Context Shapes Emergent Communication about Concepts at Different Levels of Abstraction

Kristina Kobrock, Xenia Ohmer, Elia Bruni, Nicole Gotzner
Institute of Cognitive Science, University of Osnabrück
Wachsbleiche 27, 49090 Osnabrück
{kristina.kobrock, xenia.ohmer, elia.bruni, nicole.gotzner}@uni-osnabrueck.de

Abstract

We study the communication of concepts at different levels of abstraction and in different contexts in an agent-based, interactive reference game. While playing a concept-level reference game, the neural network agents develop a communication system from scratch. We use a novel symbolic dataset that disentangles concept type (ranging from specific to generic) and context (ranging from fine to coarse) to study the influence of these factors on the emerging language. We compare two game scenarios: one in which speaker agents have access to context information (context-aware) and one in which the speaker agents do not have access to context information (context-unaware). First, we find that the agents learn higher-level concepts from the object inputs alone. Second, an analysis of the emergent communication system shows that only context-aware agents learn to communicate efficiently by adapting their messages to the context conditions and relying on context for unambiguous reference. Crucially, this behavior is not explicitly incentivized by the game, but efficient communication emerges and is driven by the availability of context alone. The emerging language we observe is reminiscent of evolutionary pressures on human languages and highlights the pivotal role of context in a communication system.

Keywords: emergent communication, concepts, context

1. Introduction

Referring to things in the world is crucial to effective communication. When choosing a referring expression, speakers recur to what they know about the referent's underlying concept and choose to communicate the concept at a level of abstraction that fits well with their communicative intentions. For example, a reference to the object in Figure 1 can be made at various levels of abstraction, ranging from the more *specific concept* 'watermelon' to the more *generic concept* 'food'. When communicating a more generic concept, speakers and listeners need to abstract away from properties of the individual objects and focus on what all objects belonging to a concept have in common. By choosing the utterance 'fruit', for example, a speaker abstracts away from irrelevant properties (the size, color etc. of the specific object) and stresses the properties that watermelons share with other items belonging to the concept 'fruit', for example that they are edible.

Crucially, the concepts speakers choose to communicate also depend on the situational context: While in a *coarse context* (Figure 2A) the referring expression 'melon' is sufficient to discriminate the target object, a *fine context* (Figure 2B) requires a more specific reference such as 'watermelon' to resolve ambiguity (e.g., Graf et al., 2016; Hawkins et al., 2018; Winters et al., 2015).

In this contribution, we focus on the immediate, or situational, context, which is defined as "the situational information that is relevant for producing and comprehending an utterance" (Winters et al., 2018). Previous research has shown that the immediate context influences the way humans communicate at different timescales ranging from the situational use of a referring expression in a specific context to the emergence of a communica-

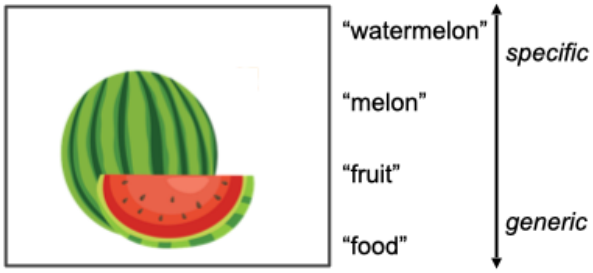

Figure 1: Example referring expressions at different levels of abstraction.

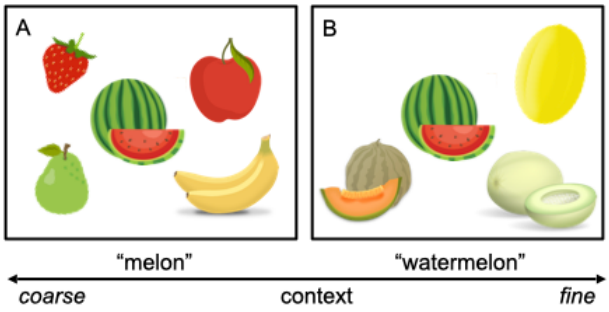

Figure 2: Example referring expressions in different contexts.

LREC-COLING 2024, pages 3831–3848

tion system. Regarding the former, research on context-based pragmatic phenomena has found that humans usually tailor their utterances to the immediate context. For example, Sedivy et al. (1999) found a referential contrast effect for items belonging to the same category such as a 'glass'. In contexts with two glasses present, participants were more likely to modify their expression with an adjective, e.g., 'tall glass' to discriminate the target object from a contrast object that belonged to the same category, e.g. a short glass (see also Sedivy, 2005, 2003). Regarding the latter, artificial language learning studies with human participants have investigated how context shapes an emerging language (e.g., Winters et al., 2015; Hawkins et al., 2018). Winters et al. (2015), for example, manipulated the specific dimensions which were relevant for discrimination in situational utterances and found that emerging languages encode specifically these dimensions. In other words, if participants were presented with contexts in which the target object differed from the distractor object in the shape dimension during most iterations, the emerging language would encode shape, but no irrelevant dimensions. Hawkins et al. (2018) found context to shape an artificial language between humans when they had to communicate hierarchically organized stimuli with novel expressions. Contextual pressures shaped the emerging lexicon in a way that when participants were presented with mostly fine contexts, each word was paired with a single meaning. In contrast, when participants communicated in coarser contexts, polysemous meanings emerged, allowing a word to have more than one object as a referent. In other words, the number of words referring to several objects was shown to be higher for participants in the coarse context condition (Hawkins et al., 2018). These findings are in line with Grice's maxim of quantity which predicts that speakers should choose utterances that are optimally informative for the listener.

However, experimental evidence shows that speakers' use of referring expressions is not always straightforward: Over- and underinformative expressions are frequent in natural conversation, and the focus of ongoing research (e.g., Degen et al., 2020; Tourtouri et al., 2019; Rubio-Fernandez, 2021; Rubio-Fernández, 2016). These studies with human participants shed light on their behavior, but the question of which communicative strategies speakers follow when referring to concepts at different levels of abstraction remains unanswered until now. We use an agent-based model with systematic manipulations to study under which circumstances a specific behavior is beneficial to communicative success.

We investigate the role of pragmatics in the communication of concepts at different levels of abstraction and in different contexts with an emergent communication paradigm using a reference game. Reference games, where a speaker describes a target and a listener has to identify the correct target among a set of distractors, are ideal for studying references at different levels of abstraction because they allow for systematic manipulation of the context (e.g., Frank et al., 2016; Hawkins et al., 2018; Graf et al., 2016; Degen et al., 2020). More recently, this game setup has been adapted to computational studies of emergent communication between deep neural network agents (e.g., Lazaridou et al., 2018, 2017; Ohmer et al., 2022; Mu and Goodman, 2021). Such computational methods allow for rigorous manipulations, and for simulating language on various time scales from evolution to situational use. They are therefore increasingly used to answer questions in the field of pragmatics (e.g., Monroe et al., 2017; White et al., 2020; Ohmer et al., 2021; Fang et al., 2022; Hu et al., 2022; Yuan et al., 2021; Andreas and Klein, 2016; Kang et al., 2020). Going beyond previous work, we systematically study the influence of concept and context type on the choice of referring expressions during emergent communication.

2. Method

2.1. General Setup

A speaker and a listener agent develop a communication system while playing a concept-level reference game (see Figure 3). Other than in a classical reference game (Lewis, 1969), the speaker has to communicate not a single but multiple targets belonging to the same concept (Mu and Goodman, 2021). The neural network agents are trained in a Reinforcement Learning paradigm with the Gumbel-Softmax relaxation (Jang et al., 2017) and are rewarded when the listener picks the correct target objects after having decoded a message generated by the speaker. A similar setup has been used in previous related work (Ohmer et al., 2022; Mu and Goodman, 2021).

