

# Emergence of Grounded Compositional Language in Multi-Agent Populations

Igor Mordatch<sup>1</sup> Pieter Abbeel<sup>1,2</sup>

## Abstract

By capturing statistical patterns in large corpora, machine learning has enabled significant advances in natural language processing, including in machine translation, question answering, and sentiment analysis. However, for agents to intelligently interact with humans, simply capturing the statistical patterns is insufficient. In this paper we investigate if, and how, grounded compositional language can emerge as a means to achieve goals in multi-agent populations. Towards this end, we propose a multi-agent learning environment and learning methods that bring about emergence of a basic compositional language. This language is represented as streams of abstract discrete symbols uttered by agents over time, but nonetheless has a coherent structure that possesses a defined vocabulary and syntax. We also observe emergence of non-verbal communication such as pointing and guiding when language communication is unavailable.

## 1. Introduction

Development of agents that are capable of communication and flexible language use is one of the long-standing challenges facing the field of artificial intelligence. Agents need to develop communication if they are to successfully coordinate as a collective. Furthermore, agents will need some language capacity if they are to interact and productively collaborate with humans or make decisions that are interpretable by humans. If such a capacity were to arise artificially, it could also offer important insights into questions surrounding development of human language and cognition.

But if we wish to arrive at formation of communication from first principles, it must form out of necessity. The approaches that learn to plausibly imitate language from examples of human language, while tremendously useful, do not learn *why* language exists. Such supervised approaches

can capture structural and statistical relationships in language, but they do not capture its functional aspects, or that language happens for purposes of successful coordination between humans. Evaluating success of such imitation-based approaches on the basis of linguistic plausibility also presents challenges of ambiguity and requirement of human involvement.

Recently there has been a surge of renewed interest in the pragmatic aspects of language use and it is also the focus of our work. We adopt a view of (Gauthier & Mordatch, 2016) that an agent possesses an understanding of language when it can use language (along with other tools such as non-verbal communication or physical acts) to accomplish goals in its environment. This leads to evaluation criteria that can be measured precisely and without human involvement.

In this paper, we propose a physically-situated multi-agent learning environment and learning methods that bring about emergence of a basic compositional language. This language is represented as streams of abstract discrete symbols uttered by agents over time, but nonetheless has a coherent structure that possesses a defined vocabulary and syntax. The agents utter communication symbols alongside performing actions in the physical environment to cooperatively accomplish goals defined by a joint reward function shared between all agents. There are no pre-designed meanings associated with the uttered symbols - the agents form concepts relevant to the task and environment and assign arbitrary symbols to communicate them.

There are similarly no explicit language usage goals, such as making correct utterances, and no explicit roles agents are assigned, such as speaker or listener, or explicit turn-taking dialogue structure as in traditional language games. There may be an arbitrary number of agents in a population communicating at the same time and part of the difficulty is learning to refer specific agents. A population of agents is situated as moving particles in a continuous two-dimensional environment, possessing properties such as color and shape. The goals of the population are based on non-linguistic objectives, such as moving to a location and language arises from the need to coordinate on those goals. We do not rely on any supervision such as human demonstrations or text corpora.

<sup>1</sup>OpenAI <sup>2</sup>UC Berkeley. Correspondence to: Igor Mordatch <mordatch@openai.com>.

Similar to recent work, we formulate the discovery the action and communication protocols for our agents jointly as a reinforcement learning problem. Agents perform physical actions and communication utterances according to an identical policy that is instantiated for all agents and fully determines the action and communication protocols. The policies are based on neural network models with an architecture composed of dynamically-instantiated recurrent modules. This allows decentralized execution with a variable number of agents and communication streams. The joint dynamics of all agents and environment, including discrete communication streams are fully-differentiable, the agents' policy is trained end-to-end with backpropagation through time.

The languages we observe forming exhibit interpretable compositional structure that in general assigns symbols to separately refer to environment landmarks, action verbs, and agents. However, environment variation leads to a number of specialized languages, omitting words that are clear from context. For example, when there is only one type of action to take or one landmark to go to, words for those concepts do not form in the language. Considerations of the physical environment also have an impact on language structure. For example, a symbol denoting *go* action is typically uttered first because the listener can start moving before even hearing the destination. This effect only arises when linguistic and physical behaviors are treated jointly and not in isolation.

The presence of a physical environment also allows for alternative strategies aside from language use to accomplish goals. A visual sensory modality provides an alternative medium for communication and we observe emergence of non-verbal communication such as pointing and guiding when language communication is unavailable. When even non-verbal communication is unavailable, strategies such as direct pushing may be employed to succeed at the task. It is important to us to build an environment with a diverse set of capabilities which language use develops alongside with.

Our work offers insights into why compositional structure emerges in our formed languages in the first place. In part, we find composition to emerge when we explicitly encourage active vocabulary sizes to be small through a soft penalty. This is consistent with analysis in evolutionary linguistics (Nowak et al., 2000) that finds composition to emerge only when number of concepts to be expressed becomes greater than a factor of agent's symbol vocabulary capacity. Another important component leading to composition is training on a variety of tasks and environment configurations simultaneously. Training on cases where most information is clear from context (such as when there is only one landmark) leads to formation of atomic concepts

that are reused compositionally in more complicated cases. Further investigation is required to determine whether this is an artifact specific to our training setup, or a more fundamental requirement for compositional syntax formation.

