Agents generalize to novel levels of abstraction by using adaptive linguistic strategies

Kristina Kobrock and **Xenia Ohmer** and **Elia Bruni** and **Nicole Gotzner**
Institute of Cognitive Science, Osnabrück University
Wachsbleiche 27, 49090 Osnabrück, Germany

Abstract

We study abstraction in an emergent communication paradigm. In emergent communication, two artificial neural network agents develop a language while solving a communicative task. In this study, the agents play a concept-level reference game. This means that the speaker agent has to describe a concept to a listener agent, who has to pick the correct target objects that satisfy the concept. Concepts consist of multiple objects and can be either more specific, i.e. the target objects share many attributes, or more generic, i.e. the target objects share fewer attributes. We tested two directions of zero-shot generalization to novel levels of abstraction: When generalizing from more generic to very specific concepts, agents utilized a compositional strategy. When generalizing from more specific to very generic concepts, agents utilized a more flexible linguistic strategy that involves reusing many messages from training. Our results provide evidence that neural network agents can learn robust concepts based on which they can generalize using adaptive linguistic strategies. We discuss how this research provides new hypotheses on abstraction and informs linguistic theories on efficient communication.

1 Introduction

One of the most fundamental goals of Artificial Intelligence (AI) and Natural Language Processing (NLP) research is to build models which can generalize well to unseen data. This is, after all, one of the crucial abilities observed in human intelligence. Abstraction has been argued to be a necessary first step towards achieving generalization (Yee, 2019). But there are also alternative views such as the exemplar-based model of categories where generalization is achieved without abstraction (Ambridge, 2020; Daelemans, 2008). We believe that understanding abstraction and how it interacts with generalization, is fundamental to building well-generalizing models.

Humans naturally use abstraction to solve complex tasks and to communicate about strategies and solutions. Well-designed AI and NLP systems cannot only benefit from good abstraction abilities in, for example, reasoning and solving complex tasks (e.g. Ho et al., 2019; Zheng et al., 2024), but interactive systems should also be able to deal with human language inputs which involve abstractions (e.g. Lachmy et al., 2022). Many researchers studying human abstraction argue for a role of language therein (see e.g., Yee, 2019; Sloutsky and Deng, 2019; Gentner and Asmuth, 2019; Lupyan and Lewis, 2019). The main idea of these accounts is that the lexicalization of concepts, i.e. having a label for a concept, helps to acquire and structure information we obtain about an entity and to observe commonalities within members of a concept in the first place. The role of language in abstraction can also be tested in computational systems. The goals of the current research are tounderstand how abstraction and generalization interact and ultimately to inform the improvement of AI and NLP systems towards achieving human-like abstraction and generalization abilities.

Starting from the assumption that language is useful for abstraction, we study abstraction in a communicative setting and investigate how abstraction is achieved with the help of linguistic strategies such as compositionality and the reuse of previously established messages. To gain insights into the principled mechanisms of abstraction and the role of language for abstraction, we use a language emergence scenario. In language emergence research, the idea is to define a set of assumptions and then observe how these assumptions change a language-like system that emerges during interaction (see e.g. Lazaridou et al., 2017; Galke et al., 2022; Rodríguez Luna et al., 2020; Chaabouni et al., 2020). This modeling framework is ideal for investigating whether human-like behavior and communication strategies can emerge even

Findings of the Association for Computational Linguistics: ACL 2025, pages 8685–8699
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

in a comparatively simple communicative setup between two artificial neural network agents.

In our model, two artificial neural network agents solve a reference game, where a speaker agent has to communicate a target concept to a listener agent who needs to select the correct target concept in a context. We operationalize a concept as a set of target objects following previous work (Kobrock et al., 2024a,b; Mu and Goodman, 2021). Our target concepts are designed in a hierarchical fashion, ranging from very specific concepts consisting of objects where all attributes (e.g., size, color and shape) are fixed to a certain value, e.g. 'small blue circle', to very generic concepts consisting of objects where only one attribute is fixed, e.g. 'circle'. We can study abstraction by making use of this abstraction hierarchy. Here, we are interested in a specific kind of abstraction, namely the zero-shot generalization to concepts at novel levels of the concept hierarchy, or, to concepts at *novel levels of abstraction* (following the terminology of seminal research from Cognitive Psychology by Rosch et al., 1976). We will not only look at the generalization performance of the trained models, but also at the linguistic strategies the agents employ. Specifically, we investigate the properties of the emergent protocol and the use of novel vs. established messages during abstraction.

While previous work in emergent communication has highlighted the role of compositionality for generalization (see e.g. Hazra et al., 2021; Kottur et al., 2017; Lazaridou et al., 2018), in our experiments we disentangle two directions of generalization and propose that they require different linguistic strategies. We find that agents use a compositional strategy only when generalizing to specific concepts, but not when generalizing to generic concepts. These results highlight that compositionality is not the only way to achieve generalization, which is in line with recent findings from Chaabouni et al. (2020) and Kharitonov and Baroni (2020).

2 Method

2.1 General Setup

We use an emergent communication paradigm (e.g. Lazaridou et al., 2018; Chaabouni et al., 2019) and build on the concept-level reference game developed in previous work (Mu and Goodman, 2021; Kobrock et al., 2024a). We train two artificial neural network agents, one speaker and one listener agent. Over several iterations, these agents develop a communication system by solving the following task: The speaker agent S has to communicate a concept, i.e. a set of target objects $T = \{t_1, ..., t_g\}$, to the listener agent L whose task is to identify the correct targets among a set of distractors $D = \{d_1, ..., d_g\}$. We call the set of target objects the *concept* and the set of distractor objects the *context*. The listener's task is to identify the target concept in a certain context given a message generated by the speaker. The message is a vector of symbols generated by the speaker neural network which does not have a pre-specified meaning. Rather, the meaning of a message emerges over several interactions between the agents and is defined by its usage (see e.g. Lazaridou et al., 2017). Concepts vary in specificity, ranging from specific, where all attributes are shared among the target objects, to generic, where only one attribute is shared among the targets. Contexts can range from being fine, where all but one attributes are shared between targets and distractors, to being coarse, where no attribute is shared between targets and distractors. Both agent networks are trained in a Reinforcement Learning paradigm with the Gumbel-Softmax relaxation (Jang et al., 2017) on a joint loss that depends on whether the listener correctly identifies the targets and distractors given the speaker-generated message.