2.2. Dataset

We train the agents on a novel symbolic dataset that disentangles *concept type*, ranging from specific to generic, from *context type*, ranging from fine to coarse. The most specific concept is defined by target objects where all attributes have a fixed value (e.g., *'blue circle'*). Objects that define the most generic concept have only one fixed attribute (e.g., *'circle'*). Distractors in a fine context share more attributes with the target concept, whereas distractors in a coarser context condition

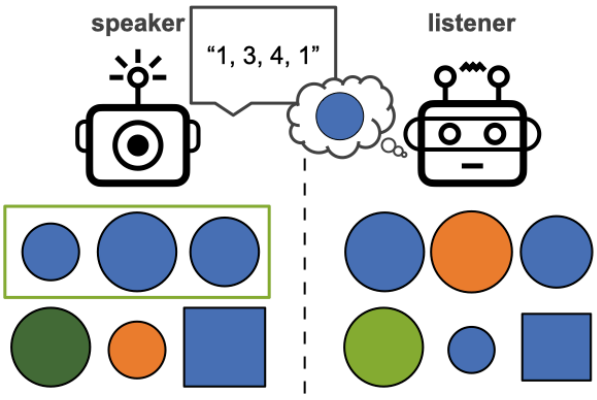

Figure 3: Schematic illustration of the concept-level reference game with the *specific* target concept "blue circle" (fixing both shape and color attributes) and a *fine* context condition (distractors share one attribute, either shape or color, with the target concept). Note that the objects that satisfy the concepts can differ between agents, in this case, they have different sizes. The speaker knows which objects are the targets (here displayed in the green box) and the listener receives a shuffled input from which it has to select the objects that satisfy the target concept.

share fewer attributes with the target concept. Following previous work (e.g., Ohmer et al., 2022), the concepts are represented by an object vector that has n attributes which can each take k values, and a concept-defining, binary vector $\vec{d} \in \{0,1\}^n$ that specifies which attributes are fixed to a specific value (1) and which can vary (0). To generate a dataset, we sample all possible concepts (restricted by the number of attributes and values) ranging from the most specific concept with a concept-defining vector consisting only of ones to the most generic concept, where the concept-defining vector is a one-hot vector, which fixes exactly one attribute value. We then sample all possible context conditions for these concepts by changing between one (fine context) and $n-1$ (coarse context) attributes relative to the target concept. For instance, if objects are defined by $n = 3$ attributes (color, shape, scale) and the target concept is defined as "blue circle" (two out of three attributes fixed), in the fine context condition, the distractors would differ in only one of the fixed dimensions (e.g., "blue square" or "green circle", see Figure 3). Note that this procedure is only used for constructing the dataset. The speaker and listener agents are trained on the target and distractor objects alone and have to figure out the concepts from these inputs.

2.3. Concept-level Reference Game

Following Mu and Goodman (2021) and Ohmer et al. (2022), we define a communication game between speaker S and listener L as a communication game $G = (T^S, D^S, T^L, D^L)$, where $T^S = \{t_1^T, ..., t_g^T\}$ is a set of game size g target objects presented to the speaker, $D^S = \{d_1^S, ..., d_g^S\}$ is a set of g distractor objects presented to the speaker, and T^L and D^L are defined analogously for the listener. The concept-level reference game is defined as a game where $T^S \neq T^L$ and $D^S \neq D^L$. This setup has been shown to increase systematicity in the emerging communication protocol and the agents' ability to generalize compared to the standard reference game (Mu and Goodman, 2021). Each round, S receives both targets T^S and distractors D^S in an ordered fashion. Based on this input, 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.[1] L receives m and their own set of targets T^L and distractors D^L shuffled together (hereafter $X^L = \{x_1^L, ..., x_i^L\}$, where $i = 2 \cdot g$, because L does not know which are the targets and which are the distractors). Based on these inputs, L predicts a label $y_i^L \in \{0,1\}$ (0: distractor, 1: target) for each object x_i^L in its input.

2.4. Architecture and Training

Our implementation[2] makes use of the EGG framework for emergent communication games (Kharitonov et al., 2019). Both agents are implemented as single-layer Gated Recurrent Units (GRUs) (Cho et al., 2014) as in previous related work (Mu and Goodman, 2021; Ohmer et al., 2022). Typically, either GRUs or LSTMs are used in the emergent communication paradigm because on the one hand, such recurrent neural networks are a better choice for modeling language than simple feed-forward neural networks because they can deal with sequential input of any length (Jurafsky and Martin, 2024), and on the other hand, they have a simpler architecture and are thus easier to train than, for example, Transformers (Vaswani et al., 2017). The speaker input is processed by two dense layers that embed the targets and distractors separately, and a third dense layer that concatenates both embeddings. The listener input is also embedded with a dense layer. After having decoded the message from the speaker, the listener returns the dot product be-

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

[2] All code and analysis scripts are available at https://github.com/kristinakobrock/context-shapes-language.

tween the received message and each of the embedded input objects. We jointly train a speaker-listener pair to maximize the listener's likelihood of selecting the correct targets. We train with binary cross entropy loss allowing the listener to predict a label $y_i \in \{0, 1\}$ (0: distractor, 1: target) for each object x_i. The loss is

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

where $\hat{m} \sim p^S(m|T^S, D^S)$ and $p^L(y_i^L|x_i^L, \hat{m}) = \sigma(\text{GRU}^L(\hat{m}) \cdot \text{embed}(x_i^L))$. We use the straight-through Gumbel-Softmax trick (Jang et al., 2017) with temperature $\tau = 2$ and a decay rate of 0.99 for training to ensure differentiability for backpropagation. We split the data in training (60%), validation (20%) and test (20%) datasets. All splits contain different concepts, i.e. unique object- and concept-defining vector combinations, and all possible context conditions for these concepts. We evaluate performance on the validation dataset after each training epoch and on the test dataset once after training. We train five runs for both game settings on six datasets for 300 epochs using the Adam optimizer. We conducted a grid search to find hyperparameters that led to a high performance on the validation sets of all datasets.[3] We train with batch size 32, learning rate 0.001 and game size 10, i.e., there are ten target and ten distractor objects in a game. Agents have an embedding layer with 64 units and a hidden layer with 128 units. The maximum message length M is set to the number of attributes in a dataset plus the End of Sequence (EOS) symbol 0. The vocabulary size for each dataset is determined by the number of attribute values in the dataset. We define a minimal vocabulary size for each dataset as the number of attribute values plus one additional symbol to encode additional information like position or relevance. The vocabulary size is calculated by multiplying this minimal vocabulary size with a factor $f = 3$ according to previous work (Ohmer et al., 2022).

2.5. Game Scenarios and Hypotheses

We implement two game scenarios to investigate the agents' communicative strategies, and specifically, whether they develop and use pragmatic behavior in the sense of context-based pragmatics (e.g., Sedivy, 2003). The basic setup involves speaker and listener agents which learn to communicate about concepts ranging from specific to generic in all context conditions. We compare two game settings: In the context-unaware setting, the speaker agent only has access to the target objects and the listener has access to both targets and distractors. In the context-aware setting, on the other hand, both speaker and listener agents have access to the target concept and to the context defined by distractor objects. The distractor objects that the listener receives can be different from those that the speaker receives, but they satisfy the same context condition. Similarly, the target objects may be different between speakers and listeners, as long as they satisfy the same target concept. We expect the speakers to use different production strategies depending on the game, as well as the concept and context type.

We formulate the following (non-exclusive) hypotheses mapping to the two games described above:

- **H1: Context-unaware literal agents (L)** have to communicate concepts on the most specific level of abstraction to be successful, thus may be overinformative (non-pragmatic baseline).
- **H2: Context-aware literal agents (L-aware)** can communicate concepts on other than the most specific level of abstraction and can rely on the context to resolve ambiguities (context-based pragmatics).

In our setup, overinformative communication is defined as mentioning specific concepts in coarse contexts, e.g. saying "green circle" in a context where no other circles are present.

3. Evaluation

We report training, validation and test accuracies as a proof of concept that the agents are trained successfully to communicate in the concept-level reference game. We use entropy-based metrics to measure information contained in the emerging messages. The Normalized Mutual Information (NMI) quantifies the degree to which messages and concepts have a one-to-one correspondence. It is calculated as follows:

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

with C being the set of concepts and M being the set of messages. If the NMI score is maximal (1.0), then each message in the emerged lexicon of the agents corresponds to exactly one concept and this concept is only referred to with this message. Additionally, we report efficiency and consistency scores as defined in Ohmer et al. (2022). The consistency score measures whether the agents consistently use the same messages to refer to the

[3] The grid search was conducted for the smallest dataset D(3,4), the one with the highest number of attributes D(5,4), and the one with the highest number of values D(3,16).

same concepts and is calculated as follows:

$$\text{consistency}(C, M) = 1 - \frac{H(M|C)}{H(M)}.$$

The effectiveness score measures whether agents effectively use messages that uniquely identify the target concept and is calculated as follows:

$$\text{effectiveness}(C, M) = 1 - \frac{H(C|M)}{H(C)}.$$

The datasets are implemented as described above and named by the number of attributes (think 'shape', 'color', etc.) and values (think 'square', 'circle', etc.) an object in this dataset can take. For example, 'D(3,4)' means that objects in this dataset have three attributes that take four values each. We run simulations for six datasets that span a range of three to five attributes and four to 16 values (see Table 1). We report means and bootstrapped 95% Confidence Intervals (CIs) from five simulations per dataset and 300 training epochs. To statistically analyze the relevant contrast between context-aware and context-unaware agents for evaluating our hypotheses, we performed a Bayesian analysis of the NMI scores between these conditions. Specifically, we evaluate NMI scores in the edge cases of concept and context conditions, i.e. the most specific and most generic concepts and the finest and coarsest contexts.