## 2. Related Work

Recent years have seen substantial progress in practical natural language applications such as machine translation (Sutskever et al., 2014; Bahdanau et al., 2014), sentiment analysis (Socher et al., 2013), document summarization (Durrett et al., 2016), and domain-specific dialogue (Dhingra et al., 2016). Much of this success is a result of intelligently designed statistical models trained on large static datasets. However, such approaches do not produce an understanding of language that can lead to productive cooperation with humans.

An interest in pragmatic view of language understanding has been longstanding (Austin, 1962; Grice, 1975) and has recently argued for in (Gauthier & Mordatch, 2016; Lake et al., 2016; Lazaridou et al., 2016b). Pragmatic language use has been proposed in the context of two-player reference games (Golland et al., 2010; Vogel et al., 2014; Andreas & Klein, 2016) focusing on the task of identifying object references through a learned language. (Winograd, 1973; Wang et al., 2016) ground language in a physical environment and focusing on language interaction with humans for completion of tasks in the physical environment. In such a pragmatic setting, language use for communication of spatial concepts has received particular attention in (Steels, 1995; Ullman et al., 2016).

Aside from producing agents that can interact with humans through language, research in pragmatic language understanding can be informative to the fields of linguistics and cognitive science. Of particular interest in these fields has been the question of how syntax and compositional structure in language emerged, and why it is largely unique to human languages (Kirby, 1999; Nowak et al., 2000; Steels, 2005). Models such as Rational Speech Acts (Frank & Goodman, 2012) and Iterated Learning (Kirby et al., 2014) have been popular in cognitive science and evolutionary linguistics, but such approaches tend to rely on pre-specified procedures or models that limit their generality.

The recent work that is most similar to ours is the application of reinforcement learning approaches towards the purposes of learning a communication protocol, as exemplified by (Foerster et al., 2016; Sukhbaatar et al., 2016; Lazaridou et al., 2016a).

### 3. Problem Formulation

The setting we are considering is a cooperative partially observable Markov game (Littman, 1994), which is a multi-agent extension of a Markov decision process. A Markov game for  $N$  agents is defined by set of states  $\mathcal{S}$  describing the possible configurations of all agents, a set of actions  $\mathcal{A}_1, \dots, \mathcal{A}_N$  and a set of observations  $\mathcal{O}_1, \dots, \mathcal{O}_N$  for each agent. Initial states are determined by a distribution  $\rho : \mathcal{S} \mapsto [0, 1]$ . State transitions are determined by a function  $\mathcal{T} : \mathcal{S} \times \mathcal{A}_1 \times \dots \times \mathcal{A}_N \mapsto \mathcal{S}$ . For each agent  $i$ , rewards are given by function  $r_i : \mathcal{S} \times \mathcal{A}_i \mapsto \mathbb{R}$ , observations are given by function  $\mathbf{o}_i : \mathcal{S} \mapsto \mathcal{O}_i$ . To choose actions, each agent  $i$  uses a stochastic policy  $\pi_i : \mathcal{O}_i \times \mathcal{A}_i \mapsto [0, 1]$ .

In this work, we assume all agents have identical action and observation spaces, and all agents act according to the same policy  $\pi$ . We consider a finite horizon setting, with episode length  $T$ . In a cooperative setting, the problem is to find a policy that maximizes the expected shared return for all agents:

$$\max_{\pi} R(\pi), \quad \text{where} \quad (1)$$

$$R(\pi) = \mathbb{E} \left[ \sum_{t=0}^T \sum_{i=0}^N r(\mathbf{s}_i^t, \mathbf{a}_i^t) \right] \quad (2)$$

The cooperative setting allows us to pose the problem as a joint minimization across all agents, as opposed to minimization-maximization problems resulting from competitive settings.

### 4. Grounded Communication Environment

As argued in the introduction, grounding multi-agent communication in a physical environment is crucial for interesting communication behaviors to emerge. In this work, we consider a physically-simulated two-dimensional environment in continuous space and discrete time. This environment consists of  $N$  agents and  $M$  landmarks. Both agent and landmark entities inhabit a physical location in space  $\mathbf{p}$  and possess descriptive physical characteristics, such as color and shape type. In addition, agents can direct their gaze to a location  $\mathbf{v}$ . See Figure 1 for an example of environments we consider. Agents can act to move in the environment and direct their gaze, but may also be affected by physical interactions with other agents. We denote the physical state of an entity (including descriptive characteristics) by  $\mathbf{x}$  and describe its precise details and transition dynamics in the Appendix.

