The Efficiency of Ambiguity: Abstract References Emerge in Coarse Contexts

Anonymous ACL submission

Abstract

We investigate how context granularity, i.e. whether fine or coarse distinctions need to be made, influences an emerging lexicon. We conduct an agent-based simulation of a concept-level reference game, in which agents learn to communicate about concepts that are operationalized by combining multiple objects. We create three experimental conditions by manipulating the context in which the instances of the target concept appear: In the fine context condition, agents must make precise distinctions between similar targets and distractors. In the coarse context condition, targets are easy to discriminate because they share no overlapping features with the distractors. In the mixed baseline condition, both fine and coarse distinctions are necessary. Our results suggest that agents adapt their communication strategies to the granularity of the context in which they learned the concepts. In the fine context and baseline conditions, agents develop a communication protocol heavily based on one-to-one mappings between messages and concepts. Conversely, in the coarse context condition agents communicate more efficiently by vastly relying on abstract references that may refer to more than a single concept but are unambiguous in context. These results show that ambiguity emerges in coarse contexts and that ambiguous abstract terms are used for more efficient communication.

1 Introduction

Context plays a crucial role when communicating information. Not only does the immediate context help to constrain the meaning of an ambiguous utterance, but also recent research suggests that context shapes the formation of lexical conventions (see e.g. Hawkins et al., 2018; Winters et al., 2018). The question of how context granularity, i.e. whether a target needs to be discriminated in a fine or coarse context, influences the emergence of abstract and specific references has been investigated in a recent study with human participants and a small set of hierarchically organized targets (Hawkins et al., 2018). The goal of our current research is to investigate the scalability of their findings to larger conceptual hierarchies and larger lexica by adapting their setup to a language emergence simulation between artificial neural network agents.

Our work is based on two previous lines of research. On the one hand, research on the evolution of artificial languages between human participants has shown that lexical conventions are shaped by communicative pressures, such as the communicative environment and pragmatic demands of context (Nölle et al., 2020; Hawkins et al., 2018; Winters et al., 2018; Silvey et al., 2015; Winters et al., 2015; Tinits et al., 2017). Specifically, Hawkins et al. (2018) found that when fine-grained distinctions are necessary to disambiguate a target from the context, the emerging lexical systems contain more one-to-one mappings between words and meanings. Contrastingly, when such fine-grained distinctions are not necessary, emerging lexical systems contain more abstract references which can be used to refer to more than one object (Hawkins et al., 2018). This line of research makes use of the artificial language learning paradigm (see e.g. Smith and Wonnacott, 2010; Kirby et al., 2008): Two participants play a reference game. The speaker's task is to communicate a target to the listener who has to select the target from a context, i.e. a set of distractor objects. The speaker select their messages from a small set of artificial words or syllables. After several interactions, an artificial lexicon has emerged. In other words, speakers and listeners have converged on the meanings of the artificial words in the context of the reference game.

The task design and setup bears close resemblance to the second line of research we base our work on: In language emergence research, two neu-

ral network agents play a similar reference game and the emerging message-meaning mappings can be investigated for their language-like properties (e.g. Lazaridou et al., 2018). The rapid advancement of artificial intelligence and deep learning techniques has created new opportunities for exploring emergent communication - now between artificial agents instead of human participants. Agent-based simulations provide valuable insights into language's intrinsic properties and evolution by enabling researchers to observe the emergence of communication in a controlled environment, where variables can be precisely manipulated and no prior knowledge exists. Numerous studies have used neural network models to investigate the development of artificial communication protocols through reference games similar to those employed in artificial language learning studies (see e.g. Ohmer et al., 2022; Mu and Goodman, 2021; Lazaridou et al., 2018; Dagan et al., 2021; Bernard et al., 2024). In the reference games employed in these simulations, a sender agent describes a target object to help a receiver agent identify it among distractors. Going beyond the communication of single objects, in Mu and Goodman (2021) and Kobrock et al. (2024), the sender describes groups of objects with shared features, forming a concept that guides the receiver's selection (see also Akkerman et al., 2024, for a different approach to communication about multiple targets). We build on the simulations by Kobrock et al. (2024) where agents learned to communicate about concepts at various levels of abstraction. Their main finding suggests that when agents are provided with contextual information, they take this context information into account when communicating. This leads to the development of a more efficient and natural communication protocol.

The research gap that our study addresses is the successful synthesis between the findings of Hawkins et al. (2018) on the influence of different context granularities (fine vs. coarse distinctions) on human language and the modeling approach by Kobrock et al. (2024) which allows us to investigate whether the results from Hawkins et al. (2018) scale to larger conceptual hierarchies and to larger lexica. Scalability is important to provide further evidence on mechanisms of natural languages because natural languages typically consist of very large lexica and can be used to refer to basically any object or concept. The aim of our study is to investigate how the simple context manipulation as in Hawkins et al. (2018) scales to larger conceptual hierarchies and to a larger set of possible messages. We do this by building on the concept-level reference game simulations from Kobrock et al. (2024) and systematically manipulating the context granularity, i.e. whether agents have to make fine or coarse distinctions between targets and distractors during training. Based on Hawkins et al. (2018), we have the following expectations: Fine-grained contexts will result in precise, one-to-one mappings between messages and concepts. Coarse-grained contexts will encourage abstract references, with fewer one-to-one-mappings but greater lexical efficiency through context-based disambiguation.