	$k = 4$	$k = 8$	$k = 16$
$n = 3$	$D(3, 4)$	$D(3, 8)$	$D(3, 16)$
$n = 4$	$D(4, 4)$	$D(4, 8)$	
$n = 5$	$D(5, 4)$		

Table 1: Datasets with n attributes and k values, labeled as $D(n, k)$.

We additionally conducted a small qualitative analysis of the messages. For this, we randomly sampled one specific concept from the last interaction of training on the D(4,4) dataset. We report all unique messages that the agents used to describe this concept for each context.

4. Results

4.1. Performance

We calculate accuracy as the average number of correct predictions by the listener, rather than the average number of games without any mistakes. As a result, the agents can achieve high accuracies when the listener correctly identifies most objects per game. First, we observe very high training and validation accuracies for all game settings and datasets (mean training and validation accuracies across runs > 0.96 for all datasets and both settings).[4] Mean test accuracies across runs on concepts that the agents have never encountered during training are 0.89 (SD=0.07) for context-unaware and 0.87 (SD=0.11) for context-aware agents. This suggests that the agents learn to successfully communicate about concepts on various levels of abstraction and in various context conditions.[5]

To get a better understanding of the agents' strategies and where communication is especially (un)successful, we performed an additional analysis of the errors, i.e. those cases where the listener agents predict some of the labels wrongly.[6] We find that most errors occur when targets and distractors share many attributes, making it more likely that they are confused with each other. In other words, most mistakes happen in the fine context conditions.[7]

4.2. Qualitative communication analysis

Second, we use a qualitative analysis of the messages to see whether we can observe differences between context-unaware and context-aware settings. Tables 2 and 3 show the results of our qualitative analysis on the D(4,4) dataset for context-unaware and context-aware, respectively. We report all unique messages for a randomly chosen specific concept ($[0, 0, 0, 3]$, all attributes fixed) and each context condition. Context-unaware agents tend to use the same messages in all context conditions (in this case, "$[11, 1, 11, 14, 0]$" is used consistently across contexts). On the other hand, context-aware agents use a larger set of messages (four unique messages over all games), and they tend to vary the messages more depending on context. In coarser contexts, the set of messages used to describe the target concept is larger than in the finest context, where the best strategy is to communicate the most specific concept. We observe the same pattern for other randomly selected concepts across different datasets.[8]

[4] It is important to note that achieving such high scores is intentional. Only with a high success score does the rest of the evaluation become meaningful. This ensures that the language we analyze can be assumed to effectively communicate what is intended in the referential game.

[5] Detailed accuracy scores can be inspected in Table 4 in Appendix A.

[6] Plots for these analyses can be inspected in Appendix B.

[7] Additional plots of the distribution of false positive and false negative errors can be found in Appendix B.2 and B.3.

[8] More examples are given in Appendix C to show that these are not cherry-picked.

Object	Context (# Shared Attributes)	Unique Messages
$[0, 0, 0, 3]$	0	"$[11, 1, 11, 14, 0]$"
	1	"$[11, 1, 11, 14, 0]$"
	2	"$[11, 1, 11, 14, 0]$"
	3	"$[11, 1, 11, 14, 0]$"

Table 2: **Context-unaware:** Unique messages used to refer to a randomly picked specific concept in the D(4,4) dataset over different context conditions.

Object	Context (# Shared Attributes)	Unique Messages
$[0, 0, 0, 3]$	0	"$[6, 2, 10, 14, 0]$" "$[6, 2, 14, 10, 0]$"
	1	"$[6, 2, 10, 14, 0]$" "$[6, 2, 14, 10, 0]$"
	2	"$[6, 2, 10, 1, 0]$" "$[6, 2, 10, 14, 0]$" "$[6, 2, 10, 5, 0]$"
	3	"$[6, 2, 10, 5, 0]$"

Table 3: **Context-aware:** Unique messages used to refer to a randomly picked specific concept in the D(4,4) dataset over different context conditions.

4.3. Quantitative communication analysis

Mappings between concepts and messages Third, we use information-theoretic scores and compare the context-unaware to the context-aware setting to quantify the results we obtained from our qualitative analysis. In the context-unaware setting, we observe high overall information scores (NMI scores ranging from 0.94 [0.9, 0.98][9] for D(5,4) to 0.97 [0.96, 0.98] for D(3,8)). This suggests that concepts and messages tend to have one-to-one mappings. While the mutual information between messages and concepts is also relatively high for context-aware agents, it is slightly lower than for context-unaware agents (NMI scores ranging from 0.86 [0.78, 0.92] for D(3,16) to 0.9 [0.88, 0.93] for D(3,8)). This could mean that context-aware trained agents adapt to the context, making strict one-to-one mappings impractical.

Figure 4 shows for the context-unaware setting how the mutual information varies when it is calculated for all concept and context conditions for dataset D(4,4).[10] Here, we observe two patterns: On the one hand, the NMI increases with the number of fixed attributes. In other words, the more specific the concepts are, the more one-to-one mappings between concepts and messages emerge. On the other hand, the NMI scores stay relatively constant across different numbers of shared attributes. This suggests that context-unaware trained speaker agents adapt their choice of reference to a concept's levels of abstraction, but not to the context (of which they are not aware).

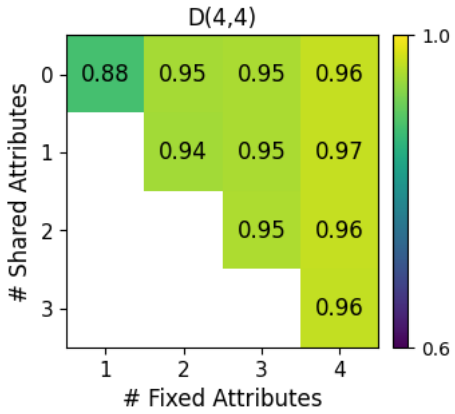

Figure 4: **Context-unaware:** Mean NMI scores across all datasets for different concept (# fixed attributes) and context conditions (# shared attributes). From top to bottom context becomes finer and from left to right concepts become more specific.

When looking at the NMI for the context-aware setting in Figure 5, we observe the opposite pattern: While changes in the concept level (i.e., the number of fixed attributes) are not reflected in changing NMI scores, we do observe increasing NMI scores with an increasing number of shared attributes. In other words, the finer the context, the more one-to-one mappings between concepts and messages can be found in the agents' communication system.

Effect of the level of abstraction We will now look first at the effect of a concept's level of abstraction and then at the effect of the context on the emerging language in more detail. The effect of a concept's level of abstraction on the emerging language is visualized in Figures 6 and 7 which plot the entropy-based scores over different concept levels aggregated over all datasets and simulation runs for context-unaware and context-aware, respectively. In Figure 6, we observe that the NMI is largely constant for more specific concepts (three fixed attributes and more) and slightly drops toward more generic concepts with one or two fixed attributes. This effect is largely driven by a corresponding drop in the consistency score when it

[9] The intervals reported here are bootstrapped 95% Confidence Intervals.

[10] Plots for all datasets are available in Appendix D.

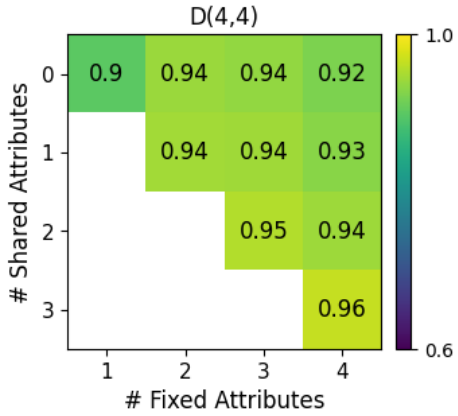

Figure 5: **Context-aware:** Mean NMI scores across all datasets for different concept (# fixed attributes) and context conditions (# shared attributes). From top to bottom context becomes finer and from left to right concepts become more specific.

comes to more generic concepts, which suggests that more than one unique message is used to refer to the same generic concept. We can think of two reasons for this: One reason might be that the agents are overly specific when referring to the generic target concept, for example, they might use "red circle" or "blue circle" to refer to "circle". Another reason is that the emerging language contains more synonymous words that refer to more generic concepts, for example the invented messages "1, 1, 2" and "2, 3, 4" both mean "circle".