2.2 Zero-shot Conditions and Hypotheses

We test the zero-shot generalization abilities of the trained networks in two conditions (see Figure 1): The first condition, "to specific", tests whether agents are able to generalize to the most specific concepts when having seen more generic concepts during training. In this condition, we expect the emerging communication system to encode more generic concepts (such as "blue" or "circle"). For a successful zero-shot generalization, these more generic concepts would need to be combined to describe a specific concept (such as "blue circle"). Here, agents will need to combine previously learned attributes compositionally to describe a more specific concept. The second condition, "to generic", tests whether agents are able to generalize to the most generic concepts when having seen more specific concepts during training. In this condition, we expect the emerging communication system to encode more specific concepts (such as "blue circle" or "orange circle"). For a successful zero-shot generalization, agents will need to abstract away from contextually irrelevant features and find the common attribute that all targets share (e.g. "circle").

2.3 Dataset

The agents are trained on six symbolic datasets developed in previous work (Kobrock et al., 2024a). These datasets contain all possible concepts, ranging from specific to generic, and contexts, ranging from fine to coarse, for a given number of attributes and values. For example, dataset D(3,4) contains all possible concepts and contexts given that objects in this dataset have three attributes and each attribute can take four different values. If we think of the three attributes as shape, color and size, an example for a specific concept would be "small blue circle" and an example for a generic concept would be "square". In a fine context, objects belonging to the concept "small blue circle" would need to be discriminated against objects that are also small and blue. In a coarse context, distractor objects do not share any attributes with the target concept. This also means that there are more possible contexts for specific concepts than for generic concepts and the datasets reflect this relationship. We use a scaling factor of 10 to construct the datasets, i.e. each concept is included in a dataset 10 times.[1] This ensures that the datasets contain enough training data.

For the zero-shot dataset generation, we manipulate the training, validation and test splits of the data. In the "to specific" condition, the test split contains all most specific concepts available, i.e. those where all attributes are shared among the targets. The training and validation splits are composed of the remaining concepts which are more generic with 75% of the data used for training and 25% of the data used for validation. In the "to generic" condition, the test split contains all most generic concepts available, i.e. those where only one attribute is shared among the targets. The training and validation sets contain the remaining more specific concepts with 75% of the data used for training and 25% of the data used for validation. Dataset sizes can be inspected in Tables 10 and 11 in Appendix D and are comparable between zero-shot conditions.

[1]We use this scaling factor only to construct the train and validation dataset splits. The zero-shot test is performed on a test split that contains the novel concepts only once.

2.4 Architecture and Training

A communication game between a speaker S and a listener L is defined as $G = (T^S, D^S, T^L, D^L)$, where $T^S = \{t_1^S, ..., t_g^S\}$ and $D^S = \{d_1^S, ..., d_g^S\}$ are the inputs to the speaker, i.e. sets of game size g targets and distractors, and T^L and D^L are the analogously defined inputs to the listener. For these inputs, $T^S \neq T^L$ and $D^S \neq D^L$ hold, i.e. the targets and distractors presented to the speaker differ from the targets and distractors presented to the listener to ensure communication of higher-level concepts (Mu and Goodman, 2021; Kobrock et al., 2024a). In each round of the game, S generates a message $m = (s_j)_{j \leq M}$, where s_j is a symbol from vocabulary V and M is the maximal message length[2], based on the inputs T^S and D^S. L in turn, receives m and an input $X^L = \{x_1^L, ..., x_i^L\}$, where $i = 2 \cdot g$ which contains the targets T^L and distractors D^L shuffled. L then predicts a label $y_i^L \in \{0, 1\}$ (0: distractor, 1: target) for each object x_i^L in its input (see e.g. Mu and Goodman, 2021; Kobrock et al., 2024a; Ohmer et al., 2022). We visualize the setup in Figure 2.

For the implementation[3], we use the EGG framework for emergent communication games (Kharitonov et al., 2019, MIT license). Both agents are implemented in a similar fashion: Feed-forward layers with 64 units serve as embedding layers for the input objects. The speaker targets and distractors are embedded separately and then concatenated into a joint embedding. The listener input objects are processed by just one embedding layer. For message encoding and decoding, both speaker and listener networks use single-layer Gated Recurrent Units (GRU, Cho et al., 2014) with a hidden layer size of 128 that can deal with sequential inputs of varying lengths. A speaker-listener pair is trained with binary cross entropy loss

$$\mathcal{L}_{BCE}(S, L, G) = -\sum_i \log p^L(y_i^L | x_i^L, \hat{m}), \quad (1)$$

where $\hat{m} \sim p^S(m | T^S, D^S)$ and $p^L(y_i^L | x_i^L, \hat{m}) = \text{ReLU}(\text{GRU}^L(\hat{m}) \cdot \text{embed}(x_i^L))$ maximizing the probability that the listener correctly identifies targets and distractors with a label $y_i \in \{0, 1\}$ (0: distractor, 1: target) for each object x_i. To ensure differentiability for backpropagation, we use the

[2]The end-of-sequence symbol 0 can be used to terminate a message before M is reached.

[3]All code and analysis scripts are available at https://github.com/kristinakobrock/zero-shot-abstraction.

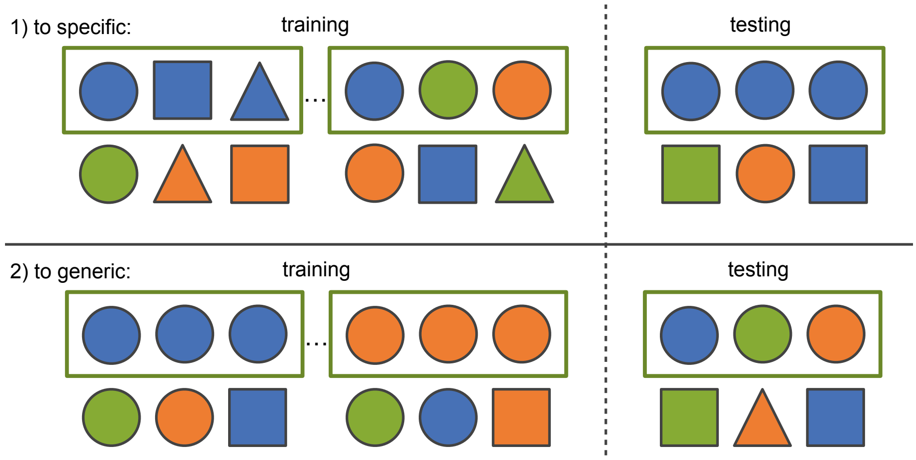

Figure 1: Examples for speaker inputs for training and testing in the two zero-shot test conditions “to specific” and “to generic”. Each input consists of targets (i.e., concepts) in the green bounding box and distractors (i.e., context).