2 Methods

2.1 General setup and game scenarios

We extend the framework proposed by Kobrock et al. (2024) to examine the effects of context granularity on the lexicalization of abstract and specific references in an emerging language. Following Hawkins et al. (2018), *specific references* are used to refer to only one specific entity and *abstract references* are references with more than one meaning. In Kobrock et al. (2024), agents iteratively learn to communicate about concepts at different levels of abstraction through a concept-level reference game. Drawing on this foundation, we design two novel game scenarios, systematically manipulating context granularity in similar manner to (Hawkins et al., 2018). In a concept-level reference game $G = (T^S, D^S, T^R, D^R)$ between sender S and receiver R, target concepts $T^S = \{t_1^T, ..., t_g^T\}$ need to be communicated in a given context $D^S = \{d_1^S, ..., d_g^S\}$, where g is the game size, i.e. the number of target and distractor objects in the input. T^R and D^R are defined analogously for the receiver and $T^S \neq T^L$ and $D^S \neq D^R$, as proposed in (Mu and Goodman, 2021). These target concepts are operationalized by combining multiple target objects that share a specific amount of fixed attributes. The number of fixed attributes within a target concept determines how specific (all attributes fixed) or generic (one attribute fixed) a target concept is. The target concepts are presented in a context determined by the distractor objects (Kobrock et al., 2024). In each round of the game, sender S receives targets T^S and distractors D^S, presented in this order. Utilizing this information, S constructs a message $m = (s_j)_{j \leq M}$, where s_j signifies a symbol from

the vocabulary V, and M refers to the maximum length of the message. The receiver R obtains the message m along with their own set of targets T^L and distractors D^R, which are mixed together (denoted subsequently as $X^R = x_1^R, \ldots, x_i^R$, where $i = 2 \cdot g$, reflecting the fact that R is unaware of which items are targets and which are distractors). From these inputs, R generates a prediction for each object x_i^R in its input, producing a label $y_i^R \in 0, 1$ (where 0 indicates a distractor and 1 indicates a target). Following Hawkins et al. (2018), we design two novel game scenarios, a coarse and a fine context condition, in which we train the agents in only coarse or only fine contexts. In the coarse context condition, the objects belonging to the target concept do not share any relevant features with the distractors. In the fine context condition, each distractor differs from the target objects by only one relevant feature. We compare these conditions to a baseline "mixed" condition that was introduced in previous work (see Kobrock et al., 2024)[1]. The sender's task is to produce a message to guide the receiver in the identification of the target concept. The receiver agent must assign a label to each one of the input objects to distinguish them between targets and distractors. The agents are implemented as neural networks and are trained in a reinforcement learning paradigm, where the agents are rewarded depending on whether the receiver agent assigns the correct label (target/distractor) to each one of the input objects.

The code for the experiment is available in the GitHub repository: `https://anonymous.4open.science/r/context-granularity-BE08`

2.2 Datasets

We train the agents on six symbolic datasets which have been introduced in previous work (Kobrock et al., 2024). They are denoted by the number of attributes n and values k that objects in each dataset can take. For example, objects in the dataset D(3,4) have three attributes that can each take four values. We train the agents on datasets D(3,4), D(3,8), D(3,16), D(4,4), D(4,8), and D(5,4) with varying numbers of attributes and values to ensure generalizability of results. The datasets feature hierarchical concepts ranging from specific to generic concepts. The specificity of a concept depends on the number of fixed attribute values. For instance, consider a world where objects are geometrical figures characterized by three attributes: size, color, and shape. The most specific concepts have fixed values for all attributes. Examples of such concepts are SMALL BLUE TRIANGLE and BIG RED SQUARE. In contrast, concepts like RED, SMALL, and SQUARE are the most generic, since objects are required to satisfy only one attribute value to belong to those concepts. Concepts defined in this way are hierarchical: for instance, the object *small blue triangle*: belongs to the concept SMALL BLUE TRIANGLE, but also to the concepts that fix only one or two of those attributes values, such as *small*, *blue*, *small blue*, and *blue triangle*.