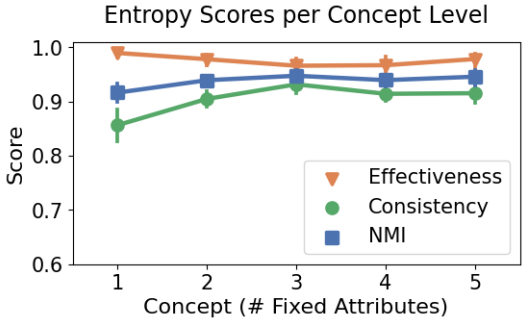

Figure 6: **Context-unaware:** Mean entropy scores across all datasets for different concept levels indicated by the number of fixed attributes. From left to right concepts become more specific. Error bars indicate bootstrapped 95% confidence intervals.

Figure 7 shows that we observe a drop in the consistency score when it comes to more generic concepts also for languages developed by context-aware agents. Additionally, we find that consistency decreases again for more specific concepts (i.e., when the number of fixed attributes is larger than three). This can be explained by the availability of context in the context-aware setting: For more specific concepts with three or more attributes, there are more context conditions possible, i.e. $n-1$ context conditions. Thus, context-aware trained speakers adapt to use different messages to refer to the same concepts when they take context into account. The butterfly shape we observe in Figure 7, where effectiveness increases for specific and for generic concepts and consistency, on the other hand, decreases for specific and for generic concepts, can thus be explained by the two factors that the agents take into account when constructing messages, both concept specificity and context.

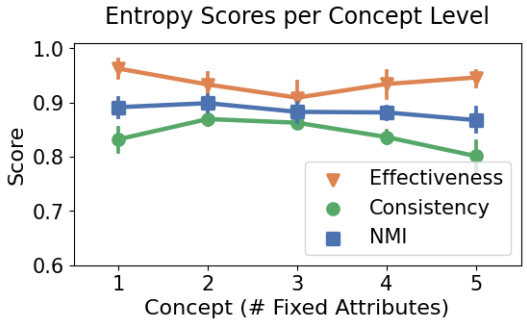

Figure 7: **Context-aware:** Mean entropy scores across all datasets for different concept levels indicated by the number of fixed attributes. From left to right concepts become more specific. Error bars indicate bootstrapped 95% confidence intervals.

We used Bayesian estimation to statistically analyze these observed differences between conditions across all five runs, following Kruschke (2013). We find no substantial difference in NMI scores between the context-unaware (M=0.92, CrI=[0.9, 0.93][11]) and the context-aware (M=0.89, CrI=[0.88, 0.91]) setting for generic concepts with an estimated difference in means of M=0.023 (CrI=[-0.003, 0.048], pd=95.9%[12], 6% *in ROPE*[13]). The difference for specific concepts between the context-unaware (M=0.94, CrI=[0.92, 0.97]) and the context-aware (M=0.87, CrI=[0.84, 0.91]) setting on the other hand is substantial with an estimated difference in means of M=0.07 (CrI=[0.026,

[11] Credible Intervals (CrIs) were computed on the posterior via the Highest Density Intervals.

[12] The probability of direction (pd) can be interpreted as the probability that a parameter's posterior distribution is strictly positive or negative (Makowski et al., 2019).

[13] The Region Of Practical Equivalence with zero (ROPE) was calculated by using one-tenth of the standard deviation of the response variable around the null following recommendations by (Kruschke, 2018): ROPE = [-0.004, 0.004].

0.109], *pd*=99.4%, 0% *in ROPE*). While these effects are rather small, we do find reliable differences. These results are in line with our observations above, specifically that specific concepts can appear in a wider range of contexts (coarse to fine). Thus, context-aware agents use a wider range of messages to refer to the same specific concept than context-unaware agents because they can make use of the context.

Effect of the context The effect of the context on the emerging language is especially evident when we compare Figure 8 and Figure 9 which plot the entropy-based scores over different context conditions for context-unaware and context-aware settings. In the context-unaware setting, the NMI stays at a constant level across different context conditions. We observe a small drop in consistency and an increase in effectiveness for fine contexts (i.e., for 3 or 4 shared attributes) in the datasets with at least 4 attributes. These results are in line with the hypothesis that context-unaware speakers communicate concepts on the most specific level in all contexts, including coarser contexts. This behavior can be referred to as over-informative from the listener's perspective. For example, in a coarse context where no other circles are present, communicating a specific concept like "red circle" is considered overinformative.

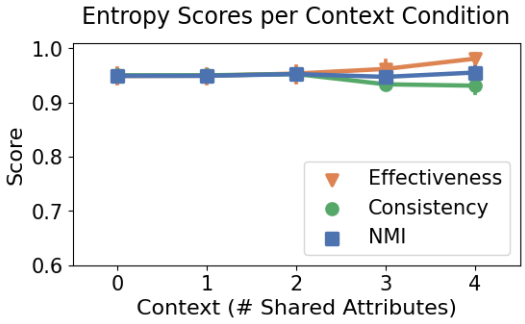

Figure 8: **Context-unaware:** Mean entropy scores across all datasets for different context conditions indicated by the number of shared attributes. From left to right context becomes finer. Error bars indicate bootstrapped 95% confidence intervals.

When agents are trained context-aware, on the other hand, we observe that the information-theoretic scores differ more between context conditions (see Figure 9). Specifically, we observe a pattern where the coarser the context (i.e., the fewer shared attributes), the lower the NMI and the finer the context (i.e. the more shared attributes), the higher the NMI. When agents develop fewer one-to-one mappings between messages and concepts in the coarse context conditions, this might indicate that they adapt more to the context which makes one-to-one mappings impractical. The reason for this might be that in coarse contexts, both more and less specific messages can be successful (e.g., "circle" can mean 'red circle', 'blue circle' etc.) because when less specific messages are used, the target concept can still be disambiguated by the context. In fine contexts, on the other hand, the messages need to contain more information on more specific levels of abstraction to be sufficiently discriminative in the context, which intuitively results in more one-to-one mappings (e.g., a more specific utterance like "red circle" is only used for the more specific concept 'red circle').

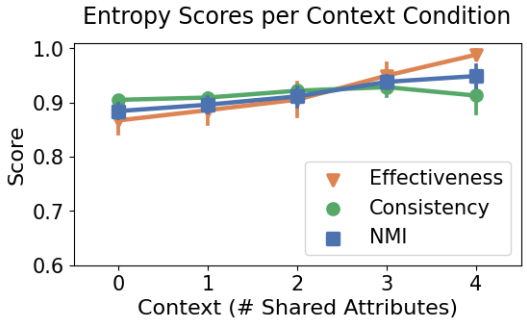

Figure 9: **Context-aware:** Mean entropy scores across all datasets for different context conditions indicated by the number of shared attributes. From left to right context becomes finer. Error bars indicate bootstrapped 95% confidence intervals.

In line with these observations, we find a substantial difference in NMI scores between the context-unaware (*M*=0.95, *CrI*=[0.94, 0.96]) and context-aware setting (*M*=0.89, *CrI*=[0.87, 0.9]) only for coarse contexts with a difference in means of *M*=0.064 (*CrI*=[0.046, 0.811], *pd*=100%, 0% *in ROPE*). For fine contexts, the difference in NMI scores between the context-unaware (*M*=0.95, *CrI*=[0.94, 0.97]) and the context-aware setting (*M*=0.95, *CrI*=[0.92, 0.97]) is not significant (*M*=0.008, *CrI*=[-0.026, 0.041], *pd*=70.1%, 20% *in ROPE*).

Looking at effectiveness and consistency scores in the context-aware setting, we observe higher consistency and lower effectiveness scores for coarse contexts and higher effectiveness and lower consistency scores for fine contexts. This means that agents tend to consistently use the same messages to refer to the same concepts (i.e. no synonyms) in coarser contexts and that agents tend to effectively use messages that uniquely identify the target concept (i.e. non-polysemous expressions) in finer contexts. This makes sense because the finer the context gets, the more it is necessary to distinguish the target concepts from the distractors.

5. Discussion

With our interactive agent-based model, we were able to generate three main insights about concept communication in various contexts and how this setup shapes an emerging language.

First, we show that artificial agents can learn to communicate successfully about concepts at different levels of abstraction and in different contexts in a concept-level reference game. Previous work has explicitly encoded concept information in the form of relevance vectors (Ohmer et al., 2022) or prototype embeddings (Mu and Goodman, 2021). For humans, however, abstracting the relevant concept, or level of reference, happens without such explicit information. Here, we show that agents can learn higher-level concepts from the object inputs alone, providing a more natural model for the emergence of abstraction.