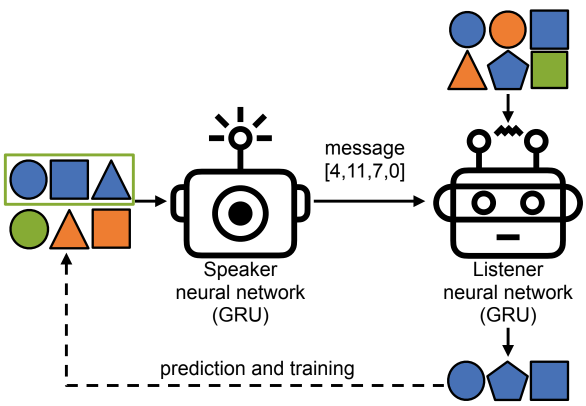

Figure 2: Architecture: Speaker and listener neural networks receive separate inputs where target objects satisfy the same target concept (here “blue”) and distractor objects (i.e., the context) share the same number of attributes with the target concept (here 0). They are trained on successful communication, i.e. when the listener identifies the correct target objects.

straight-through Gumbel-Softmax trick (Jang et al., 2017) with temperature $\tau = 2$ and a decay rate of 0.99. These and other hyperparameters were determined in a grid search that we conducted for all parameters over the different dataset sizes aiming for maximal validation accuracy. We train with batch size 32 and learning rate 0.001. For our simulations, we use game size 10, i.e. 10 target objects form a concept and 10 distractor objects form the context. The maximum message length M is defined as the total number of attributes in a dataset plus the End of Sequence (EOS) symbol 0. The vocabulary size for each dataset corresponds to the total number of attribute values present. We establish a minimal vocabulary size for each dataset as the sum of the number of attribute values plus one additional symbol. This minimal vocabulary size is then scaled by a factor of $f = 3$, as suggested by Ohmer et al. (2022) to ensure a sufficiently large communication channel (Chaabouni et al., 2020).

3 Results

We trained the models on six symbolic datasets with varying numbers of attributes and values. In a dataset $D(n, k)$, objects have n attributes which each can take k different values. For all metrics, we report means and standard deviations over five individual runs per dataset.

3.1 Generalization Performance

We evaluate the agents' performance on the test datasets to assess their zero-shot generalization abilities.[4] Accuracies are calculated as a percentage over the objects that the listener classifies as targets or distractors. An accuracy of 0.9 means that 90% of the objects, i.e. 18 objects with a game size of 10, have been classified correctly as targets or distractors. Or, in other words, two objects have been misclassified.

Table 1 summarizes the mean test accuracies over the five runs conducted on each dataset for both conditions. All zero-shot test accuracies are >=0.63 indicating that the listeners correctly identify more than 60% of the 20 objects as targets or

[4]Training and validation accuracies for both conditions are >=0.97 indicating that the agents have learned the task and achieved high performance on both the training and the validation data splits - a necessary prerequisite for a valid interpretation of the zero-shot test accuracies (see Tables 4 and 5 in Appendix A).

	to specific	to generic
D(3,4)	0.92 ± 0.02	0.71 ± 0.04
D(3,8)	0.85 ± 0.01	0.68 ± 0.07
D(3,16)	0.82 ± 0.03	0.63 ± 0.03
D(4,4)	0.95 ± 0.00	0.82 ± 0.02
D(4,8)	0.95 ± 0.01	0.82 ± 0.07
D(5,4)	0.96 ± 0.01	0.84 ± 0.06

Table 1: Zero-shot test accuracies for both conditions.

distractors. This corresponds to a number of 12 correctly identified objects. Agents achieve higher performance in the "to specific" condition compared to the "to generic" condition in all datasets. Comparing test accuracies between datasets, generalization performance is better for datasets with more attributes. Specifically, on datasets with at least four attributes, agents achieve generalization accuracies of 0.82 or higher in both conditions. This means that speakers choose expressions to describe the held-out concepts at novel levels of abstraction that enable listeners to classify at least 16 of the 20 objects correctly.

3.2 Concept reference

We investigate the emergent mappings between concepts and messages during training with the Normalized Mutual Information (NMI) score calculated over messages M and concepts C:

$$\text{NMI}(C, M) = \frac{H(M) - H(M|C)}{0.5 \cdot (H(C) + H(M))}, \qquad (2)$$

The NMI score is maximal (i.e., 1.0) if for all messages and concepts seen during training, every message maps to exactly one concept and vice versa. In other words, a maximal score indicates that the agents developed a protocol that includes only one-to-one mappings between messages and concepts, i.e. no ambiguity. We expect high but not maximal NMI scores which would indicate that the agents have learned a structured but not unambiguous mapping between concepts and messages. The mean NMI scores calculated for messages and concepts during training in five runs range between 0.84 and 0.95 in the "to specific" condition, i.e. when trained on more generic concepts, and between 0.77 and 0.87 in the "to generic" condition, i.e. when trained on more specific concepts (see Table 2). This indicates that a structured communication protocol has emerged in both conditions, while more ambiguity arises when training the agents on more specific concepts in the "to generic" condition.

	to specific	to generic
D(3,4)	0.93 ± 0.03	0.87 ± 0.04
D(3,8)	0.95 ± 0.01	0.82 ± 0.02
D(3,16)	0.87 ± 0.01	0.77 ± 0.02
D(4,4)	0.94 ± 0.01	0.87 ± 0.05
D(4,8)	0.84 ± 0.03	0.83 ± 0.03
D(5,4)	0.87 ± 0.02	0.83 ± 0.04

Table 2: NMI scores for both conditions.

3.3 Generalization strategies

When agents generalize to novel concepts in the zero-shot test, there are two conceivable strategies. Firstly, agents might reuse messages that have been successfully used during training also on the test dataset. Secondly, agents might invent novel messages to describe the novel concepts in the test dataset. We define reuse rates and novelty rates, respectively, to investigate the use of these strategies in our simulations. Next, we lay out our predictions for the reuse and novelty rates of agents trained in the "to specific" and "to generic" conditions.

In the "to specific" condition, we expect a high novelty rate and a lower reuse rate. However, we hypothesize that this strategy can only be effective if the agents use compositionality to combine learned meanings into novel messages.[5] An alternative strategy would be that the agents mainly reuse messages that have been uttered during training and do not produce many novel messages (i.e., high reuse rate and low novelty rate). But we hypothesize that this strategy is not the most prevalent strategy because it leads to the production of underinformative messages in certain contexts, i.e. messages that do not provide enough information for the listener to unambiguously identify the target concept (e.g. Engelhardt et al., 2006; Deutsch and Pechmann, 1982; Grice, 1975). For example, a message that has been used to refer to the concept "blue circle" during training might be used to refer to a "small blue circle" during testing. In some contexts, e.g. when all blue circles are small blue circles, this is efficient. In other contexts, however, e.g. when small blue circles need to be discriminated from large blue circles, this strategy is underinformative and not effective.

[5]See section 3.4 for an investigation of the messages' compositionality.

