Reinforcement Communication Learning in Different Social Network Structures

Marina Dubova [1]   **Arseny Moskvichev** [2]   **Robert L. Goldstone** [1 3]

Abstract

Social network structure is one of the key determinants of human language evolution. Previous work has shown that the network of social interactions shapes decentralized learning in human groups, leading to the emergence of different kinds of communicative conventions. We examined the effects of social network organization on the properties of communication systems emerging in decentralized, multi-agent reinforcement learning communities. We found that the global connectivity of a social network drives the convergence of populations on shared and symmetric communication systems, preventing the agents from forming many local "dialects". Moreover, the agent's degree is inversely related to the consistency of its use of communicative conventions. These results show the importance of the basic properties of social network structure on reinforcement communication learning and suggest a new interpretation of findings on human convergence on word conventions.

1. Introduction

Human languages evolve as complex adaptive systems, driven by micro-level processes and constraints (such as individual learning mechanisms and perceptual biases), macro-level factors (such as a topology of social interactions), and the history of their development (Steels, 2000; Christiansen & Chater, 2008; "Five Graces Group" et al., 2009).

Linguistic communication depends on the shared knowledge of word-to-meaning mapping conventions (Lewis, 1975), upon which the population converges through the local interactions between agents, often with no central controller

[1]Cognitive Science Program, Indiana University, Bloomington, IN, USA [2]Department of Cognitive Sciences, University of California, Irvine, CA, USA [3]Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA. Correspondence to: Marina Dubova <mdubova@iu.edu>.

available (Baronchelli, 2018). Empirical studies of human learning demonstrate that groups quickly converge on new communicative conventions in "decentralized" settings (Garrod & Doherty, 1994; Selten & Warglien, 2007).

Multi-agent reinforcement learning to communicate (MARLC), however, faces instability challenges if no central optimization is introduced (Bernstein et al., 2002; Laurent et al., 2011; Matignon et al., 2012). This influences the ability of groups consisting of reinforcement learners to converge on efficient and stable communication systems which are shared by all their members. Therefore, different methods of centralized control and optimization have been proposed to stabilize MARLC (Sukhbaatar et al., 2016; Foerster et al., 2016; Lowe et al., 2017; Pesce & Montana, 2020; Foerster et al., 2018). Central optimization makes the simulations more brittle and less flexibly adaptive, and, potentially, less promising in developing communication systems as freely expressive and well-optimized for their users (Gibson et al., 2019) as natural languages.

We argue that empirical evidence on individual- and population-level factors that drive decentralized learning in human groups can guide simulations of language evolution in MARLC settings. In this work, we explore whether the social network organization shapes the properties of communication systems that arise through decentralized MARLC in simplified settings.

1.1. Human Learning in Different Social Network Structures

Convergence of human groups on word conventions is dramatically affected by the social topology that determines the
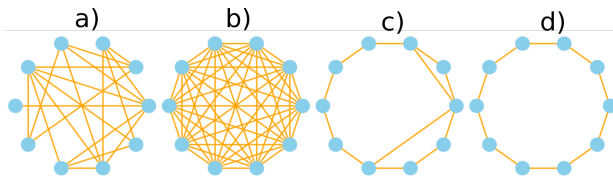

