Learning to Refer: How Scene Complexity Affects Emergent Communication in Neural Agents

Dominik Künkele[1] **and Simon Dobnik**[1,2]
Department of Philosophy, Linguistics and Theory of Science[1]
Centre for Linguistic Theory and Studies in Probability (CLASP)[2]
University of Gothenburg, Sweden
contact@dominik-kuenkele.de and simon.dobnik@gu.se

Abstract

We explore how neural network-based agents learn to map continuous sensory input to discrete linguistic symbols through interactive language games. One agent describes objects in 3D scenes using invented vocabulary; the other interprets references based on attributes. We extend the CLEVR dataset with more complex scenes to study how increased referential complexity impacts language acquisition and symbol grounding in artificial agents.

1 Introduction

We investigate a core challenge in artificial intelligence and cognitive science: how systems can bridge the gap between **continuous sensory input** (like vision) and **discrete symbolic communication** (like language) known as the **symbol grounding problem** (Harnad, 1990). It refers to the difficulty of connecting abstract symbols to real-world referents in a meaningful way, especially in artificial systems where symbols must acquire meaning through interaction rather than pre-programmed associations. We study symbol grounding through **generation and interpretation of referring expressions** which require a system to map visual attributes (like color, shape, and size) to symbolic representations that can be communicated and understood by another agent. We explore how neural agents can develop such referential abilities through **language games**—interactive scenarios where communication protocols emerge from repeated coordination attempts in interaction—by exchanging discrete messages toi solve a visual diuscrimiantion task.

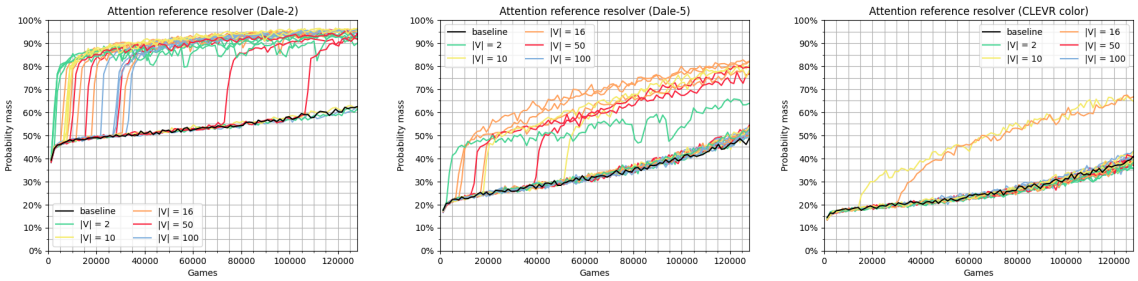
In this setup one can study the proprties of **artifical languages** the agents develop and whether these resemble human languages (Bartlett and Kazakov, 2005; Kirby et al., 2008; Steels and Loetzsch, 2009; Kharitonov et al., 2019; Lazaridou et al., 2017). However, our focus here is investigation of **conditions and protocols that lead learning successful interaction**. These include different configurations and complexities of discriminating features between the target object and distractors and between different scenes as well different configurations of grounded langauge models. This gives us important insights what neural models like these are capable of learning in inetractive scenarios with natural, human language.

2 Dataset

Our dataset conists of images of contrasting scenes and objects. The scenes are generated from an adapted code that was used to generate the **CLEVR dataset**. Instead of focusing on compositional properties of descriptions, we generate scenes with increasing complexity and control over object attributes, inspired by (Dale and Reiter, 1995), but we used the **feature hirerachy** to generate visual scenes rather than generate descriptions. We create the following datasets:

In **CLEVR color**, the target object is uniquely identifiable by **color alone**. All distractors share the same shape and size as the target. This allows the study of how agents learn to use a single attribute for reference. **Dale-2** includes **one target and one distractor**. The target is uniquely identifiable by a minimal combination of attributes (color, shape, size). This setup introduces more variability and requires agents to learn which attributes are most informative in each context. **Dale-5** increases complexity by including **one target and four distractors**. The target may share multiple attributes with different distractors, requiring more complex referring expressions. This setup closely mirrors real-world scenarios where objects often share overlapping features. Each dataset contains 10,000 images, with up to 10 non-overlapping objects per image. The images are 480×320 pixels, and objects are placed to ensure visibility and spatial separation.

Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue, September 3–5, 2025, Bielefeld, Germany.



(a) 'Dale-2' dataset (b) 'Dale-5' dataset (c) 'CLEVR color' dataset

Figure 1: Learning curves of all language games on each dataset. The colors correspond to different vocabulary sizes $|V|$. The baseline is marked in black.

Images are processed using a **ResNet-101** model, followed by two convolutional layers with ReLU activations. These layers reduce the feature maps to 128 channels.

3 Experiments

The games are set up through the **EGG framework** (Kharitonov et al., 2019) that allows communication through a discrete channel with an LSTM. Backpropagation is enabled through **Gumbel-Softmax relaxation**.