In the “to generic” condition, we expect a low novelty rate and a higher reuse rate. If the speaker agents produce novel messages to refer to novel concepts, they might come up with a highly efficient mapping, but they also run into the risk that the listener might not work out what the novel message refers to. This is due to the fact that the agents cannot draw on a compositional strategy when communicating only a single relevant attribute. However, if the speaker agents reuse messages from training, this will result in overinformative messages, i.e. messages that provide more information than necessarily required for the listener to unambiguously identify the target concept in certain contexts, e.g. when producing “small blue circle” in reference to a CIRCLE that needs to be discriminated against other shapes (e.g. Grice, 1975; Degen et al., 2020; Rubio-Fernandez, 2021). In other contexts, however, reused messages are highly efficient: As concepts are presented in a variety of contexts during training, there are communicative situations (namely coarse contexts) in which the speaker agent can choose to communicate only a single relevant attribute and rely on context to resolve ambiguity. This might lead to the emergence of messages that encode the meaning of generic concepts such as CIRCLE already during training, even though they are never explicitly presented. Reusing these messages during testing will thus be highly efficient.

To test these predictions, we look at the messages generated during testing on the novel concepts. First, we define the set of test concepts C_{test} and the set of test messages M_{test}. These are the messages produced and the concepts described during interactions on the test data split. Next, we define a message-concept ratio as the ratio between the number of test messages and the number of test concepts M_{test}/C_{test}. The resulting ratio is 1.0 if the number of messages is equal to the number of novel concepts, or, in other words, if for each novel concept, the agents produce one message during testing. Scores lower than 1.0 indicate that the agents produce fewer distinct messages than there are novel concepts. We calculate the ratios to ensure comparability between the “to specific” and “to generic” conditions because the test sets contain different amounts of novel concepts (see Tables 6 and 7 in Appendix B). We define the reuse and novelty rates by looking at the overlap between messages used during training and validation $M_{trainval}$ and messages used when generalizing to the test data split M_{test}.[6] We define novel messages as those messages that have been produced during testing but have not been produced during training and validation, i.e. the set difference $M_{test} - M_{trainval}$. We calculate the novelty rate as the ratio between the number of novel messages and the total number of unique messages used during testing $|M_{test} - M_{trainval}|/|M_{test}|$. We define reused messages as those messages that have been used in training and validation and then reused in testing, i.e. the intersection of the two sets of messages $M_{trainval} \cap M_{test}$ and calculate the reuse rate $|M_{trainval} \cap M_{test}|/|M_{test}|$. If reuse rate and novelty rate are balanced, this means that agents invent equally many new messages as they reuse old messages from training. If the percentages shift to one or the other extreme, this means that agents reuse more old messages than they invent new ones or vice versa.

In the “to specific” condition, we find that ratios between distinct messages and novel concepts in the test set range between 0.23 and 0.92 (see Table 3). The dataset with the highest ratio close to 1.0 is D(4,4) with a score of 0.92, where almost for each novel concept, a distinct message is produced. Strikingly, there are many datasets with a low message-concept ratio, which suggests that one message is used to refer to many concepts. As test accuracies are generally high (see Table 1), this does not impact generalization performance, but rather reflects the emergence of a very efficient language, where many concepts can be described with a small set of messages. As accuracies are not maximal, though, some objects are being misclassified by the listener. This suggests that this small set of messages can be underinformative. One reason for high message-concept ratios that we will test later in section 3.4 is that the emerging language might be very structured. A structured language allows the agents to use previously established meanings and combine them in a compositional fashion to novel meanings. As the novel meanings are generated on the fly, the agents might vary which previously established meanings they use, how they combine them and in which order, leading to a large amount of novel messages. Indeed, when looking at the reuse and novelty rates, we find that agents trained in the “to specific” condition invent at least 34% new messages during testing (see Table 3). For half of the datasets, the novelty rate even

[6] Both sets contain only unique message counts.

exceeds the reuse rate, suggesting that the agents come up with more new messages than they reuse old messages.

In the "to generic" condition, we observe message-concept ratios of distinct test messages to novel concepts that are very close to 1.0, indicating that the agents produce almost exactly one message for each novel concept. These agents also reuse more messages than invent new ones (see Table 3). There are two kinds of messages that can be reused from training with high communicative success: The first kind of messages have encoded all relevant attributes of a target concept during training. These are necessarily overinformative when produced during testing, but might still lead to a high number of objects being classified correctly by the listener. The second kind of messages have not encoded all relevant attributes during training, but were underinformative during training and required the context to resolve ambiguity (see Kobrock et al., 2024a). These messages would be just on the appropriate level of reference during testing. We have seen in section 3.2 that such ambiguous messages emerge in the "to generic" condition. This might explain the highly efficient reuse of messages in the "to generic" condition.

3.4 Compositionality

In the previous section, we hypothesized that agents in the "to specific" condition use a compositional strategy, whereas agents trained in the "to generic" condition do not rely on composition but rather reuse messages from training to describe novel concepts. To test the compositionality in the emerging languages, we use topographic similarity (also called "topographic ρ", Brighton and Kirby, 2006). The idea behind this metric is that emerging languages should exhibit structure. Specifically, regarding the mapping between meanings (in our case concepts) and messages, messages which are highly similar to each other should refer to concepts which are also highly similar to each other. This relationship can be measured with the topographic similarity metric (e.g., Ohmer et al., 2022; Lazaridou et al., 2018; Brighton and Kirby, 2006; Mu and Goodman, 2021). We calculate topographic similarity between messages and concepts by first calculating two distance vectors: one containing the pairwise Hausdorff distances between concepts and one containing the pairwise Edit, specifically Levenshtein, distances between messages (as in Mu and Goodman, 2021). Then we correlate these two distance vectors by using Spearman correlation to obtain the topographic similarity score between 0 and 1.0. The higher the score, the more compositional are the messages.

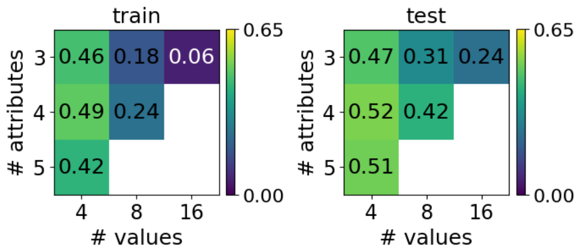

Figure 3: **To specific:** Topographic similarity scores calculated on messages from the train and test splits.

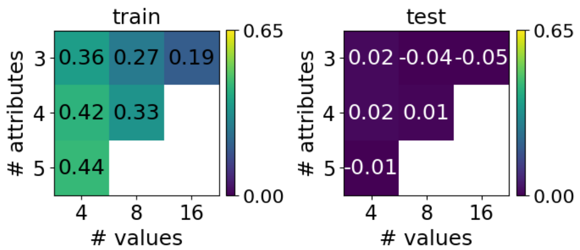

Figure 4: **To generic:** Topographic similarity scores calculated on messages from the train and test splits.