In addition to performing physical actions, agents utter verbal communication symbols  $c$  at every timestep. These utterances are discrete elements of an abstract symbol vocabulary  $\mathcal{C}$  of size  $K$ . We do not assign any significance or meaning to these symbols. They are treated as abstract

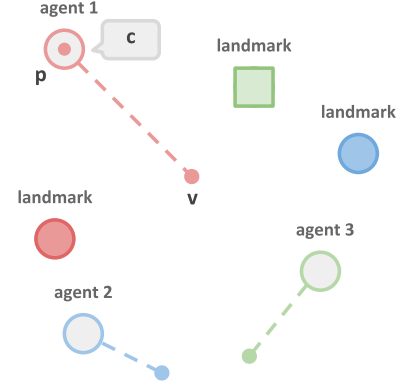


Figure 1. An example environment we consider, consisting of 3 agents and 3 landmarks of random shapes and colors. Every agent and landmark is located at  $\mathbf{p}$  and an agent’s gaze is directed at  $\mathbf{v}$ .

categorical variables that are emitted by each agent and observed by all other agents. It is up to agents at training time to assign meaning to these symbols. As shown in Section 7, these symbols become assigned to interpretable concepts. Agents may also choose not to utter anything at a given timestep, and there is a cost to making an utterance, loosely representing the metabolic effort of vocalization. We denote a vector representing one-hot encoding of symbol  $c$  with boldface  $\mathbf{c}$ .

Each agent has internal goals specified by vector  $\mathbf{g}$  that are private and not observed by other agents. These goals are grounded in the physical environment and include tasks such as moving to or gazing at a location. These goals may involve other agents (requiring the other agent to move to a location, for example) but are not observed by them and thus necessitate coordination and communication between agents. Verbal utterances are one tool which the agents can use to cooperatively accomplish all goals, but we also observe emergent use of non-verbal signals and altogether non-communicative strategies.

To aid in accomplishing goals, each agent has internal recurrent memory bank  $\mathbf{m}$  that is also private and not observed by other agents. This memory bank has no pre-designed behavior and it is up to the agents to learn to utilize it appropriately.

The full state of the environment is given by

$$\mathbf{s} = [\mathbf{x}_{1,\dots,(N+M)} \ \mathbf{c}_{1,\dots,N} \ \mathbf{m}_{1,\dots,N} \ \mathbf{g}_{1,\dots,N}] \in \mathcal{S}$$

Each agent observes physical states of all entities in the environment, verbal utterances of all agents, and its own private memory and goal vector. The observation for agent  $i$  is

$$\mathbf{o}_i(\mathbf{s}) = [{}_i\mathbf{x}_{1,\dots,(N+M)} \ \mathbf{c}_{1,\dots,N} \ \mathbf{m}_i \ \mathbf{g}_i]$$

Where  $x_j$  is the observation of entity  $j$ 's physical state in agent  $i$ 's reference frame. More intricate observation models are possible, such as physical observations solely from pixels or verbal observations from a single input channel. These models would require agents learning to perform visual processing and source separation, which are orthogonal to this work. Despite the dimensionality of observations varying with the number of physical entities and communication streams, our policy architecture as described in Section 5.2 allows a single parameterization across these variations.

## 5. Policy Learning with Backpropagation

Each agent acts by sampling actions from a stochastic policy  $\pi$ , which is identical for all agents and defined by parameters  $\theta$ . There are several common options for finding optimal policy parameters. The model-free framework of Q-learning can be used to find the optimal state-action value function, and employ a policy that acts greedily to according to the value function. Unfortunately, Q function dimensionality scales quadratically with communication vocabulary size, which can quickly become intractably large. Alternatively it is possible to directly learn a policy function using model-free policy gradient methods, which use sampling to estimate the gradient of policy return  $\frac{dR}{d\theta}$ . The gradient estimates from these methods can exhibit very high variance and credit assignment becomes an especially difficult problem in the presence of sequential communication actions.

Instead of using model-free reinforcement learning methods, we build an end-to-end differentiable model of all agent and environment state dynamics over time and calculate  $\frac{dR}{d\theta}$  with backpropagation. At every optimization iteration, we sample a new batch of 1024 random environment instantiations and backpropagate their dynamics through time to calculate the total return gradient. Figure 2 shows the dependency chain between two timesteps. A similar approach was employed by (Foerster et al., 2016; Sukhbaatar et al., 2016) to compute gradients for communication actions, although the latter still employed model-free methods for physical action computation.

The physical state dynamics, including discontinuous contact events can be made differentiable with smoothing. However, communication actions require emission of discrete symbols, which present difficulties for backpropagation.

### 5.1. Discrete Communication and Gumbel-Softmax Estimator

In order to use categorical communication emissions  $c$  in our setting, it must be possible to differentiate through

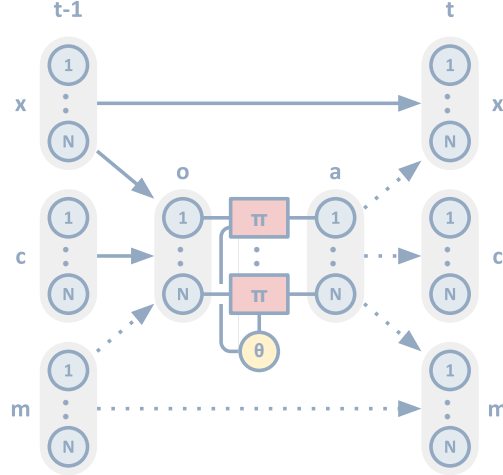


Figure 2. The transition dynamics of  $N$  agents from time  $t - 1$  to  $t$ . Dashed lines indicate one-to-one dependencies between agents and solid lines indicate all-to-all dependencies.