Figure 1. Types of social topologies tested in our experiments. a) Random (Erdos, 1959), b) Fully-connected (clique), c) Small-world (Newman & Watts, 1999), d) Ring
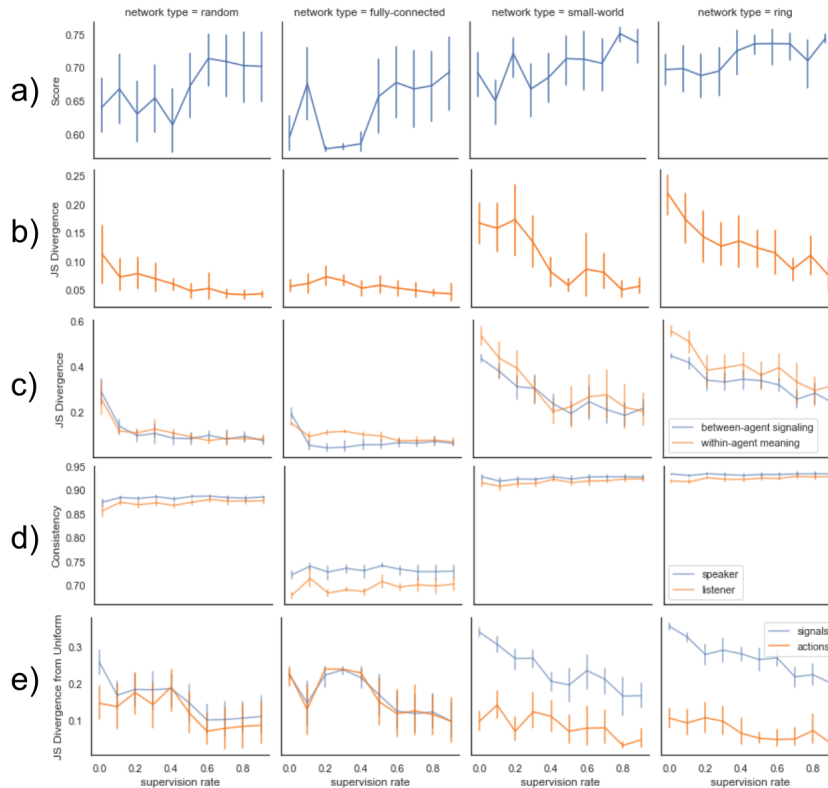
Figure 2. The Communication Analysis Metrics with 95% Confidence Intervals for Experiment 1. a) Average rewards b) Between-agent signal-action mapping divergence c) Signaling divergence (blue) and within-agent signal-action mapping divergence (orange) d) Average speaking (blue) and listening (orange) consistencies e) Average predictability of agents' signals (blue) and actions (orange).

possible interactions between participants, as demonstrated by the naming game experiment (Centola & Baronchelli, 2015). In particular, when arranged in a social network with many local connections (e.g. "ring" topology (Fig. 1d)) or a randomly-connected network (Fig. 1a), large groups converge on many local word conventions, reaching no global consensus. However, if each person is equally likely to interact with any other person in the group ("clique" (Fig. 1b)), global consensus is easily achieved with no centralization. Other studies on decentralized problem solving in human groups demonstrated that social network organization shapes the multi-agent optimization process, with different network types being beneficial for different types of optimization landscapes. High local connectivity supports independent local exploration, whereas high global connectivity helps groups to converge on a shared solution, choosing among the local ones (Fang et al., 2010; Mason et al., 2008; Mason & Watts, 2012; Lazer & Friedman, 2007; Wisdom & Goldstone, 2011).

In this study, we looked at how the type of social network organization, its average degree, and local connectivity affect the results of communication learning in groups of deep reinforcement learning agents.

2. Method

2.1. Coordination Game

Every game episode involves two agents, randomly assigned to speaker and listener roles. The speaker produces a message, which is transmitted to the listener. Then, both agents independently choose an action[1]. If the actions match, the agents receive a reward. We add an additional penalty for overusing any specific action to avoid degenerate solutions that ignore the communication channel. This setting encapsulates the most basic form of a coordination game that benefits from the formation of communicative conventions. Please, see supplementary materials for more detail.

2.2. Types of Social Network Organization

Social network determines how the agents are sampled to play games with one another. For each game round, one agent is selected randomly, and then its partner is selected from its neighbors. Thus, an agent can only play a game with one of its immediate neighbors in the social network. In all the simulations, the social network size was set to

---

[1]Action and signaling sizes were set to 4 in all the simulations.

10. The networks were undirected, and the self-connections were not allowed. We tested 4 types of social networks in our experiments (Fig. 1):

1. Random (ER). Random graph is generated by connecting the nodes with equal probability p (Erdos, 1959).

2. Fully-connected (clique). In clique, all the nodes are connected to each other.

3. Ring. In the ring network, all the nodes have exactly two neighbors, and connections form a single continuous path.

4. Small-world. The small-world network is generated by adding new, "global", connections to the ring network with constant probability p (Newman & Watts, 1999).

2.3. Agents

We used simple feed-forward neural networks to represent the agents. The networks were trained using a vanilla deep Q-learning algorithm (Mnih et al., 2015) with an added proportion of bottom-up driven "supervising" feedback (see supplementary materials). We tried to avoid any centralized optimization, popular in MARLC settings, to study the ability of different networks to self-organize conventions.

2.4. Metrics

We used a number of information-theoretic metrics, developed in Dubova & Moskvichev (2020) and Lowe et al. (2019) to comprehensively evaluate the communication protocols: speaker & listener consistency, between- and within-agent signal-action mapping divergence, signaling divergence, and behavioral predictability (see supp. materials).