The receiver's task is to predict a 3×3 region around the target object in a 14×14 grid over the image (see Appendix A). The model outputs a probability distribution over all regions, and performance is measured by the **probability mass** assigned to the correct region. The sender encodes bounding boxes of all objects and passes them through an LSTM to generate a message. The receiver decodes the message and combines it with its own visual representation of the scene to predict the target region. The receiver does not have enough information to solve the task on its own. A total of 128.000 games are played. Furthermore, we allow different message lengths vocabulary sizes. All results are compared to a baseline in which the sender is generating random messages.

On the 'Dale-2' dataset, almost all configurations outperform the baseline, with top-configurations achieving over 96% probability mass (see Appendix B). Message length primarily influences performance, with $n \in \{3, 4\}$ yielding consistent results. While $n = 6$ configurations can succeed, they are less reliable. Vocabulary size shows less impact, though $|V| = 2$ performs slightly worse. No clear correlation between n and $|V|$ emerges. On the 'Dale-5' dataset, only 8 out of 30 configurations beat the baseline. Best models reach 84%, but many struggle due to the increased complexity. Shorter messages ($n \in \{2, 3\}$) and medium vocabularies ($|V| \in \{10, 16, 50\}$) are most effective. The increased number of distractors complicates the task: objects share more attributes, requiring more complex descriptions, and their spatial proximity can lead to confusion in region identification. Performance is weakest on the 'CLEVR color' dataset, with only two configurations beating the baseline (64-67%), both using medium message lengths ($n \in \{3, 4\}$) and vocabularies ($|V| \in \{10, 16\}$). Notably, short messages ($n = 2$) often mislead the receiver. The presence of up to 10 objects increases the likelihood of focusing on incorrect targets.

4 Findings and future directions

Our study shows (i) that **emergent communication is possible** and in the studied environments and network configurations but (ii) **scene complexity** significantly affects learning. More distractors and overlapping attributes make it harder for agents to learn effective communication. (iii) **Medium-sized vocabularies and message lengths** strike the best balance between expressiveness and learnability. (iv) **Attribute difficulty**: shape is easiest to learn, followed by color, with size being the hardest. This mirrors findings in human cognition and suggests that (v) **neural networks may benefit from similar inductive biases**. The findings suggest that that successful language-vision models must go beyond mere observation of pixels and words where such biases would be provided. They must incorporate **structured representations**, **attention mechanisms**, and **pragmatic reasoning** to handle real-world complexity.

References

Mark Bartlett and Dimitar Kazakov. 2005. The origins of syntax: from navigation to language. *Connection Science*, 17(3-4):271–288.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D 42: 335-346*.

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 2019. EGG: a toolkit for research on emergence of lanGuage in games. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 55–60, Hong Kong, China. Association for Computational Linguistics.

Simon Kirby, Hannah Cornish, and Kenny Smith. 2008. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations*.

Luc Steels and Martin Loetzsch. 2009. Perspective alignment in spatial language. In Kenny R. Coventry, Thora Tenbrink, and John A. Bateman, editors, *Spatial Language and Dialogue*, volume 3 of *Explorations in language and space*, pages 70–88. Oxford University Press.

A Architecture of the language game

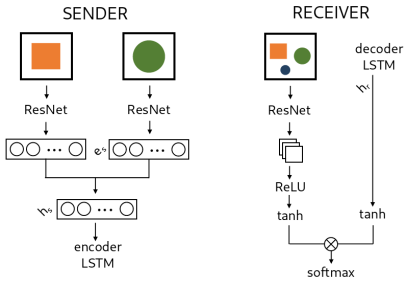


Figure 2: Simplified architecture of the attention predictor game.

B Results

		Dale-2	Dale-5	color
n	$\|V\|$	P **mass**	P **mass**	P **mass**
baseline		62,16%	49,61%	41,68%
2	2	92,27%	52,15%	33,64%
3	2	94,52%	51,97%	37,09%
4	2	89,15%	51,98%	39,68%
6	2	59,68%	53,57%	38,43%
2	10	**96,16%**	80,26%	36,53%
3	10	94,9%	53,47%	38,24%
2	16	95,84%	**84,03%**	39,65%
4	10	**96,08%**	48,03%	**64,31%**
3	16	94,59%	**81,46%**	**67,88%**
6	10	63,46%	**82,12%**	40,11%
4	16	94,14%	49,81%	40,84%
6	16	95,86%	50,71%	40,61%
2	50	93,78%	52,24%	39,56%
3	50	93,88%	79,65%	40,36%
2	100	92,43%	53,23%	37,68%
4	50	**96,24%**	48,79%	43,61%
3	100	95,25%	48,52%	42,55%
6	50	91,27%	52,55%	40,21%
4	100	95,55%	49,65%	42,85%
6	100	60,27%	46,92%	41,98%

Table 1: Probability masses of the attention reference resolver after 128.000 games: n are different maximum message lengths and $|V|$ are different vocabulary sizes. Results in red didn't pass the baseline. The results are sorted by the product of n and $|V|$ which corresponds to available space for the message. The best results are achieved with a medium-sized message space across all datasets.