them. There has been a wealth of work in machine learning on differentiable models with discrete variables, but we found recent approach in (Jang et al., 2016; Maddison et al., 2016) to be particularly effective in our setting. The approach proposes a Gumbel-Softmax distribution, which is a continuous relaxation of a discrete categorical distribution. Given  $K$ -categorical distribution parameters  $p$ , a differentiable  $K$ -dimensional one-hot encoding sample  $G$  from the Gumbel-Softmax distribution can be calculated as:

$$G(\log p)_k = \frac{\exp((\log p_k + \epsilon)/\tau)}{\sum_{j=0}^K \exp((\log p_j + \epsilon)/\tau)}$$

Where  $\epsilon$  are i.i.d. samples from Gumbel(0, 1) distribution,

$$\epsilon = -\log(-\log(u)), \quad u \sim \mathcal{U}[0, 1]$$

And where  $\tau$  is a softmax temperature parameter. We did not find it necessary to anneal the temperature and set it to 1 in all our experiments for training and sample directly from the categorical distribution at test time. To emit a communication symbol, our policy is trained to directly output  $\log p \in \mathcal{R}^K$ , which is transformed to a symbol emission sample  $c \sim G(\log p)$ . The resulting gradient can be estimated as  $\frac{dc}{d\theta} = \frac{dG}{dp} \frac{dp}{d\theta}$ .

### 5.2. Policy Architecture

The policy class we consider in this work are stochastic neural networks. The policy outputs samples of an agent's physical actions  $u$ , communication symbol utterance  $c$ , and internal memory updates  $\Delta m$ . The policy must consolidate multiple incoming communication symbol streams emitted by other agents, as well as incoming observations of physical entities. Importantly, the number of agents (and thus the

number of communication streams) and number of physical entities can vary between environment instantiations. To support this, the policy instantiates a collection of identical processing modules for each communication stream and each observed physical entity. Each processing module is a fully-connected multi-layer perceptron. The weights between all communication processing modules are shared (and similarly for all physical observation modules). The outputs of individual processing modules are pooled with a softmax operation into feature vectors  $\phi_c$  and  $\phi_x$  for communication and physical observation streams, respectively. Such weight sharing and pooling makes it possible to apply the same policy parameters to any number of communication and physical observations. We found that feeding  $\phi_c$  as an additional input to each physical observation processing module can help training, but is not critical and is omitted for simplicity.

The pooled features and agent’s private goal vector are passed to the final processing module that outputs distribution parameters  $[\psi_u \ \psi_c]$  from which action samples are generated as  $\mathbf{u} = \psi_u + \varepsilon$  and  $\mathbf{c} \sim G(\psi_c)$ , where  $\varepsilon$  is a zero-mean Gaussian noise.

Unlike communication games where agents only emit a single utterance, our agents continually emit a stream of symbols over time. Thus processing modules that read and write communication utterance streams benefit greatly from recurrent memory that can capture meaning of a stream over time. To this end, we augment each communication processing and output module with an independent internal memory state  $\mathbf{m}$ , and each module outputs memory state updates  $\Delta\mathbf{m}$ . In this work we use simple additive memory updates for simplicity and interpretability, but other memory architectures such as LSTMs can be used.

$$\mathbf{m}^t = \tanh(\mathbf{m}^{t-1} + \Delta\mathbf{m}^{t-1} + \varepsilon)$$

We build all fully-connected modules with 256 hidden units and 2 layers each in all our experiments, using exponential-linear units and dropout with a rate of 0.1 between all hidden layers. Size is feature vectors  $\phi$  is 256 and size of each memory module is 32. The overall policy architecture is shown in Figure 3.

### 5.3. Auxiliary Prediction Reward

To help policy training avoid local minima, we found it helpful to include auxiliary goal prediction tasks, similar to recent work in reinforcement learning (Dosovitskiy & Koltun, 2016; Silver et al., 2016). In agent  $i$ ’s policy, each communication processing module  $j$  additionally outputs a prediction  $\hat{\mathbf{g}}_{i,j}$  of agent  $j$ ’s goals. At the end of the episode, we add a reward for predicting other agent’s goals, which in turn encourages communication utterances that convey the agent’s goals clearly to other agents. Across all agents

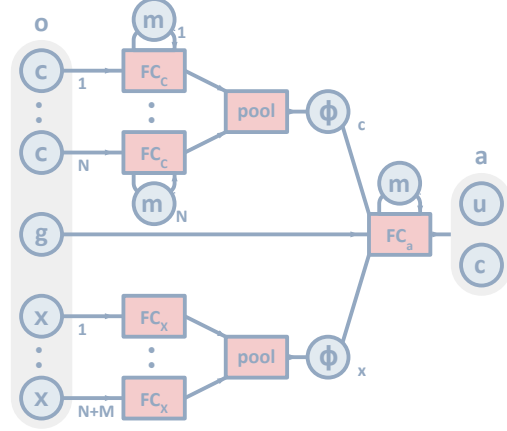


Figure 3. Overview of our policy architecture, mapping observations to actions at every point time. *FC* indicates a fully-connected processing module that shares weights with all others of its label. *pool* indicates a softmax pooling layer.

this reward has the form

$$r_g = - \sum_{\{i,j|i \neq j\}} \|\hat{\mathbf{g}}_{i,j}^T - \mathbf{g}_j^T\|^2$$

Note that we do not use  $\hat{\mathbf{g}}$  as an input in calculating actions. It is only used for the purposes of auxiliary prediction task.