3. Results

3.1. Experiment 1: types of social network organization

We simulated multi-agent learning in 4 types of social networks: **ring** (avg degree=2, var=0), **random** (avg degree=2, var=1.6, p(connection)=0.2), **small-world** (avg degree=2.2, p(add global connection)=0.2), and **clique** (avg degree=9, var=0). We also varied the supervision rate (from 0.0 to 0.9 with a step of 0.1). MARLC in each combination of these two conditions was simulated 10 times to get statistical estimates of the communication metrics. In all the experiments, each simulation consisted of 120000 game rounds.

The average speaker and listener consistencies were highest in the ring and small-world networks, and the lowest in the clique structure. Consistency scores did not vary with the supervision rate (Fig. 2d). These patterns indicated the potential dependency of the consistency scores on a single factor: agent's degree (the clique had the highest possible average degree). This hypothesis is tested in Experiment 2.

The agents in random and fully-connected social networks developed almost perfectly symmetric and homogeneous
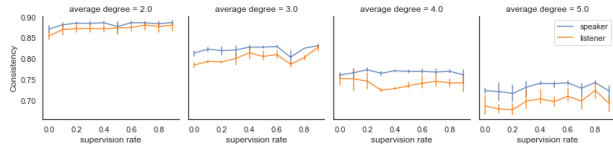

Figure 3. Speaker and listener consistency estimates with 95% Confidence Intervals and the average degree of agents in the random social network (Experiment 2)

communication systems according to all three communication asymmetry metrics (Fig. 2b and 2c). Small-world and ring social networks lead the populations to develop asymmetrical and local communication patterns. Moreover, supervised feedback helped the agents in all topology conditions to develop more shared and symmetric communication systems. We hypothesized that the effect of network type on asymmetry scores is driven by local connectivity of the populations: fully-connected and random networks do not form local communities, whereas the ring-shaped and small-world networks in our simulations mainly consisted of local connections. We tested this hypothesis in Experiment 3.

Lastly, agents in fully-connected and random networks produced less diversified actions, but their action and signaling distributions were much more coordinated than in the ring-shaped and small-world network structures (see Fig. 2). It suggests that agents in the last two conditions overfitted to the "action diversity" part of the reward function, while failing to coordinate using the communication channel.

3.2. Experiment 2: average degree

To test the effect of network's average degree on speaker and listener consistencies, we focused on the random (ER) network, which produces desired variation in the average degrees of the nodes, while keeping most of the other network properties constant. We varied the probability of connecting nodes in the network from 0.2 to 0.9, and supervision rate from 0 to 0.9 with a step of 0.1. This resulted in 80 conditions total, each of which was simulated 5 times to obtain consistency estimates with confidence intervals.

We found that the average degree negatively affected listener ($p < 0.001$) and speaker ($p < 0.001$) consistencies (Fig. 3). Consistency scores for the average degrees higher than 5.0 stayed similar to the estimate we obtained for the fully-connected topology in Experiment 1.

3.3. Experiment 3: global and local connections

We examined the effect of locally-connected groups in a social network on homogeneity and symmetry of the developed communication systems. For this, we tested MARLC in the small-world network with different probabilities of adding a "global" connection to the initial ring-shaped struc-

ture. We classify a link as "global" if it connects agents further than one link apart in the ring network. We varied both the probability of global connection and the supervision rate from 0 to 0.9 with a step of 0.1. Each combination of these two factors was simulated 5 times.

Global connection probability was inversely related to the signaling divergence ($p < 0.001$), between- ($p < 0.001$) and within-agent signal-action mapping divergence ($p < 0.001$) of the developed communication systems (Fig. 3). Higher proportion of global connections led the groups to converge on communication systems that are shared by all their agents. Supervised feedback also helped the agents to develop homogeneous and symmetric communication patterns. The cumulative effect of both high supervision rate and probability of global connections lead to communication systems that are shared by all the agents in a group.
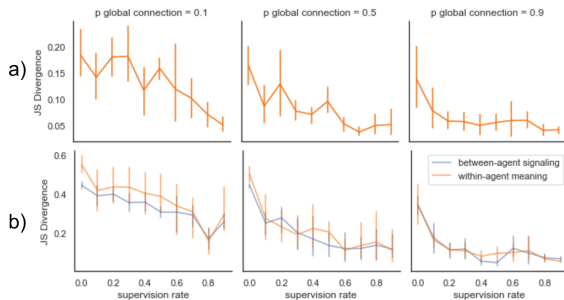