Second, we find that only context-aware agents learn to communicate efficiently by adapting their messages to the context conditions. While context-unaware agents use the same messages to refer to concepts in all context conditions, context-aware agents adapt their messages successfully to the context. Overinformative communication, in the sense that specific concepts are communicated also in coarse contexts where they contain more information than necessary for disambiguation, is reduced in the context-aware game scenario. This might indicate that context-aware agents communicate more efficiently (Piantadosi et al., 2012). It should be noted, though, that these agents do not share the same biases as humans. Future work should focus on the biases and pressures that shape the emerging language between artificial agents towards the kind of efficient overinformative communication we often observe in humans (e.g., Degen et al., 2020; Rubio-Fernandez, 2021; Tourtouri et al., 2019; Kreiss et al., 2017).

Third, we conclude that the availability of context alone shapes the emerging language towards being more efficient (i.e. less overinformative) without additional pressures. The agents were not explicitly incentivized to use the context but they share the same architecture and training procedure with the context-unaware agents, the only difference being that they also receive distractor objects as input. Because we have not incentivized the context-aware agents to use context, they could follow the same strategy as context-unaware agents and be maximally specific all the time. Instead, we find that the agents develop a strategy that makes use of the context in which they communicate. Although the differences we observe between the context-aware and context-unaware settings are rather small, they are reliable and they do indicate that the mere presence of context already drives its use in communication. Future work can investigate whether pressures, such as increasing cognitive load for longer messages, would even intensify these differences.

Our results are in line with previous work on how an emerging vocabulary depends on the contexts in which the targets are presented. Hawkins et al. (2018) found a similar pattern in an artificial language learning paradigm with human participants: The finer the context, the more one-to-one mappings are established in an emerging language, and the coarser the context, the more synonyms can be found. Further, they also found that an emerging language contains more words that refer to only one concept and fewer that refer to more than one concept when participants only encounter fine contexts.

Our modeling results add to this evidence and highlight the role of context from a different angle. We treat neural network models as testbeds for hypotheses on human cognition. Here, we show that context in itself is a pressure that drives efficiency in an evolving language. Even though our neural network agents lack human cognitive biases, they develop more efficient protocols when they can (but do not have to!) access information about the context compared to when they cannot. This finding demonstrates that the presence of context alone may drive aspects of pragmatic communication without any additional pressures and cognitive prerequisites. We can take this as evidence for the role of external factors such as context for the emergence of an efficient communication system. In line with that, Piantadosi et al. (2012) argue that ambiguity, as we see it in the emerging communication system in the context-aware setting, makes a language efficient because it can usually be resolved by context. Our simulations provide evidence for this hypothesis.

In conclusion, the here presented models and analyses contribute to our understanding of referential communication and the role of pragmatics in communicating concepts through a systematic manipulation of communicative needs. Our results show that the speaker's access to the context shapes the emerging communication system, reproducing a pattern that was observed in humans (e.g., Hawkins et al., 2018; Winters et al., 2018, 2015). These findings have implications both for linguistics research with the questions of how human language evolved and how we make use of language efficiently, as well as for emergent communication research with the question of how we can build artificial models that communicate in a human-like way. More generally, our work illustrates how language emergence simulations with neural network agents can be used to explore questions about human cognition.

6. Acknowledgements

We thank three anonymous reviewers for their helpful comments and feedback.

The simulations were run on a high-performance computing cluster funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 456666331. Kristina Kobrock is 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, Visualization. **Xenia Ohmer:** Conceptualization, Methodology, Software, Writing - Review & Editing, Visualization. **Elia Bruni:** Conceptualization, Methodology, Writing - Review & Editing, Supervision. **Nicole Gotzner:** Conceptualization, Methodology, Writing - Review & Editing, Supervision, Project Administration.

7. Bibliographical References

Jacob Andreas and Dan Klein. 2016. Reasoning About Pragmatics with Neural Listeners and Speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1182.

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.

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.

Fei Fang, Kunal Sinha, Noah D. Goodman, Christopher Potts, and Elisa Kreiss. 2022. Color Overmodification Emerges from Data-Driven Learning and Pragmatic Reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, volume 44, pages 1796–1803.

Michael C. Frank, Andrés Gómez Emilsson, Benjamin Peloquin, Noah D. Goodman, and Christopher Potts. 2016. Rational speech act models of pragmatic reasoning in reference games. PsyArXiv.

Caroline Graf, Judith Degen, Robert D. Hawkins, and Noah D. Goodman. 2016. Animal, dog, or dalmatian? Level of abstraction in nominal referring expressions. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society (CogSci)*, pages 2261–2266.

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, New York.

Robert D. Hawkins, Michael Franke, Kenny Smith, and Noah D Goodman. 2018. Emerging abstractions: Lexical conventions are shaped by communicative context. In *Proceedings of the 40th annual conference of the cognitive science society (CogSci)*, pages 463–468.

Jennifer Hu, Roger Levy, and Noga Zaslavsky. 2022. Scalable pragmatic communication via self-supervision. In *ICML Workshop on Self-Supervised Learning for Reasoning and Perception*.

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

Daniel Jurafsky and James H. Martin. 2024. RNNs and LSTMs. In *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, third edition draft edition.

Yipeng Kang, Tonghan Wang, and Gerard de Melo. 2020. Incorporating Pragmatic Reasoning Communication into Emergent Language. In *Advances in Neural Information Processing Systems*, volume 33, pages 10348–10359.

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.

Elisa Kreiss, Robert D. Hawkins, Judith Degen, and Noah D. Goodman. 2017. Mentioning atypical properties of objects is communicatively efficient. *Cognitive Science*.

John K. Kruschke. 2013. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2):573–603.

John K. Kruschke. 2018. Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280.

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)*.

David K Lewis. 1969. *Convention*. Harvard Univ. Press, Cambridge, Mass.

Dominique Makowski, Mattan S. Ben-Shachar, S. H. Annabel Chen, and Daniel Lüdecke. 2019. Indices of effect existence and significance in the bayesian framework. *Frontiers in Psychology*, 10.

Will Monroe, Robert D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in Context: A Pragmatic Neural Model for Grounded Language Understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

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.

Xenia Ohmer, Michael Franke, and Peter König. 2021. Mutual Exclusivity in Pragmatic Agents. *Cognitive science*, 46(1):e13069.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

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

Paula Rubio-Fernández. 2016. How Redundant Are Redundant Color Adjectives? An Efficiency-Based Analysis of Color Overspecification. *Frontiers in Psychology*, 7:153.

Julie C. Sedivy. 2003. Pragmatic Versus Form-Based Accounts of Referential Contrast: Evidence for Effects of Informativity Expectations. *Journal of Psycholinguistic Research*, 32(1):3–23.

Julie C. Sedivy. 2005. Evaluating Explanations for Referential Context Effects: Evidence for Gricean Mechanisms in Online Language Interpretation. In John C Trueswell and Michael K Tanenhaus, editors, *Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions*, pages 345–364. The MIT Press, Cambridge, Massachusetts; London, England.

Julie C. Sedivy, Michael K. Tanenhaus, Craig G. Chambers, and Gregory N. Carlson. 1999. Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71(2):109–147.

Elli N. Tourtouri, Francesca Delogu, Les Sikos, and Matthew W. Crocker. 2019. Rational over-specification in visually-situated comprehension and production. *Journal of Cultural Cognitive Science*, 3(2):175–202.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Julia White, Jesse Mu, and Noah D. Goodman. 2020. Learning to refer informatively by amortizing pragmatic reasoning. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (CogSci)*, pages 994–1000.

James Winters, Simon Kirby, and Kenny Smith. 2015. Languages adapt to their contextual niche. *Language and Cognition*, 7(3):415–449.

James Winters, Simon Kirby, and Kenny Smith. 2018. Contextual predictability shapes signal autonomy. *Cognition*, 176:15–30.

Luyao Yuan, Zipeng Fu, Jingyue Shen, Lu Xu, Junhong Shen, and Song-Chun Zhu. 2021. Emergence of Pragmatics from Referential Game between Theory of Mind Agents. In *Emergent Communication Workshop, 33rd Conference on Neural Information Processing Systems (NeurIPS)*.