## 6. Compositionality and Vocabulary Size

What leads to compositional syntax formation? One known constructive hypothesis requires modeling the process of language transmission and acquisition from one generation of agents to the next iteratively as in (Kirby et al., 2014). In such iterated learning setting, compositionality emerges due to poverty of stimulus - one generation will only observe a limited number of symbol utterances from the previous generation and must infer meaning of unseen symbols. This approach requires modeling language acquisition between agents, but when implemented with pre-designed rules was shown over multiple iterations between generations to lead to formation of a compositional vocabulary.

Alternatively, (Nowak et al., 2000) observed that emergence of compositionality requires the number of concepts describable by a language to be above a factor of vocabulary size. In our preliminary environments the number of concepts to communicate is still fairly small and is within the capacity of a non-compositional language. We use a maximum vocabulary size  $K = 20$  in all our experiments. We tested a smaller maximum vocabulary size, but found that policy optimization became stuck in a poor local minima where concepts became conflated. Instead, we propose to use a large vocabulary size limit but use a soft penalty

function to prevent the formation of unnecessarily large vocabularies. This allows the intermediate stages of policy optimization to explore large vocabularies, but then converge on an appropriate active vocabulary size. As shown in Figure 6, this is indeed what happens.

How do we penalize large vocabulary sizes? (Nowak et al., 2000) proposed a word population dynamics model that defines reproductive ratios of words to be proportional to their frequency, making already popular words more likely to survive. Inspired by these rich-get-richer dynamics, we model the communication symbols as being generated from a Dirichlet Process (Teh, 2011). Each communication symbol has a probability of being symbol  $c_k$  as

$$p(c_k) = \frac{n_k}{\alpha + n - 1}$$

Where  $n_k$  is the number of times symbol  $c_k$  has been uttered and  $n$  is the total number of symbols uttered. These counts are accumulated over agents, timesteps, and batch entries.  $\alpha$  is a Dirichlet Process hyperparameter corresponding to the probability of observing an out-of-vocabulary word. The resulting reward across all agents is the log-likelihood of all communication utterances to independently have been generated by a Dirichlet Process:

$$r_c = \sum_{i,t,k} \mathbb{I}[\mathbf{c}_i^t = c_k] \log p(c_k)$$

Maximizing this reward leads to consolidation of symbols and the formation of compositionality. This approach is similar to encouraging code population sparsity in autoencoders (Ng, 2011), which was shown to give rise to compositional representations for images.

## 7. Experiments<sup>1</sup>

We experimentally investigate how variation in goals, environment configuration, and agents physical capabilities lead to different communication strategies.

In this work, we consider three types of actions an agent needs to perform, *go to* location, *look at* location, and *do nothing*. Goal for agent  $i$  consists of an action to perform, a location to perform it on  $\bar{\mathbf{r}}$ , and an agent  $r$  that should perform that action. These goal properties are accumulated into goal description vector  $\mathbf{g}$ . These goals are private to each agent, but may involve other agents. For example, agent  $i$  may want agent  $r$  to go to location  $\bar{\mathbf{r}}$ . This goal is not observed by agent  $r$ , and requires communication between agents  $i$  and  $r$ . The goals are assigned to agents such that no agent receives conflicting goals. We do however show generalization in the presence of conflicting goals in Section 7.3.

<sup>1</sup> Videos of our experimental results can be viewed at <https://sites.google.com/site/multiagentlanguage/>

Agents have different reference frames and can only communicate in discrete symbols, and so cannot directly send goal position vector. Even if they could, agents observe the environment in different reference frames and have no shared global positioning reference. What makes the task possible is that we place goal locations  $\bar{\mathbf{r}}$  on landmark locations of which are observed by all agents (in their independent reference frames). The strategy then is for agent  $i$  to unambiguously communicate landmark reference to agent  $r$ . Importantly, we do not provide explicit association between goal positions and landmark reference. It is up to the agents to learn to associate a position vector with a set of landmark properties and communicate them with discrete symbols.

In the results that follow, agents do not observe other agents. This disallows capacity for non-verbal communication, necessitating the use of language. In section 7.4 we report what happens when agents are able to observe each other and capacity for non-verbal communication is available.

Despite training with continuous relaxation of the categorical distribution, we observe very similar reward performance at test time. No communication is provided as a baseline (again, non-verbal communication is not possible). The no-communication strategy is for all agents go towards the centroid of all landmarks.

Condition	Train Reward	Test Reward
No Communication	-0.919	-0.920
Communication	-0.332	-0.392

Table 1. Training and test physical reward for setting with and without communication.

### 7.1. Syntactic Structure

We observe a compositional syntactic structure emerging in the stream of symbol uttered by agents. When trained on environments with only two agents, but multiple landmarks and actions, we observe symbols forming for each of the landmark colors and each of the action types. A typical conversation and physical agent configuration is shown in first row of Figure 4 and is as follows:

```
Green Agent: GOTO, GREEN, ...
Blue Agent: GOTO, BLUE, ...
```

The labels for abstract symbols are chosen by us purely for interpretability and visualization and carry no meaning for training.