Figure 4. The Communication Analysis Metrics with 95% Confidence Intervals for the Experiment 3. a) Between-agent signal-action mapping divergence b) Signaling divergence (blue), within-agent signal-action mapping divergence (orange) and the probability of global connection in the small-world network

4. Discussion and Conclusion

We conducted three experiments to explore and isolate the factors of social network organization that drive the effectiveness, homogeneity, and symmetry of communication systems developed in MARLC settings. Following the suggestions of Lowe et al. (2019), we used a set of information-theoretic metrics to evaluate our results. This allowed us to determine which particular properties of communication systems are affected by our interventions.

The results of our first experiment partially replicated the effects of social network type on learning word conventions by human groups (Centola & Baronchelli, 2015). This suggests that the simple domain-general reinforcement learning model can capture core regularities found in human convergence on linguistic conventions. Our further experiments provide a more detailed account of particular social network properties that are responsible for these patterns in communication learning.

The second experiment demonstrated that the average degree of a social network is a central factor affecting how deterministically agents use the communication channel. The more communication partners an agent needs to adapt to, the more variable are its signaling and listening patterns. Key finding of the third experiment is that the proportion of global connections is the primary determining factor for three variables: homogeneity, within- and between-agent symmetry of the developed communication system. In particular, high proportion of local social connections led to the emergence of local communication patterns ("dialects") within the population, whereas adding more global connections forced the agents to find global consensus. Similarly, high local connectivity resulted in asymmetric communicative patterns, where even the same agent in two different roles uses completely different vocabularies.

At present, social network effects on MARLC have been largely under-investigated. To the best of our knowledge, there is only one work in this direction; L. Graesser and her colleagues (2019) looked at how communication systems developed within communities change if one community is introduced to another, depending on their inter- and intra-connectivity. Unfortunately, the insights from that work are only applicable for the specific scenario of community merging. In our study, we focused on the social network factors that are applicable to almost any MARLC setting.

The results of our analysis corroborate Lowe et al.'s (2019) suggestion that overall scores do not reflect whether the agents use the communication channel to solve the task, and how efficiently they do so (see Fig. 2). For example, the performance scores of the agents learning to coordinate in a ring-shaped social network were the highest among all the network types. However, more detailed analysis revealed that agents in this condition developed many local communication patterns, which varied across agents and their roles. This, again, illustrates the importance of thorough analysis of communication protocols learned in MARLC settings.

Our approach has a number of limitations that are important to mention. Firstly, while we aimed to test a broad spectrum of social network factors, our experiments are by no means comprehensive. There are other important aspects of social network structure that may play a key role in determining the properties of emerging communication protocols. We believe that studying the effects of variance of network's degree distribution and its modularity on MARLC is especially promising. Secondly, we used a very simplified setting of vanilla Q-learning in an "amodal" coordination game to minimize the number of assumptions that might make our results less representative to the MARLC problem in general. We suggest testing these results on more advanced reinforcement learning models and realistically perceptually grounded game settings.

5. Data Availability

The code and data for this work are available online at the project's github repository.

6. Acknowledgements

References

Baronchelli, A. The emergence of consensus: a primer. *Royal Society open science*, 5(2):172189, 2018.

Bernstein, D. S., Givan, R., Immerman, N., and Zilberstein, S. The complexity of decentralized control of markov decision processes. *Mathematics of operations research*, 27(4):819–840, 2002.

Centola, D. and Baronchelli, A. The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences*, 112(7):1989–1994, 2015.

Christiansen, M. H. and Chater, N. Language as shaped by the brain. *Behavioral and brain sciences*, 31(5):489–509, 2008.

Dubova, M. and Moskvichev, A. Effects of supervision, population size, and self-play on multi-agent reinforcement learning to communicate. *Artificial Life Conference Proceedings*, 32:678–686, 2020.

Erdos, P. On random graphs. *Publicationes mathematicae*, 6:290–297, 1959.

Fang, C., Lee, J., and Schilling, M. A. Balancing exploration and exploitation through structural design: The isolation of subgroups and organizational learning. *Organization Science*, 21(3):625–642, 2010.

Foerster, J., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pp. 122–130. International Foundation for Autonomous Agents and Multiagent Systems, Jul 2018.