A. Accuracy scores across all datasets

Datasets	Condition	Accuracy		
		training	validation	test
D(3,4)	context-unaware	0.995 *(0.002)*	0.99 *(0.003)*	0.84 *(0.036)*
	context-aware	0.993 *(0.003)*	0.983 *(0.004)*	0.784 *(0.035)*
D(3,8)	context-unaware	0.993 *(0.003)*	0.989 *(0.003)*	0.778 *(0.068)*
	context-aware	0.984 *(0.006)*	0.977 *(0.006)*	0.686 *(0.061)*
D(3,16)	context-unaware	0.981 *(0.007)*	0.979 *(0.008)*	0.896 *(0.005)*
	context-aware	0.969 *(0.005)*	0.968 *(0.006)*	0.874 *(0.007)*
D(4,4)	context-unaware	0.992 *(0.002)*	0.989 *(0.002)*	0.922 *(0.028)*
	context-aware	0.995 *(0.003)*	0.993 *(0.005)*	0.942 *(0.048)*
D(4,8)	context-unaware	0.961 *(0.011)*	0.961 *(0.011)*	0.943 *(0.012)*
	context-aware	0.984 *(0.004)*	0.982 *(0.006)*	0.976 *(0.007)*
D(5,4)	context-unaware	0.98 *(0.011)*	0.979 *(0.012)*	0.964 *(0.014)*
	context-aware	0.985 *(0.007)*	0.984 *(0.008)*	0.979 *(0.01)*

Table 4: Accuracy means for agents trained in the context-unaware and context-aware setting averaged over five runs with standard deviations.

B. Errors across all datasets

B.1. Errors per game round

These plots show the errors on the validation dataset across all datasets for different concept (# fixed attributes) and context conditions (# shared attributes). Game rounds in which at least one object was incorrectly classified count as errors and are normalized with the number of occurrences of the specific condition in the dataset. This means that a value of 1.0 indicates that listeners incorrectly classified at least one object in each game round in this condition.

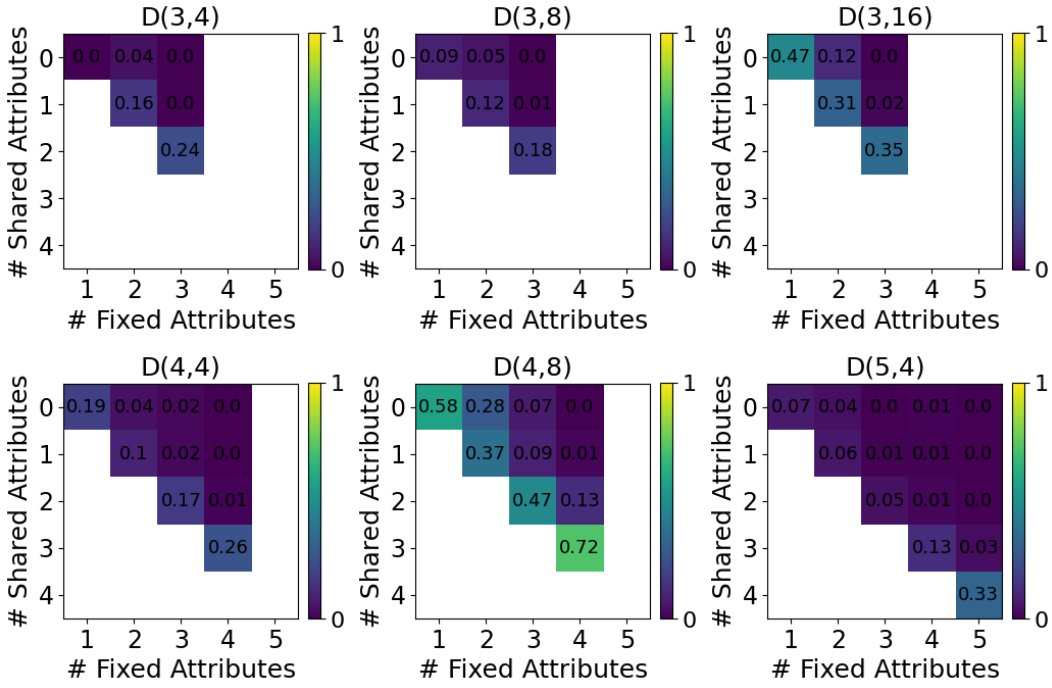

Figure 10: **Context-unaware:** Most errors occur on the diagonal from top left to bottom right, i.e. in the finest possible context conditions.

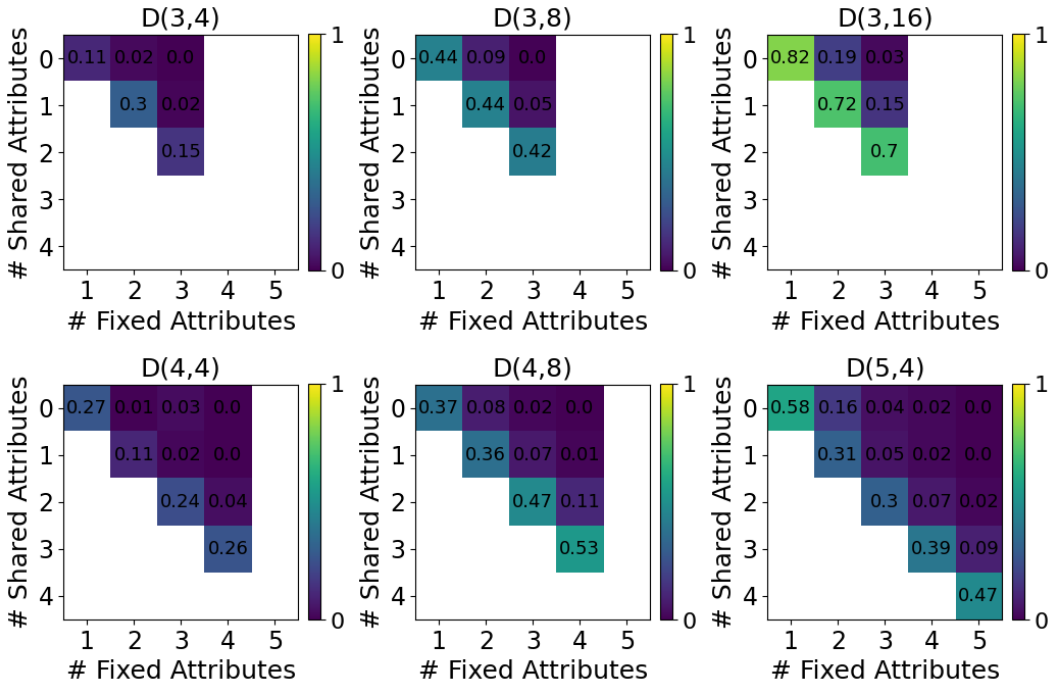

Figure 11: **Context-aware:** Most errors occur on the diagonal from top left to bottom right, i.e. in the finest possible context conditions.

As can be seen in the figures, errors occur mostly in fine context conditions, i.e. where the maximally possible number of attributes is shared between targets and distractors. Some of these errors are false positives, i.e. distractors are incorrectly classified as targets, and some of these errors are false negatives, i.e. targets are incorrectly classified as distractors. We find that false negative errors occur mainly with more generic concepts and fine contexts. This is probably due to the target concepts being very heterogenous and thus, harder to discriminate against distractors. False positive errors, on the other hand, occur in the finest contexts when the concept is very specific. This can be explained by the distractors being very similar to the targets in these conditions. In other words, false positive errors might indicate that the learned target concept is a bit too wide, and false negative errors might indicate that the learned target concept is a bit too narrow.

B.2. False negative errors

Here, we plot object-based false negative errors on the validation dataset across all datasets for different concept (# fixed attributes) and context conditions (# shared attributes). Errors are calculated object-based, i.e. the higher the score the more objects have been incorrectly classified in one game round. These scores are again normalized with the number of occurrences of the specific condition in the dataset. A value of 1.0 indicates that listeners incorrectly classified one target as a distractor in each game round on average in this condition.