Physical environment considerations play a part in the syntactic structure. The action type verb *GOTO* is uttered first because actions take time to accomplish in the grounded environment. When the agent receives *GOTO* symbol it starts moving toward the centroid of all the landmarks (to



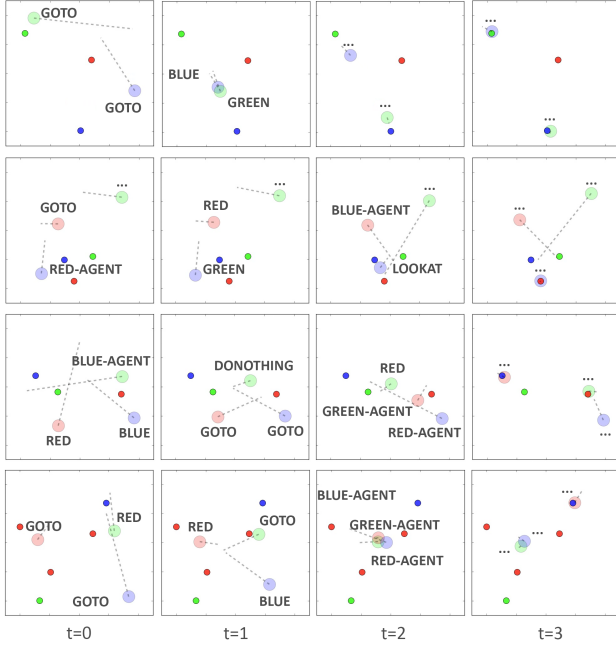


Figure 4. A collection of typical sequences of events in our environments shown over time. Each row is an independent trial. Large circles represent agents and small circles represent landmarks. Communication symbols are shown next to the agent making the utterance. The labels for abstract communication symbols are chosen purely for visualization.

be equidistant from all of them) and then moves towards the specific landmark when it receives its color identity.

When the environment configuration can contain more than three agents, agents need to form symbols for referring to each other. Three new symbols form to refer to agent colors that are separate in meaning from landmark colors. The typical conversations are shown in second and third rows of Figure 4.

Red Agent: GOTO, RED, BLUE-AGENT, ...  
 Green Agent: ..., ..., ..., ...  
 Blue Agent: RED-AGENT, GREEN, LOOKAT, ...

Agents may not omit any utterances when they are the subject of their private goal, in which case they have access to that information and have no need to announce it. In this language, there is no set ordering to word utterances. Each symbol contributes to sentence meaning independently, similar to case marking grammatical strategies used in many human languages (Beuls & Steels, 2013).

The agents largely settle on using a consistent set of symbols for each meaning, due to vocabulary size penalties and that discourage synonyms. We show the aggregate streams of communication utterances in Figure 5.

In simplified environment configurations when there is only one landmark or one type of action to take, no sym-

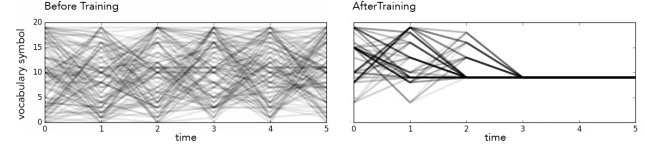


Figure 5. Communication symbol streams emitted by agents over time before and after training accumulated over 10 thousand test trials.

bols are formed to refer to those concepts because they are clear from context.

## 7.2. Symbol Vocabulary Usage

We find word activation counts to settle on the appropriate compositional word counts. That early during training large vocabulary sizes are being taken advantage of to explore the space of communication possibilities before settling on the appropriate effective vocabulary sizes as shown in Figure 6. In this figure,  $1 \times 1 \times 3$  case refers to environment with two agents and a single action, which requires only communicating one of three landmark identities.  $1 \times 2 \times 3$  contains two types of actions, and  $3 \times 3 \times 3$  case contains three agents that require explicit referencing.

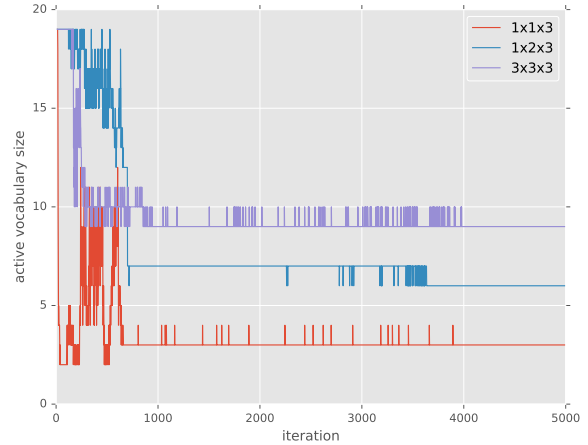


Figure 6. Word activations counts for different environment configurations over training iterations.

## 7.3. Generalization to Unseen Configurations

One of the advantages of decentralised execution policies is that trained agents can be placed into arbitrarily-sized groups and still function reasonably. When there are additional agents in the environment with the same color identity, all agents of the same color will perform the same task if they are being referred to. Additionally, when agents of a particular color are asked to perform two conflicting tasks (such as being asked go to two different landmarks by two

different agents), they will perform the average of the conflicting goals assigned to them. Such cases occur despite never having been seen during training.