We find that the mean compositionality scores over five runs for concepts and messages seen during training range between 0.06 and 0.49 for the "to specific" condition, and that they range between 0.19 and 0.44 for the "to generic" condition. When looking at the topographic similarity scores calculated on the sets of messages and concepts from the test data split, we see a diverging picture: For the "to specific" condition, compositionality scores are higher during testing, suggesting that the agents use a highly compositional strategy when describing very specific concepts. In Appendix C, we present a qualitative analysis of the messages that shows that agents use established symbol-attribute mappings and compositionally combine these symbols into novel messages. For the "to generic" condition, on the other hand, we observe a drop of compositionality scores almost towards zero. This means that agents do not use a compositional strategy when being tested on the most generic concepts.

4 Discussion

In this study, we set out to investigate the linguistic strategies agents use to generalize to concepts at

	to specific			to generic		
	M_{test}/C_{test}	reuse rate	novelty rate	M_{test}/C_{test}	reuse rate	novelty rate
D(3,4)	0.84 ± 0.06	0.54 ± 0.07	0.46 ± 0.07	0.98 ± 0.03	0.80 ± 0.19	0.20 ± 0.19
D(3,8)	0.68 ± 0.06	0.45 ± 0.07	0.55 ± 0.07	0.93 ± 0.02	0.83 ± 0.07	0.17 ± 0.07
D(3,16)	0.23 ± 0.03	0.45 ± 0.07	0.55 ± 0.07	0.93 ± 0.02	0.75 ± 0.06	0.25 ± 0.06
D(4,4)	0.92 ± 0.04	0.44 ± 0.08	0.56 ± 0.08	0.97 ± 0.05	0.74 ± 0.14	0.26 ± 0.14
D(4,8)	0.58 ± 0.11	0.66 ± 0.12	0.34 ± 0.12	0.90 ± 0.10	0.93 ± 0.06	0.07 ± 0.06
D(5,4)	0.89 ± 0.03	0.65 ± 0.09	0.35 ± 0.09	0.98 ± 0.02	1.00 ± 0.00	0.00 ± 0.00

Table 3: The ratio of messages and concepts for the test data split M_{test}/C_{test}. Percentage of reused and novel messages from the total set of unique messages M_{test}.

novel levels of abstraction. The main interest of this research lies in investigating which strategies emerge as a function of the two different zero-shot test conditions in an interactive agent-based model and compare them to human-like communication strategies. Our results should be interpreted under this perspective, i.e. all differences we observe are emergent features of the model. We interpret the results on the agent-level and treat the differences between conditions as differences in linguistic strategies the agents are using. Because we use the exact same modeling setup in both conditions, the differences we observe have to be due to differences between conditions. Our main finding is that the abstraction abilities and linguistic strategies differ depending on the direction of the zero-shot generalization: When the agents generalize to very specific concepts, they make use of a compositional strategy and invent novel messages by combining established symbols in new ways. When they generalize to very generic concepts, the agents' strategy is mostly characterized by reusing already established messages that might have been ambiguous during training and are sufficiently informative during testing.

We observe lower performance when testing zero-shot generalization to very generic concepts than zero-shot generalization to very specific concepts. One explanation for this is that agents in the "to specific" condition make use of compositionality as their main strategy for generalization as shown by the compositionality scores and the qualitative analysis of the messages. Intuitively, combining learned symbols for attributes like "circle" or "blue" into a message "blue circle" is easier to achieve with a finite lexicon than abstracting to novel generic concepts. If a language does not encode specific concepts, novel meanings can always be generated by combining established meanings. However, if a language does not encode generic concepts, a compositional strategy is not an option. Novel meanings, however, cannot be established in a zero-shot generalization, so the only chance the agents have is to recur to already established meanings. But previously learned words that encode specific meanings are only useful to a certain extent, making the "to generic" direction of zero-shot generalization a harder task for our trained agent models which is reflected in the accuracies.

This idea is supported by evidence we obtained from an in-depth analysis of the emerging protocols. First, we have shown that NMI scores are high but not maximal in both conditions, suggesting a large number of one-to-one mappings in the emergent mapping between concepts and messages. We have observed that more ambiguity emerges when training the agents in the "to generic" condition. This ambiguity likely is what enables the agents to perform fairly well on the generalization task, where we observe that they mostly reuse messages from the training phase to refer to novel concepts. Ambiguity and the use of messages that rely on context to resolve ambiguities that remain after interpreting a message can lead to an efficient strategy that agents rely on mainly in coarse context conditions (see Kobrock et al., 2024a).

Second, we have shown that agents employ different strategies for generalizing to the most specific than to the most generic concepts. Specifically, agents come up with more novel messages that are produced by compositionally combining symbols that have been associated with a fixed concept attribute during training, when being tested on the most specific concepts. Agents trained in the "to generic" condition, on the other hand, mostly reuse entire messages the meanings of which have been established during training. These agents thus make use of the ambiguity of concept-message mappings that has emerged during training, accepting that these messages may be overinformative when describing the most generic concepts. Our findings are in line with results from Chaabouni et al. (2020) who found that compositionality is a sufficient but not a necessary condition for generalization.

A fruitful direction for research building on our results would be to further investigate the non-compositional strategy in abstraction to more generic concepts. This might be done by relating our results in the "to generic" condition to the phenomenon of overgeneralization in children acquiring language. Overgeneralization, or overextension, happens when a child uses a familiar label, for example "boot" to refer to an unfamiliar object, like SANDAL (see e.g. Gelman et al., 1998; Rescorla, 1980; Ferreira Pinto and Yang, 2021). This is similar to what our agents do when they use familiar messages to refer to novel concepts in the "to generic" condition. Future research could benefit from integrating both lines of research to develop new hypotheses on children's acquisition of concepts and overgeneralization as a pragmatic strategy for successful and efficient communication even if the correct label is not known (see e.g. Gershkoff-Stowe et al., 2006).

Another direction for future research is to investigate the role of communicative pressures during the emergent communication in our setting. Recent research in the field is dedicated to understanding better how efficient emergent communication systems emerge as a function of informativeness and utility, and highlights the role of communicative pressures for the emergence of an efficient solution that generalizes well (e.g. Gualdoni and Boleda, 2024; Tucker et al., 2022b,a). In our study, we do not use any communicative pressures except for keeping the maximum message length quite small (corresponding to the number of attributes in a dataset). As suggested by one of our reviewers, it would be interesting to test the effect of communicative pressures, such as a cost on the message length or an informativeness pressure as in Tucker et al. (2022b), in our setup. For example, in the "to generic" condition, where we observe that agents mostly reuse messages from training that might be overinformative during testing on the most generic concepts, a communicative pressure might lead to shorter and less overinformative messages.