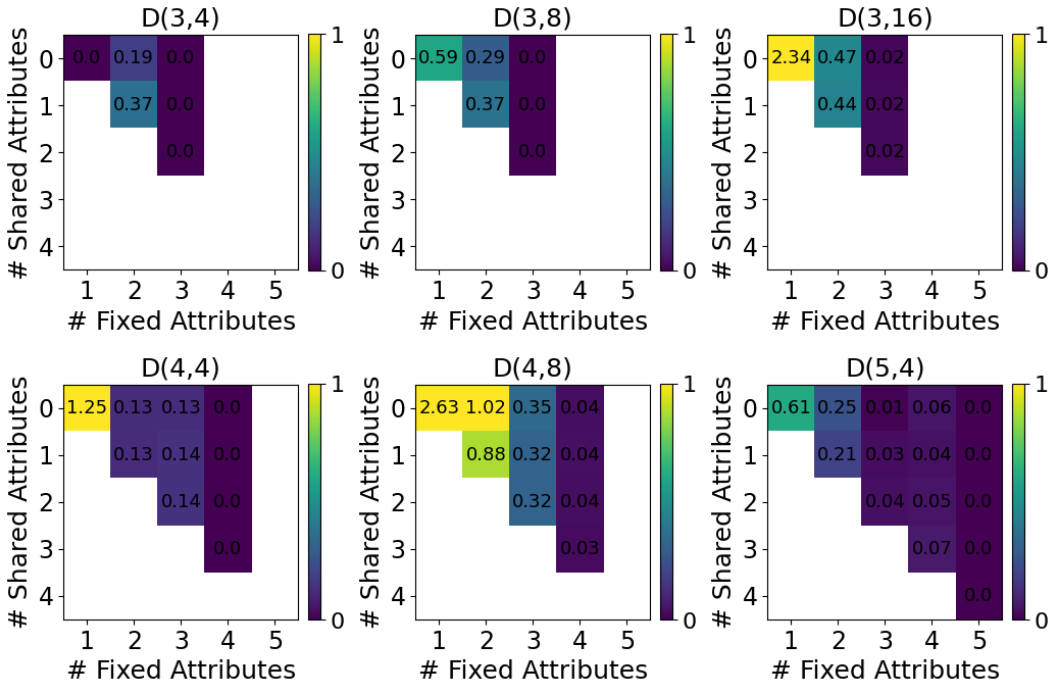

Figure 12: **Context-unaware:** Most false negative errors occur in the top left, i.e. in conditions where generic concepts need to be discriminated in fine contexts.

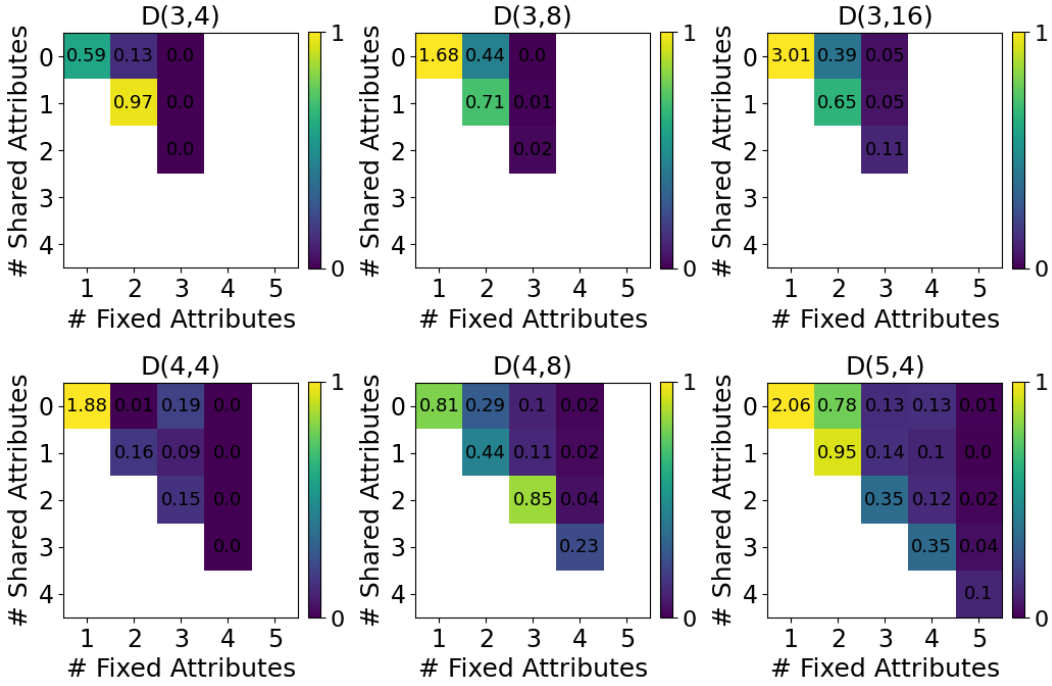

Figure 13: **Context-aware:** Most false negative errors occur in the top left, i.e. in conditions where generic concepts need to be discriminated in fine contexts.

B.3. False positive errors

Here, we plot object-based false positive errors on the validation dataset across all datasets for different concept (# fixed attributes) and context conditions (# shared attributes). Errors are calculated object-based, i.e. the higher the score the more objects have been incorrectly classified in one game round. These scores are again normalized with the number of occurrences of the specific condition in the dataset.

A value of 1.0 indicates that listeners incorrectly classified one distractor as a target in each game round on average in this condition.

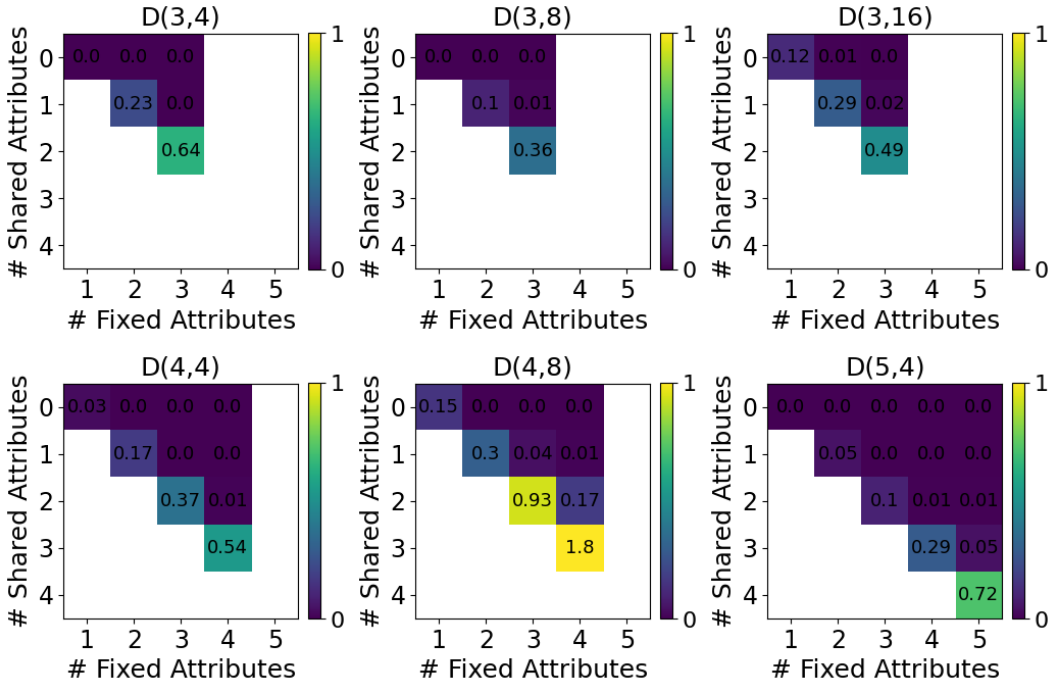

Figure 14: **Context-unaware:** Most false positive errors occur in the bottom right, i.e. in conditions where specific concepts have to be discriminated in a fine context.

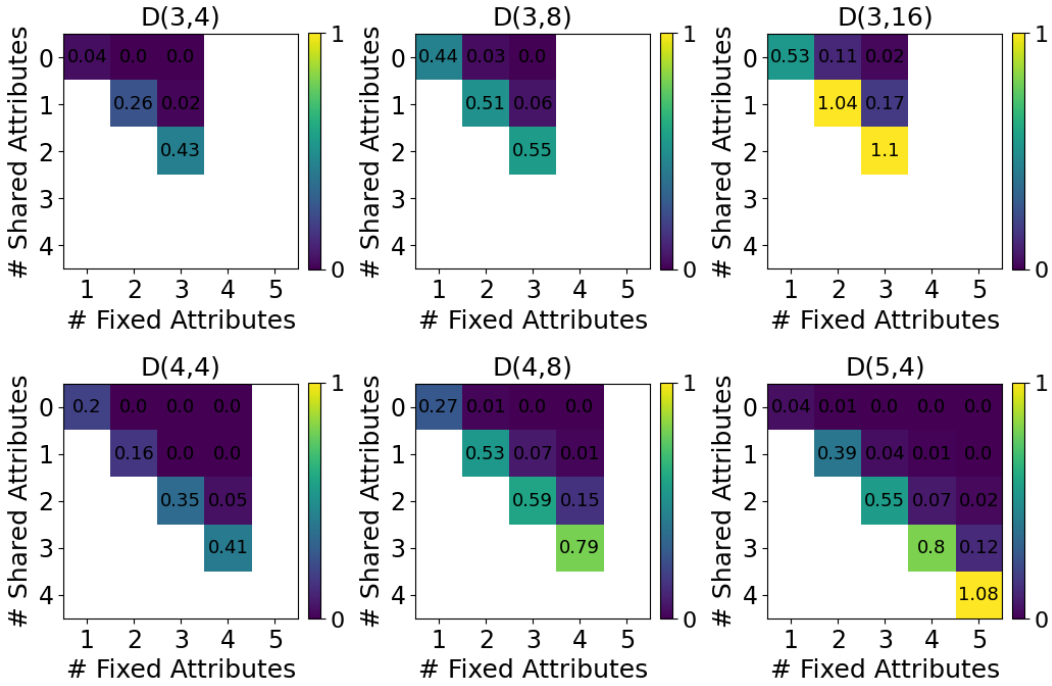

Figure 15: **Context-aware:** Most false positive errors occur in the bottom right, i.e. in conditions where specific concepts have to be discriminated in a fine context.