Foerster, J. N., Assael, Y. M., de Freitas, N., and Whiteson, S. Learning to communicate to solve riddles with deep distributed recurrent q-networks. *arXiv:1602.02672 [cs]*, Feb 2016. URL http://arxiv.org/abs/1602.02672. arXiv: 1602.02672.

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

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., and Levy, R. How efficiency shapes human language. *Trends in cognitive sciences*, 2019.

Laurent, G. J., Matignon, L., Fort-Piat, L., et al. The world of independent learners is not markovian. *International Journal of Knowledge-based and Intelligent Engineering Systems*, 15(1):55–64, 2011.

Lazer, D. and Friedman, A. The network structure of exploration and exploitation. *Administrative science quarterly*, 52(4):667–694, 2007.

Lewis, D. Languages and language. 1975.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, O. P., and Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, pp. 6379–6390, 2017.

Lowe, R., Foerster, J., Boureau, Y.-L., Pineau, J., and Dauphin, Y. On the pitfalls of measuring emergent communication. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 693–701. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

Mason, W. and Watts, D. J. Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, 109(3):764–769, 2012.

Mason, W. A., Jones, A., and Goldstone, R. L. Propagation of innovations in networked groups. *Journal of Experimental Psychology: General*, 137(3):422, 2008.

Matignon, L., Laurent, G. J., and Le Fort-Piat, N. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *The Knowledge Engineering Review*, 27(1):1–31, 2012.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Newman, M. E. and Watts, D. J. Renormalization group analysis of the small-world network model. *Physics Letters A*, 263(4-6):341–346, 1999.

Pesce, E. and Montana, G. Improving coordination in small-scale multi-agent deep reinforcement learning through memory-driven communication. *Machine Learning*, Jan 2020. ISSN 1573-0565. doi: 10.1007/s10994-019-05864-5. URL https://doi.org/10.1007/s10994-019-05864-5.

Selten, R. and Warglien, M. The emergence of simple languages in an experimental coordination game. *Proceedings of the National Academy of Sciences*, 104(18): 7361–7366, 2007.

Steels, L. Language as a complex adaptive system. In *International Conference on Parallel Problem Solving from Nature*, pp. 17–26. Springer, 2000.

Sukhbaatar, S., Fergus, R., et al. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems*, pp. 2244–2252, 2016.

Wisdom, T. N. and Goldstone, R. L. Innovation, imitation, and problem solving in a networked group. *Nonlinear Dynamics-Psychology and Life Sciences*, 15(2):229, 2011.

"Five Graces Group", Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., et al. Language is a complex adaptive system: Position paper. *Language learning*, 59: 1–26, 2009.

A. Coordination game environment

Each game consists of two-step episodes. Every episode of the game involves two agents: a speaker and a listener who need to coordinate their actions.

On the first step, the speaker produces an action and a message. On the second step, the listener receives the message and outputs an action. If the actions match, agents receive a fixed "coordination" reward R_c. There are four unique actions and messages that the agents choose from.

In order to avoid trivial solutions when agents converge on a single action and ignore the communication channel, we introduced a penalty for repeating an action too often. Specifically, an additional reward of $min(0, 1/4 - \hat{p}_A)$ is added when an agent chooses action A. Here $\hat{p}_A$ denotes an empirical proportion of selecting action A during the last H steps. We used a fixed length history $H = 100$. Current $\hat{p}_A$ are added to the agent's input state spaces to aid convergence.

On every episode, each agent can be assigned to play either the speaker or the listener role. The current role type is provided as a binary input.

Overall, the agents receive, as inputs:

1. A binary role indicator

2. Random input (for speakers, in order to allow for non-deterministic policies) or speaker's message (for listeners)

3. Proportions of different actions in agent's recent history

The agents have two output layers both stemming from the last hidden layer:

1. Action output layer (for neurons, and the action is determined as argmax)

2. Message output layer which has the same structure as the action output layer. Message outputs are ignored if the agents plays the listener role

B. Agents

Agents are implemented using simple feed-forward neural networks with two hidden layers (hidden sizes of 25 and 15) and ReLU activation function.

We used vanilla deep Q-learning algorithm with additional supervised updates to train the agents. Supervised feedback corresponds to "peeking" the other agent's action and storing it in memory (with the signaling experience from that game) as a "correct" response. In this case, the action-signal mapping is not superimposed, and the "correct" answers in the supervising trials are bottom-up driven and reflect the dynamics of the agents themselves. Supervised feedback was implemented by changing a certain proportion of negative ("miscoordinated") experiences to the "supervising" ones. We decided to control for supervision in our experiments because partial supervising feedback in different forms is often available in naturalistic language learning situations. By amplifying the reinforcement signal, supervising feedback may help to overcome the "sparcity of rewards" problem in multi-agent reinforcement learning.