Related to both research directions outlined above, another fruitful avenue will be to investigate specifically the pragmatic processes involved in selecting an efficient message for an unfamiliar referent and investigate whether, for example, reasoning about the listener's likely interpretation of a message helps speakers to identify a well-suited message, improving the agents' performance in the zero-shot test (see Zarrieß and Schlangen, 2019, for a related approach).

In summary, we have shown that the successful linguistic strategy for generalization depends on whether agents have to generalize from generic (low information) concepts to specific (high information) concepts or vice versa. Our results add to existing evidence that language in the form of labels is important for abstraction in humans. Our findings go beyond this research indicating a role of compositional messages and novel vs. established labels depending on the abstraction process. This work has important ramifications for linguistic theories on the role of compositionality and ambiguity in efficient communication and generalization: While previous work has highlighted the role of compositionality in generalization, we do not find evidence that abstraction to more generic concepts benefits from a compositional strategy in the same way as generalization to more specific concepts. Instead, generalization to more generic concepts, i.e. abstraction, benefits from reusing ambiguous messages from the emerging protocol. These results in the "to generic" condition are in line with two linguistic phenomena: First, the widening of meanings in diachronic language change where previously specific meanings are used for more generic concepts (see e.g. Wood, 2009; Díaz-Vera, 2022), and second the overgeneralization of familiar labels to unfamiliar objects of the same category in language acquisition (see e.g. Gershkoff-Stowe et al., 2006; Gelman et al., 1998).

Limitations

We would like to discuss how our experimental set-up featuring two agents with fixed speaker and listener roles resembles and differs from human communication and language evolution. For example, as one reviewer pointed out, human languages typically evolve in larger populations and seminal research on modeling language evolution has also focused on agent populations (Steels, 2000). Furthermore, we model communication only in one direction, where one agent has the speaker role and one agent has the listener role throughout the entire experiment. On the other hand, we model many of the key characteristics of human language, such as the evolution through interaction in reference situations, and there are several reasons why we believe that the here presented set-up is the ideal testbed to investigate our research question. Our experimental setup follows a typical Gricean sender-receiver model of communication which is a model that also underlies much work in (experimental) Linguistics (see e.g. Winters et al., 2018 for a language evolution experiment with dyads of human participants who have fixed speaker/listener roles). For that reason, we believe this approach is appropriate to study human-like communication. While it might be interesting to develop a more human-like version of our setup, we do not believe that this is necessary in our current experiments for two main reasons: First, while authors in the emergent communication field have argued for the use of populations of agents (e.g. Chaabouni et al., 2022), they have not found an advantage of population size for generalization. Previous work has also found that introducing populations or flexible role agents (i.e. agents which sometimes have the speaker and sometimes the listener role), does not change the results of the dyadic setup with fixed roles (see Ohmer et al., 2022). Second, one of the main reasons brought forward for using populations of agents is that this might help to make emergent languages compositional and to show that compositionality might be needed for successful generalization (and in this sense make emergent languages more comparable to human language, see e.g. Galke et al., 2022). However, we show that compositionality already emerges in our dyadic setup and that compositionality aids in generalization to specific concepts. We believe that the comparatively simple setup is a strength of our research that highlights that certain features of human-like communication already emerge when building on only few central characteristics of human communication.

The experiments presented here have been conducted as a proof-of-concept on symbolic data. On the one hand, these datasets are an ideal testbed for our hypotheses because they have been designed and constructed specifically for the purpose of studying concepts at different levels of abstraction. Using symbolic data has the advantage of total control over the manipulation and data without noise. On the other hand, we acknowledge that this is also a crucial limitation of our work. Future research needs to show whether the linguistic strategies we identify and the differences between generalizing to more specific or more generic concepts via compositionality or abstraction hold also for more naturalistic data. A validation on a more natural dataset is planned for future work and is expected to improve the generalizability of these results.

Acknowledgments

We thank our anonymous ARR reviewers for their thoughtful reviews and suggestions that helped us to improve the clarity of our paper.

The simulations were run on a high-performance computing cluster funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 456666331. Kristina Kobrock was supported by the DFG-funded Research Training Group "Computational Cognition" (DFG-GRK 2340).

Author Contributions:
Kristina Kobrock: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Xenia Ohmer:** Conceptualization, Methodology, Writing - Review & Editing. **Elia Bruni:** Conceptualization, Methodology, Writing - Review & Editing, Supervision. **Nicole Gotzner:** Conceptualization, Methodology, Writing - Review & Editing, Supervision, Project Administration.

References

Ben Ambridge. 2020. Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, 40(5-6):509–559.

Henry Brighton and Simon Kirby. 2006. Understanding Linguistic Evolution by Visualizing the Emergence of Topographic Mappings. *Artificial Life*, 12(2):229–242.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. 2020. Compositionality and Generalization In Emergent Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4427–4442.

Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. 2019. Anti-efficient encoding in emergent communication. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv preprint*.

Walter Daelemans. 2008. A comparison of Analogical Modeling to Memory-Based Language Processing. In Royal Skousen, Deryle Lonsdale, and Dilworth B. Parkinson, editors, *Analogical Modeling: An exemplar-based approach to language*, pages 157–179. John Benjamins Publishing Company.

Judith Degen, Robert D. Hawkins, Caroline Graf, Elisa Kreiss, and Noah D. Goodman. 2020. When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review*, 127(4):591–621.

Werner Deutsch and Thomas Pechmann. 1982. Social interaction and the development of definite descriptions. *Cognition*, 11(2):159–184.

Javier E. Díaz-Vera. 2022. Soft hearts and hard souls: The multiple textures of Old English feelings and emotions. *Cognitive Linguistic Studies*, 9(1):128–151. Publisher: John Benjamins.

Paul E. Engelhardt, Karl G. D. Bailey, and Fernanda Ferreira. 2006. Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, 54(4):554–573.

Renato Ferreira Pinto and Xu Yang. 2021. A computational theory of child overextension. *Cognition*, 206(104472).

Lukas Galke, Yoav Ram, and Limor Raviv. 2022. Emergent Communication for Understanding Human Language Evolution: What's Missing? In *Emergent Communication Workshop at ICLR 2022*.

Susan A. Gelman, William Croft, Panfang Fu, Timothy Clausner, and Gail Gottfried. 1998. Why is a pomegranate an *apple* ? The role of shape, taxonomic relatedness, and prior lexical knowledge in children's overextensions of *apple* and *dog*. *Journal of Child Language*, 25(2):267–291.

Dedre Gentner and Jennifer Asmuth. 2019. Metaphoric extension, relational categories, and abstraction. *Language, Cognition and Neuroscience*, 34(10):1298–1307.

Lisa Gershkoff-Stowe, Brenda Connell, and Linda Smith. 2006. Priming overgeneralizations in two- and four-year-old children. *Journal of Child Language*, 33(3):461–486.

Paul Herbert Grice. 1975. Logic and Conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and semantics*, volume 3, speech acts, pages 41 – 58. NY: Academic Press.