C. Examples for qualitative results across datasets

C.1. Dataset D(3,4)

Object	Context (# Shared Attributes)	Unique Messages
$[3, 2, 3]$	0	"$[13, 6, 6, 0]$"
	1	"$[13, 6, 6, 0]$"
	2	"$[13, 6, 6, 0]$"

Table 5: **Context-unaware:** Unique messages used to refer to a randomly picked specific concept in the D(3,4) dataset over different context conditions.

Object	Context (# Shared Attributes)	Unique Messages
$[3, 2, 3]$	0	"$[15, 15, 10, 0]$" "$[15, 15, 8, 0]$"
	1	"$[15, 15, 10, 0]$" "$[15, 2, 10, 0]$" "$[15, 2, 6, 0]$"
	2	"$[15, 2, 10, 0]$"

Table 6: **Context-aware:** Unique messages used to refer to a randomly picked specific concept in the D(3,4) dataset over different context conditions.

C.2. Dataset D(5,4)

Object	Context (# Shared Attributes)	Unique Messages
$[3, 0, 2, 1, 3]$	0	"$[5, 4, 12, 7, 2, 0]$"
	1	"$[5, 4, 12, 7, 2, 0]$"
	2	"$[5, 4, 12, 7, 2, 0]$"
	3	"$[5, 4, 12, 7, 2, 0]$"
	4	"$[5, 4, 12, 7, 2, 0]$"

Table 7: **Context-unaware:** Unique messages used to refer to a randomly picked specific concept in the D(5,4) dataset over different context conditions.

Object	Context (# Shared Attributes)	Unique Messages
$[3, 0, 2, 1, 3]$	0	"$[2, 14, 14, 14, 9, 0]$" "$[2, 14, 2, 14, 9, 0]$" "$[2, 14, 7, 14, 9, 0]$" "$[3, 2, 14, 14, 9, 0]$"
	1	"$[2, 14, 14, 7, 9, 0]$" "$[2, 14, 7, 14, 9, 0]$" "$[2, 14, 8, 14, 9, 0]$" "$[3, 2, 14, 14, 9, 0]$" "$[3, 2, 8, 14, 9, 0]$"
	2	"$[13, 8, 14, 14, 9, 0]$" "$[2, 14, 14, 14, 9, 0]$" "$[2, 14, 2, 14, 9, 0]$" "$[3, 2, 8, 14, 9, 0]$"
	3	"$[13, 2, 8, 14, 9, 0]$" "$[3, 2, 8, 14, 9, 0]$"
	4	"$[13, 8, 8, 14, 9, 0]$" "$[3, 2, 8, 14, 9, 0]$"

Table 8: **Context-aware:** Unique messages used to refer to a randomly picked specific concept in the D(5,4) dataset over different context conditions.

D. Mean NMI scores for different concept and context conditions across all datasets

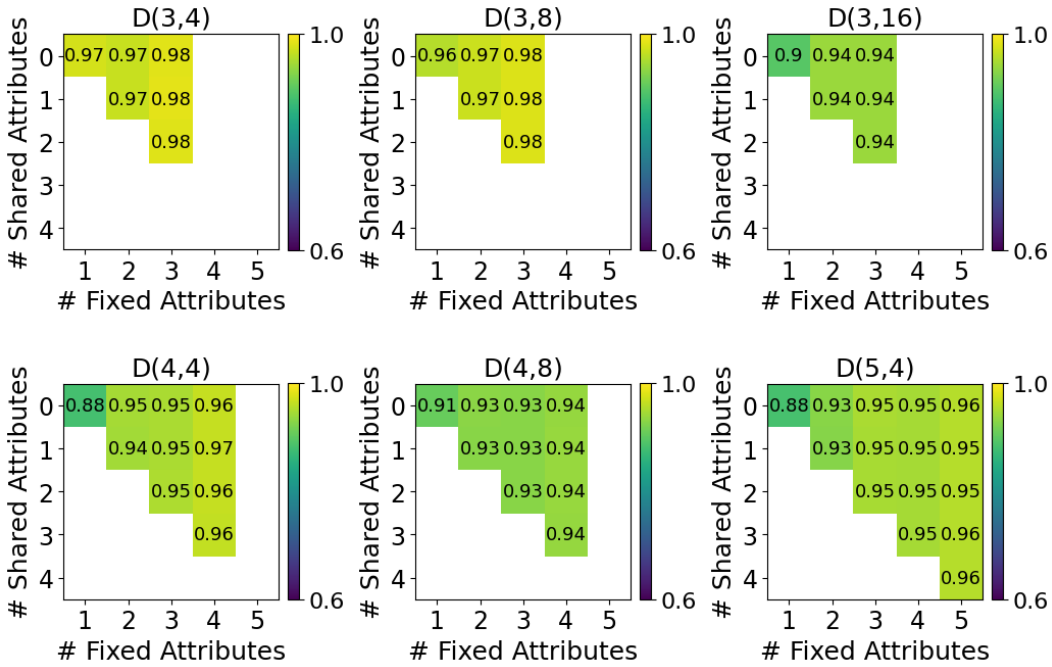

Figure 16: **Context-unaware:** Mean NMI scores across all datasets for different concept (# fixed attributes) and context conditions (# shared attributes). From top to bottom context becomes finer and from left to right concepts become more specific.

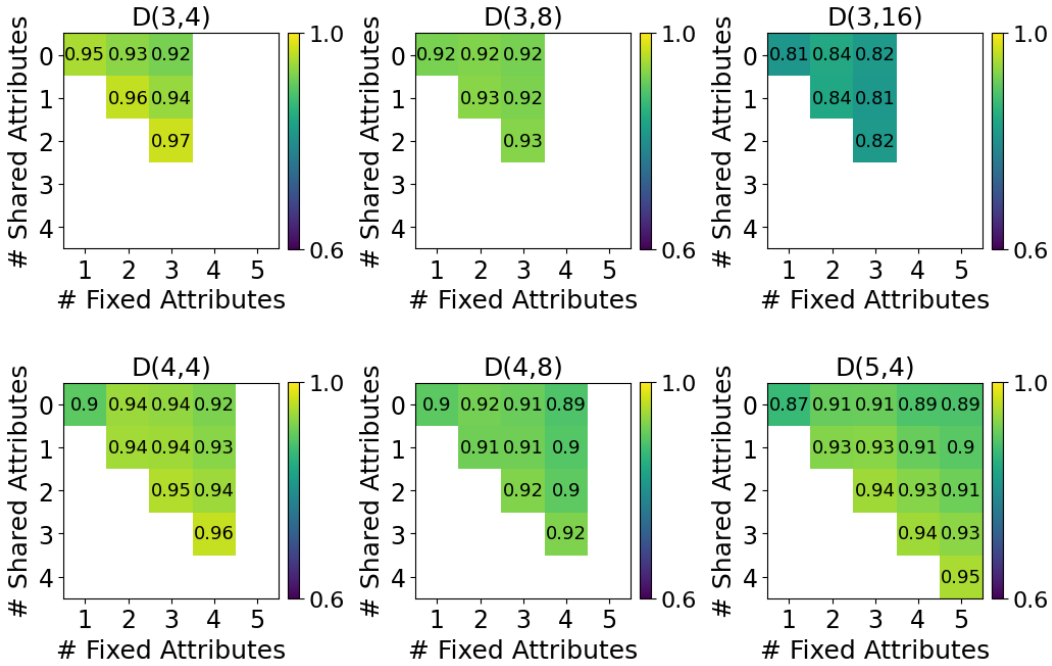

Figure 17: **Context-aware:** Mean NMI scores across all datasets for different concept (# fixed attributes) and context conditions (# shared attributes). From top to bottom context becomes finer and from left to right concepts become more specific.