C. Metrics

We follow the same set of metrics as used in (Dubova & Moskvichev, 2020).

1. **Speaking Consistency and Listening Consistency**.

   This metric provides a quantitative measure of whether the actions that the agents perform are related to the

signals that they send or receive. We follow Lowe et al. (2019) who proposed to use the normalized mutual information between the distributions over messages and actions induced by an agent. The metric can be formally described as follows:

$$C = \sum_{a \in A_l} \sum_{m \in A_c} p_{a,m}(a,m) log \frac{p_{a,m}(a,m)}{p_a(a)p_m(m)} / Z \quad (1)$$

Here, Z is the average entropy of the two marginal distributions: $Z = \frac{(H(p_a)+H(p_m))}{2}$. A_l and A_c denote the set of available actions and the set of available messages respectively.

This metric is computed twice for every agent, conditioned on the role played by the agent (speaker or listener). We average these metrics across agents in every simulation to obtain the speaking and listening consistency metrics that we report.

2. **Communication Asymmetry metrics**:
   a. Between-agent signal-action mapping divergence. While the previous metric aimed to measure whether an agent is using communication channel (i.e. whether its actions correspond to the signals in any way), the second metric aims to measure whether the communication patterns differ between agents.

We compute the average Jensen-Shannon pairwise divergence between distributions of agents' actions following a specific signal (averaged over all signals and pairs of agents).

For a pair of agents, the metric is defined as follows

$$\sum_{m \in A_c} JSD(p_{a_1|m}, p_{a_2|m})/|A_c| \quad (2)$$

Here, $p_{(a_1|m)}$ are action distributions of agent 1 conditional on the message (received or sent) being equal to m.

b. Within-agent signal-action mapping divergence.

This metric aims to capture internal inconsistencies in agent's behaviors when it switches between roles: whether the agent's behaviors differ depending on whether it receive or send a specific message.

For that, we use the average Jensen-Shannon divergence between the distributions over agent's actions (conditioned on receiving or sending a specific message) when the agent plays the speaker and the listener role. Formally, we use the same definition as in Equation 2, but now $p_{a_1|m}$ and $p_{a_2|m}$ correspond to the same agent's distributions when this agent plays different roles (as opposed to distributions of a pair of

different agents playing the same role). We average the metric scores across all agents to obtain the final measure that we report.

c. Signaling divergence.

This metric aims to measure the difference in individual agents' messaging preferences. We define *signaling divergence* as an average pairwise Jensen-Shannon divergence of marginal signaling distributions of different agents.

3. **Behavioral Predictability**. This last metric is created to assess the general diversity in agent's actions. When the agents' actions are less diverse (and hence, more predictable), it is easier to achieve successful multi-agent coordination without using the communication channel. To look at whether the diversity of actions corresponds to the diversity of signals, we also compute the predictability of agents' signaling patterns. We define *behavioural action/message predictability* as Jensen-Shannon divergence between marginal distributions of agent's actions/messages and the uniform distribution.

Many of these metrics require knowing probability distributions. We estimate all such distributions empirically.

As a short summary, if we see the learned language as a simple probabilistic dictionary that maps messages to actions, the metrics can be summarized as follows (note that every agent defines two such dictionaries: one for the speaker and one for the listener role):

1. **Speaking Consistency and Listening Consistency**. Are the dictionaries reliable? I.e. if we look up a specific message, do we consistently get the same action, or is there a lot of randomness?

2. **Communication Asymmetry**:
   a. Between-agent signal-action mapping divergence. Are the dictionaries similar for different agents?

   b. Within-agent signal-action mapping divergence. How different are the "speaker" and "listener" dictionaries that each agent defines?

   c. Talking divergence. Do different agents show different patterns in their dictionary lookups?

3. **Behavioral Predictability**. How uniformly do agents look up different words in the dictionary (speaking predictability)? How uniform are the results of their queries (behavioral predictability)?

D. Statistical analysis

All hypotheses were tested using a linear regression model with robust covariance estimation, controlling for supervi-

sion rate. Excluding the supervision rate did not qualita-
tively change the results, however.