Due to the modularized observation architecture, the number of landmarks in the environment can also vary between training and execution. The agents perform sensible behaviors with different numbers of landmarks, despite not being trained in such environments. For example, when there are distractor landmarks of novel colors, the agents never go towards them. When there are multiple landmarks of the same color, the agent communicating the goal still utters landmark color (because the goal is the position of one of the landmarks). However, the agents receiving the landmark color utterance go towards the centroid of all landmark of the same color, showing a very sensible generalization strategy. An example of such case is shown in fourth row of Figure 4.

#### 7.4. Non-verbal Communication and Other Strategies

The presence of a physical environment also allows for alternative strategies aside from language use to accomplish goals. In this set of experiments we enable agents to observe other agents' position and gaze location, and in turn disable communication capability via symbol utterances. When agents can observe each other's gaze, a pointing strategy forms where the agent can communicate a landmark location by gazing in its direction, which the recipient correctly interprets and moves towards. When gazes of other agents cannot be observed, we see behavior of goal sender agent moving towards the location assigned to goal recipient agent (despite receiving no explicit reward for doing so), in order to guide the goal recipient to that location. Lastly, when neither visual nor verbal observation is available on part of the goal recipient, we observe the behavior of goal sender directly pushing the recipient to the target location. Examples of such strategies are shown in Figure 7. It is important to us to build an environment with a diverse set of capabilities which language use develops alongside with.

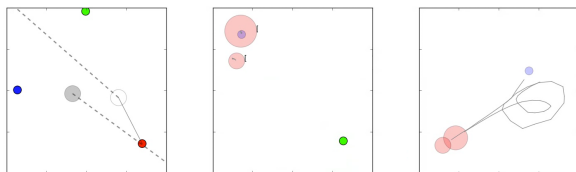


Figure 7. Examples of non-verbal communication strategies, such as pointing, guiding, and pushing.

## 8. Conclusion

We have presented a multi-agent environment and learning methods that brings about emergence of an abstract compositional language from grounded experience. This abstract language is formed without any exposure to human language use. We investigated how variation in environment configuration and physical capabilities of agents affect the communication strategies that arise.

In the future, we would like experiment with larger number of actions that necessitate more complex syntax and larger vocabularies. We would also like integrate exposure to human language to form communication strategies that are compatible with human use.

## References

- Andreas, Jacob and Klein, Dan. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 1173–1182, 2016.
- Austin, J.L. *How to Do Things with Words*. Oxford, 1962.
- Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- Beuls, Katrien and Steels, Luc. Agent-based models of strategies for the emergence and evolution of grammatical agreement. *PloS one*, 8(3):e58960, 2013.
- Dhingra, Bhuwan, Li, Lihong, Li, XiuJun, Gao, Jianfeng, Chen, Yun-Nung, Ahmed, Faisal, and Deng, Li. End-to-End Reinforcement Learning of Dialogue Agents for Information Access. *arXiv:1609.00777 [cs]*, September 2016. arXiv: 1609.00777.
- Dosovitskiy, Alexey and Koltun, Vladlen. Learning to act by predicting the future. *arXiv preprint arXiv:1611.01779*, 2016.
- Durrett, Greg, Berg-Kirkpatrick, Taylor, and Klein, Dan. Learning-based single-document summarization with compression and anaphoricity constraints. *arXiv preprint arXiv:1603.08887*, 2016.
- Foerster, Jakob N., Assael, Yannis M., de Freitas, Nando, and Whiteson, Shimon. Learning to Communicate with Deep Multi-Agent Reinforcement Learning. 2016.
- Frank, Michael C. and Goodman, Noah D. Predicting Pragmatic Reasoning in Language Games. *Science*, 336(6084):998, May 2012. ISSN 1095-9203. doi: 10.1126/science.1218633.