Eleonora Gualdoni and Gemma Boleda. 2024. Why do objects have many names? A study on word informativeness in language use and lexical systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18150–18163, Miami, Florida, USA. Association for Computational Linguistics.

Rishi Hazra, Sonu Dixit, and Sayambhu Sen. 2021. Zero-Shot Generalization using Intrinsically Motivated Compositional Emergent Protocols. In *Visually Grounded Interaction and Language (ViGIL) Workshop at NAACL 2021*, Mexico City, Mexico.

Mark K Ho, David Abel, Thomas L Griffiths, and Michael L Littman. 2019. The value of abstraction. *Current Opinion in Behavioral Sciences*, 29:111–116.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations (ICML)*.

Eugene Kharitonov and Marco Baroni. 2020. Emergent Language Generalization and Acquisition Speed are not tied to Compositionality. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 11–15, Online. Association for Computational Linguistics.

Eugene Kharitonov, Rahma Chaabouni, Marco Baroni, and Diane Bouchacourt. 2019. EGG: A toolkit for research on emergence of language in games. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Proceedings of System Demonstrations*, pages 55–60.

Kristina Kobrock, Xenia Isabel Ohmer, Elia Bruni, and Nicole Gotzner. 2024a. Context Shapes Emergent Communication about Concepts at Different Levels of Abstraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3831–3848, Torino, Italia. ELRA and ICCL.

Kristina Kobrock, Charlotte Uhlemann, and Nicole Gotzner. 2024b. Superordinate referring expressions in abstraction: Introducing the concept-level reference game. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. 2017. Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2962–2967, Copenhagen, Denmark. Association for Computational Linguistics.

Royi Lachmy, Valentina Pyatkin, Avshalom Manevich, and Reut Tsarfaty. 2022. Draw Me a Flower: Processing and Grounding Abstraction in Natural Language.

Transactions of the Association for Computational Linguistics, 10:1341–1356.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. 2018. Emergence of linguistic communication from referential games with symbolic and pixel input. In *International Conference on Learning Representations (ICML)*.

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

Gary Lupyan and Molly Lewis. 2019. From words-as-mappings to words-as-cues: the role of language in semantic knowledge. *Language, Cognition and Neuroscience*, 34(10):1319–1337.

Jesse Mu and Noah Goodman. 2021. Emergent Communication of Generalizations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 17994–18007.

Xenia Ohmer, Marko Duda, and Elia Bruni. 2022. Emergence of Hierarchical Reference Systems in Multi-agent Communication. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5689–5706.

Leslie A. Rescorla. 1980. Overextension in early language development. *Journal of Child Language*, 7(2):321–335.

Diana Rodríguez Luna, Edoardo Maria Ponti, Dieuwke Hupkes, and Elia Bruni. 2020. Internal and external pressures on language emergence: least effort, object constancy and frequency. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4428–4437, Online. Association for Computational Linguistics.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439.

Paula Rubio-Fernandez. 2021. Color discriminability makes over-specification efficient: Theoretical analysis and empirical evidence. *Humanities and Social Sciences Communications*, 8(1):147.

Vladimir M. Sloutsky and Wei (Sophia) Deng. 2019. Categories, concepts, and conceptual development. *Language, Cognition and Neuroscience*, 34(10):1284–1297.

Luc Steels. 2000. Language as a Complex Adaptive System. In *Parallel Problem Solving from Nature PPSN VI*, Lecture Notes in Computer Science, pages 17–26, Berlin, Heidelberg. Springer.

Mycal Tucker, Roger Levy, Julie A. Shah, and Noga Zaslavsky. 2022a. Trading off Utility, Informativeness, and Complexity in Emergent Communication. *Advances in Neural Information Processing Systems*, 35:22214–22228.

Mycal Tucker, Roger P. Levy, Julie Shah, and Noga Zaslavsky. 2022b. Generalization and Translatability in Emergent Communication via Informational Constraints. In *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems (InfoCog @ NeurIPS 2022)*.

Tahir Wood. 2009. Abstraction and adherence in discourse processes. *Journal of Pragmatics*, 41(3):484–496.

Eiling Yee. 2019. Abstraction and concepts: when, how, where, what and why? *Language, Cognition and Neuroscience*, 34(10):1257–1265.

Sina Zarrieß and David Schlangen. 2019. Know What You Don't Know: Modeling a Pragmatic Speaker that Refers to Objects of Unknown Categories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 654–659, Florence, Italy. Association for Computational Linguistics.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*.

A Training and validation accuracies

We present the training and validation accuracies for the "to specific" and "to generic" conditions in Table 4 and Table 5, respectively. For the interpretation of the zero-shot test accuracies, it is important that we achieve high training and validation accuracies in both conditions.

	training	validation
D(3,4)	1.00 ± 0.00	0.98 ± 0.01
D(3,8)	0.99 ± 0.00	0.98 ± 0.01
D(3,16)	0.97 ± 0.00	0.96 ± 0.01
D(4,4)	1.00 ± 0.00	1.00 ± 0.00
D(4,8)	0.98 ± 0.01	0.97 ± 0.01
D(5,4)	0.99 ± 0.00	0.99 ± 0.00

Table 4: **To specific:** Training and validation accuracies.

B Number of test concepts and messages

The zero-shot test datasets differ between conditions in the number of concepts presented due to there being more specific concepts than generic concepts in our datasets. In Table 6 and Table 7, we present the numbers of concepts in the test data splits C_{test}, as well as the number of unique messages used during testing M_{test}, for the "to

	training	validation
D(3,4)	0.98 ± 0.01	0.97 ± 0.02
D(3,8)	0.97 ± 0.00	0.97 ± 0.00
D(3,16)	0.96 ± 0.01	0.96 ± 0.01
D(4,4)	0.99 ± 0.01	0.98 ± 0.01
D(4,8)	0.97 ± 0.01	0.97 ± 0.02
D(5,4)	0.98 ± 0.01	0.98 ± 0.01

Table 5: **To generic:** Training and validation accuracies.

specific" and "to generic" condition, respectively. From these two values, we calculate the message-concept ratio M_{test}/C_{test} reported in the main paper.

	C_{test}	M_{test}	M_{test}/C_{test}
D(3,4)	64	54.0 ± 3.6	0.84 ± 0.06
D(3,8)	512	348.8 ± 28.5	0.68 ± 0.06
D(3,16)	4096	948.6 ± 123.7	0.23 ± 0.03
D(4,4)	256	236.2 ± 9.6	0.92 ± 0.04
D(4,8)	4096	2380.8 ± 437.7	0.58 ± 0.11
D(5,4)	1024	911.0 ± 34.9	0.89 ± 0.03

Table 6: **To specific:** Number of concepts and messages for the test split and their ratio.

