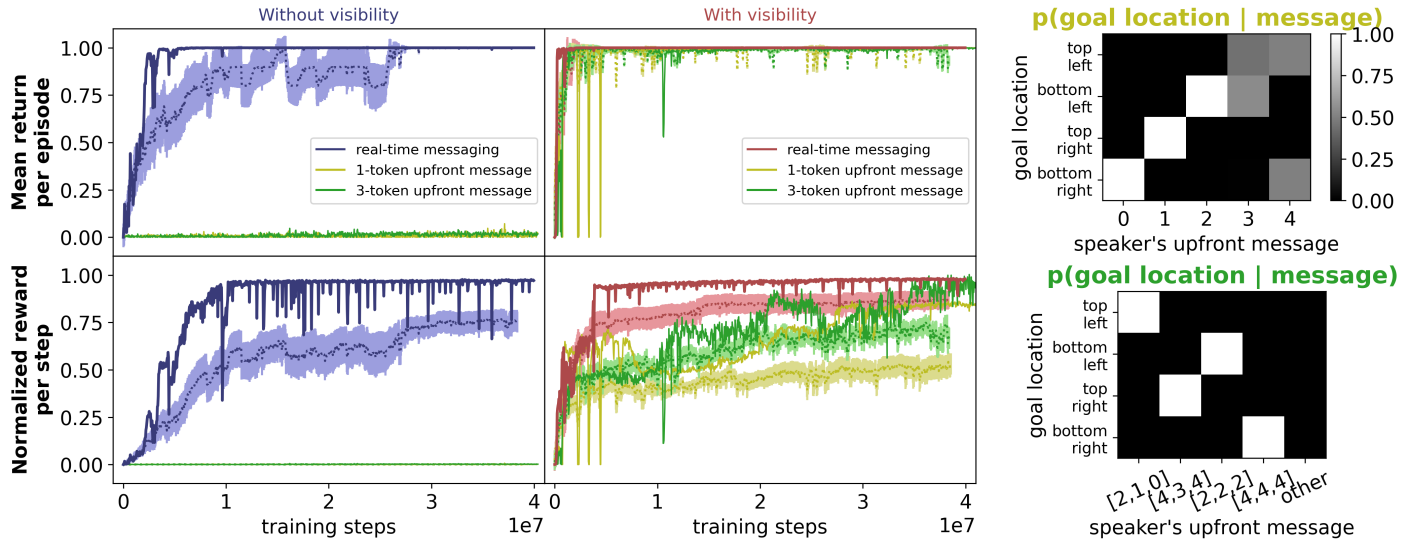

Figure 3: **Comparison of upfront and real-time messaging; agents have memory.** Real-time messaging improves convergence on a successful communication protocol. With upfront messaging, agents learn to solve the task before episode timeout when the listener has partial visibility. However, convergence is slow and agents are unlikely to solve the task in the optimal number of steps.

Our ongoing work will expand this idea and situate both the speaker and listener in the environment, allowing both agents to communicate and take actions in the gridworld environment.

References

[Cao et al., 2018] Cao, K., Lazaridou, A., Lanctot, M., Leibo, J. Z., Tuyls, K., and Clark, S. (2018). Emergent communication through negotiation. *Int. Conf. on Learning Representations*.

[Chaabouni et al., 2019] Chaabouni, R., Kharitonov, E., Dupoux, E., and Baroni, M. (2019). Anti-efficient encoding in emergent communication. *Advances in Neural Information Processing Systems*.

[Crandall et al., 2018] Crandall, J. W., Oudah, M., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M. A., Rahwan, I., et al. (2018). Cooperating with machines. *Nature Communications*.

[Evtimova et al., 2018] Evtimova, K., Drozdov, A., Kiela, D., and Cho, K. (2018). Emergent communication in a multi-modal, multi-step referential game. *Int. Conf. on Learning Representations*.

[Grice, 1975] Grice, H. P. (1975). Logic and conversation. *Speech Acts*.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*.

[Kingma and Ba, 2015] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *Int. Conf. on Learning Representations*.

[Kirby and Hurford, 2002] Kirby, S. and Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. *Simulating the Evolution of Language*.

[Kottur et al., 2017] Kottur, S., Moura, J. M., Lee, S., and Batra, D. (2017). Natural language does not emerge 'naturally' in multi-agent dialog. *Conf. on Empirical Methods in NLP*.

[Lazaridou and Baroni, 2020] Lazaridou, A. and Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*.

[Lazaridou et al., 2018] Lazaridou, A., Hermann, K. M., Tuyls, K., and Clark, S. (2018). Emergence of linguistic communication from referential games with symbolic and pixel input. *Int. Conf. on Learning Representations*.

[Li and Bowling, 2019] Li, F. and Bowling, M. (2019). Ease-of-teaching and language structure from emergent communication. *Advances in Neural Information Processing Systems*.

[Riedmiller, 2005] Riedmiller, M. (2005). Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method. *European Conf. on Machine Learning*.

[Silver et al., 2021] Silver, D., Singh, S., Precup, D., and Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*.

[Steels, 2003] Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Sciences*.

[Wagner et al., 2003] Wagner, K., Reggia, J. A., Uriagereka, J., and Wilkinson, G. S. (2003). Progress in the simulation of emergent communication and language. *Adaptive Behavior*.