- Gauthier, Jon and Mordatch, Igor. A paradigm for situated and goal-driven language learning. *CoRR*, abs/1610.03585, 2016.
- Golland, Dave, Liang, Percy, and Klein, Dan. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pp. 410–419, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- Grice, H. P. Logic and conversation. In Cole, P. and Morgan, J. L. (eds.), *Syntax and Semantics: Vol. 3: Speech Acts*, pp. 41–58. Academic Press, San Diego, CA, 1975.
- Jang, E., Gu, S., and Poole, B. Categorical Reparameterization with Gumbel-Softmax. *ArXiv e-prints*, November 2016.
- Kirby, Simon. *Syntax out of Learning: the cultural evolution of structured communication in a population of induction algorithms*. 1999.
- Kirby, Simon, Griffiths, Tom, and Smith, Kenny. Iterated learning and the evolution of language. *Current opinion in neurobiology*, 28:108–114, 2014.
- Lake, Brenden M., Ullman, Tomer D., Tenenbaum, Joshua B., and Gershman, Samuel J. Building machines that learn and think like people. *CoRR*, abs/1604.00289, 2016.
- Lazaridou, Angeliki, Peysakhovich, Alexander, and Baroni, Marco. Multi-agent cooperation and the emergence of (natural) language. *arXiv preprint arXiv:1612.07182*, 2016a.
- Lazaridou, Angeliki, Pham, Nghia The, and Baroni, Marco. Towards Multi-Agent Communication-Based Language Learning. May 2016b. *arXiv*: 1605.07133.
- Littman, Michael L. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the eleventh international conference on machine learning*, volume 157, pp. 157–163, 1994.
- Maddison, Chris J., Mnih, Andriy, and Teh, Yee Whye. The concrete distribution: A continuous relaxation of discrete random variables. *CoRR*, abs/1611.00712, 2016.
- Ng, Andrew. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.
- Nowak, Martin A., Plotkin, Joshua B., and Jansen, Vincent A. A. The evolution of syntactic communication. *Nature*, 404(6777):495–498, March 2000. doi: 10.1038/35006635.
- Silver, David, van Hasselt, Hado, Hessel, Matteo, Schaul, Tom, Guez, Arthur, Harley, Tim, Dulac-Arnold, Gabriel, Reichert, David, Rabinowitz, Neil, Barreto, Andre, et al. The predictron: End-to-end learning and planning. *arXiv preprint arXiv:1612.08810*, 2016.
- Socher, Richard, Perelygin, Alex, Wu, Jean Y, Chuang, Jason, Manning, Christopher D, Ng, Andrew Y, Potts, Christopher, et al. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, pp. 1642. Citeseer, 2013.
- Steels, Luc. A self-organizing spatial vocabulary. *Artif. Life*, 2(3):319–332, January 1995. ISSN 1064-5462. doi: 10.1162/artl.1995.2.319.
- Steels, Luc. What triggers the emergence of grammar? In *AISB'05: Proceedings of the Second International Symposium on the Emergence and Evolution of Linguistic Communication (EELC'05)*, pp. 143–150. University of Hertfordshire, 2005.
- Sukhbaatar, Sainbayar, Szlam, Arthur, and Fergus, Rob. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 2244–2252, 2016.
- Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc V. Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 3104–3112. Curran Associates, Inc., 2014.
- Teh, Yee Whye. Dirichlet process. In *Encyclopedia of machine learning*, pp. 280–287. Springer, 2011.
- Ullman, Tomer, Xu, Yang, and Goodman, Noah. The pragmatics of spatial language. In *Proceedings of the Cognitive Science Society*, 2016.
- Vogel, Adam, Gómez Emilsson, Andrés, Frank, Michael C., Jurafsky, Dan, and Potts, Christopher. Learning to reason pragmatically with cognitive limitations. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, pp. 3055–3060, Wheat Ridge, CO, July 2014. Cognitive Science Society.
- Wang, S. I., Liang, P., and Manning, C. Learning language games through interaction. In *Association for Computational Linguistics (ACL)*, 2016.
- Winograd, Terry. A procedural model of language understanding. 1973.

## Appendix

### 8.1. Physical State and Dynamics

The physical state of the agent is specified by  $\mathbf{x} = [\mathbf{p} \ \dot{\mathbf{p}} \ \mathbf{v} \ \mathbf{d}]$  where  $\dot{\mathbf{p}}$  is the velocity of  $\mathbf{p}$ .  $\mathbf{d} \in \mathcal{R}^3$  is the color associated with the agent. Landmarks have similar state, but without gaze and velocity components. The physical state transition dynamics for a single agent are given by:

$$\mathbf{x}_i^t = \begin{bmatrix} \mathbf{p} \\ \dot{\mathbf{p}} \\ \mathbf{v} \end{bmatrix}_i^t = \begin{bmatrix} \mathbf{p} + \dot{\mathbf{p}}\Delta t \\ \gamma\dot{\mathbf{p}} + (\mathbf{u}_p + \mathbf{f}(\mathbf{x}_1, \dots, \mathbf{x}_N))\Delta t \\ \mathbf{u}_v \end{bmatrix}_i^{t-1}$$

Where  $\mathbf{f}(\mathbf{x}_1, \dots, \mathbf{x}_N)$  are the physical interaction forces (such as collision) between all agents in the environment and any obstacles,  $\Delta t$  is the simulation timestep (we use 0.1), and  $(1 - \gamma)$  is a damping coefficient (we use 0.5).

### 8.2. Goal Specification and Reward

Goal for agent  $i$  consists of an action to perform, a location to perform it on  $\bar{\mathbf{r}}$ , and an agent  $r$  that should perform that action. Action type  $\in \{\text{go-to, look-at, do-nothing}\}$  is encoded as a one-hot vector  $\mathbf{g}^{\text{type}} \in \mathcal{R}^3$ . These goal properties are accumulated into goal description vector  $\mathbf{g}$ . The physical reward associated with goal  $\mathbf{g}_i$  for agent  $i$  at time  $t$  is:

$$r_i^t = -\left( \begin{bmatrix} \|\mathbf{p}_r^t - \bar{\mathbf{r}}\|^2 \\ \|\mathbf{v}_r^t - \bar{\mathbf{r}}\|^2 \\ 0 \end{bmatrix} \right)^T \mathbf{g}^{\text{type}} + \|\mathbf{u}_i^t\|^2 + \|\mathbf{c}_i^t\|^2$$

The total return for the episode is

$$R = r_c + r_g + \sum_t \sum_i r_i^t$$