	C_{test}	M_{test}	M_{test}/C_{test}
D(3,4)	12	11.8 ± 0.4	0.98 ± 0.03
D(3,8)	24	22.2 ± 0.4	0.93 ± 0.02
D(3,16)	48	44.4 ± 0.8	0.93 ± 0.02
D(4,4)	16	15.6 ± 0.8	0.97 ± 0.05
D(4,8)	32	28.8 ± 3.1	0.90 ± 0.10
D(5,4)	20	19.6 ± 0.5	0.98 ± 0.02

Table 7: **To generic:** Number of concepts and messages for the test split and their ratio.

C Example protocols and qualitative analysis

In this section, we show examples of the messages the agents used to refer to concepts during testing for both conditions. We also conduct a short qualitative analysis.

In Table 8 and Table 9, we show for each dataset one randomly picked example of a concept that the agents have seen during testing. The concept is a specific concept in the "to specific" test case. This means that all attributes are fixed to a specific value, e.g. (1,1,2) for D(3,4). These concepts are presented in a randomly sampled context condition. We define the context condition as the number of shared attributes between target concept and objects in the context, i.e. distractors. For example for D(3,4), there is one shared attribute between target concept and objects in the context. This means that the higher the context condition, the closer the context is to the target concept, i.e. the more specific a message has to be to be sufficiently discriminative in a certain context. The messages end with the EOS symbol that terminates the message, i.e. "0". These are examples from the interactions that have been gathered during testing.

Our goal for the qualitative analysis was to check whether specific symbols have been associated with a specific attribute during training. For this purpose, we constructed a mapping between fixed attributes and symbols uttered during training based on the mutual information score defined in section 3.2. This mapping is position-sensitive, i.e. we find the symbol with the highest mutual information for an attribute at a certain position in the concept. The two rightmost columns show the symbols which have been associated with a specific attribute in a specific position as well as the respective mutual information score. For example for D(3,4), the symbol associated with a value 1 in the first position of the target concept is "4". This symbol is not communicated in the message from this example. The symbols associated with values 1 and 2 in the second and third position of the target concept, however, are encoded in the messages, i.e. "14" and "11".

In the to specific condition, we generally observe quite high mutual information scores for symbols being associated with certain attributes. In addition, we observe a high tendency of these symbols being included in the speaker agents' actual messages.

In the "to generic" condition, the test dataset contains only generic concepts, i.e. concepts with only one fixed attribute. Attributes which are not fixed and thus irrelevant to the target concept are represented as _ in Table 9. By definition, generic concepts can only appear in coarse contexts, i.e. when 0 attributes are shared between the target concept and the objects in the context. Again, we randomly selected one example from the test interactions. We conducted the same qualitative analysis as for the "to specific" condition. However, as we have seen in the quantitative analyses presented in the main paper that in the "to generic" condition,

	fixed indices	fixed values	context condition	message	symbol	symbol MI
D(3,4)	(1,1,1)	(1,1,2)	1	[11,14,14,0]	4	0.7340
					14	1
					11	0.7465
D(3,8)	(1,1,1)	(3,0,4)	1	[18,14,8,0]	13	0.4077
					15	0.5491
					18	0.3100
D(3,16)	(1,1,1)	(0,14,13)	2	[31,40,20,0]	27	0.0724
					40	0.1166
					13	0.0471
D(4,4)	(1,1,1,1)	(0,0,2,0)	0	[2,13,9,14,0]	14	0.6409
					9	0.3386
					13	0.3141
					2	0.8785
D(4,8)	(1,1,1,1)	(3,7,0,5)	1	[22,10,7,19,0]	22	0.4357
					10	0.4502
					12	0.1765
					19	0.6663
D(5,4)	(1,1,1,1,1)	(3,2,1,2,1)	4	[10,15,4,12,14,0]	15	0.3144
					14	0.3086
					10	0.9573
					12	0.2455
					4	0.1043

Table 8: **To specific:** One random example for a specific concept from the test data per dataset, the context condition in which it was presented (in number of shared attributes) and the message that was used to refer to the concept. The two rightmost columns present the results of a qualitative analysis where we sampled symbols that have been associated with attributes of the target concept during training.

	fixed indices	fixed values	context condition	message	symbol	symbol MI
D(3,4)	(1,0,0)	(1,_,_)	0	[14,14,2,0]	14	0.3181
D(3,8)	(1,0,0)	(1,_,_)	0	[12,1,13,0]	8	0.2446
D(3,16)	(0,0,1)	(_,_,2)	0	[18,13,28,0]	18	0.0665
D(4,4)	(0,1,0,0)	(_,1,_,_)	0	[13,13,6,6,0]	9	0.0002
D(4,8)	(0,0,1,0)	(_,_,3,_)	0	[24,10,8,8,0]	25	0.0918
D(5,4)	(0,0,0,1,0)	(_,_,_,0,_)	0	[7,7,9,14,14,0]	2	0.0143

Table 9: **To generic:** One random example for a generic concept from the test data per dataset, the context condition in which it was presented (in number of shared attributes) and the message that was used to refer to the concept. The two rightmost columns present the results of a qualitative analysis where we sampled symbols that have been associated with attributes of the target concept during training.

agents mostly reuse messages from training, we do not expect to see the same pattern as in the "to specific" condition. Indeed, the mutual information between single symbols and attributes is rather small with the highest value being 0.31. In line with this observation, we do not find a consistent position-sensitive attribute-symbol mapping for the "to generic" condition. And the symbols with the highest mutual information are not consistently included in the messages.

These qualitative findings support the main conclusion from the quantitative analyses, namely that agents use different strategies for generalizing "to specific" or "to generic" concepts. In the "to specific" condition, position-sensitive symbol-attribute mappings emerge during training and are successfully used for generalizing via composition.

D Dataset sizes

Tables 10 and 11 show the sizes of the datasets.

	training	validation	test	total
D(3,4)	810	270	64	1144
D(3,8)	3060	992	512	4564
D(3,16)	11880	3936	4096	19912
D(4,4)	7320	2432	256	10008
D(4,8)	52080	17344	4096	73520
D(5,4)	55350	18432	1024	74806

Table 10: **To specific:** Number of unique concepts in each dataset for each dataset split.

	training	validation	test	total
D(3,4)	780	308	12	1036
D(3,8)	3435	1429	24	4376
D(3,16)	15606	7945	48	19504
D(4,4)	7429	2683	16	9871
D(4,8)	55972	21339	32	73248
D(5,4)	56140	19507	20	74643

Table 11: **To generic:** Number of unique concepts in each dataset for each dataset split.

E Computational Budget

We ran the experiments reported in this paper on a High-Performance-Computing Cluster (HPC3) on a single gpu core, using up to 400GB memory. We estimate the computing time for reproducing all results reported here, including generating the datasets and training the models on the six datasets for five runs at <72h with comparable resources.