The concepts are presented in context. We call objects which are part of the target concept targets and objects which are part of the context distractors. The context for each input is defined by the number of concept-defining attributes (the attributes whose value is fixed in the target concept) that are shared between targets and distractors. Previous work has presented concepts in all possible context conditions ranging from fine, where all but one concept-defining attributes are shared between targets and distractors [2], to coarse, where no concept-defining attribute is shared between targets and distractors (Kobrock et al., 2024). We will call this the 'mixed' context condition and use it as a baseline. To build our novel game scenarios, we generate datasets where target concepts are presented only in fine or only in coarse contexts. This way, consistently with (Hawkins et al., 2018), we create two conditions that differ by the granularity of the context, i.e. the number of concept-defining attributes that must be specified to correctly identify the target objects among the distractors. Hawkins et al. (2018) present a single-target reference game where the coarse context condition presents a target that differs by two attributes from the distractors. Differently, we decided to implement the coarse context condition as the coarsest condition possible, i.e. targets differ by all concept-defining attributes from the distractors. Additionally, we introduce the mixed baseline to capture trials of all sorts, i.e. where distinctions based on $1 - n$ attributes are necessary. Figures 1 and 2 present examples for items in coarse and fine context conditions for the target concept YELLOW TRIANGLE if attributes of the dataset are size, color, and shape. Note that the

[1]The condition we refer to was called "context-aware" in Kobrock et al. (2024).

[2]The concept-defining attribute which is not shared between targets and distractors can vary among distractors. For example, in Figure 2 one distractor differs by shape and the other by color.

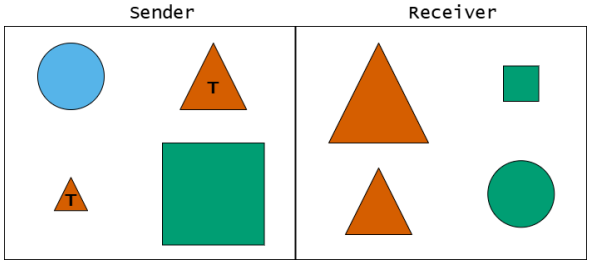

Figure 1: **Coarse context:** The target concept ("T") ORANGE TRIANGLE shares no concept-defining attributes (color and shape) with distractors.

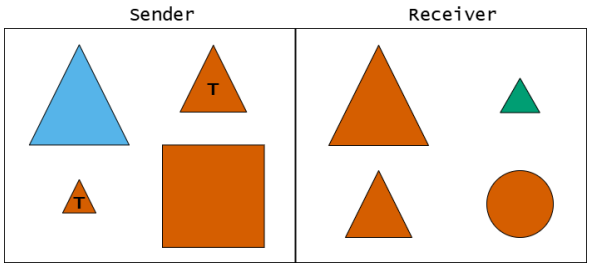

Figure 2: **Fine context:** The target concept ("T") ORANGE TRIANGLE shares all but one of the two concept-defining attributes (color and shape) with distractors.

sender and receiver are given different inputs, but in both inputs, targets are instances of the same concept and distractors fulfill the same context condition (following Mu and Goodman, 2021; Kobrock et al., 2024). For comparison, the dataset used in Hawkins et al. (2018) uses three hierarchical features (shape, color/texture and frequency/intensity) which could take two values at each hierarchy level. However, unlike their dataset, we do not use a fixed hierarchy of stimuli. Rather, all possible combinations of attributes are included in our datasets, and instead of a fixed hierarchy of objects we use concepts.

2.3 Architectures and training

The implementation of the concept-level reference game is based on Kobrock et al. (2024) and makes use of the EGG framework (Kharitonov et al., 2019, MIT license). The sender and receiver are implemented as GRUs (Gated Recurrent Units) with a single layer of size 128 and are trained in a reinforcement learning paradigm with the Gumbel-softmax relaxation[3] which ensures differentiability for backpropagation (Jang et al., 2017). To ensure comparability of our novel fine and coarse context granularity conditions to the mixed baseline from previous work, we adopt the same settings and hyperparameters as in the context-aware condition from Kobrock et al. (2024). The inputs for both the sender and receiver consist of 10 target objects T and 10 distractors D, i.e. game size g is 10. For each item, the sender can produce a message in the form of a vector, with a maximum length $M = n + 1$. Agents use a zero as the End Of Sequence (EOS) symbol. The vocabulary size V, i.e. the number of symbols that sender can use in their messages, is set to $V = 3 \cdot (k + 1)$ (as in Ohmer et al., 2022; Kobrock et al., 2024). The receiver predicts a label $y_i \in \{0, 1\}$ for each object x_i in its input based on whether it believes it to be a target (1) or a distractor (0). We use binary cross-entropy loss for training:

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

where $\hat{m} \sim p^S(m|T^S, D^S)$ and $p^L(y_i^L|x_i^L, \hat{m}) = \sigma(\text{GRU}^L(\hat{m}) \cdot \text{embed}(x_i^L))$. Following (Kobrock et al., 2024), we use 60% of each dataset for training, 20% for validation, and 20% for testing. Each split contains different concepts presented in the relevant novel context condition (coarse or fine). We run the training process five times to account for the random initialization of the parameters of the neural networks and train for 300 epochs with a batch size of 32. The testing split is used only once at the end of the training to evaluate the agents' generalization capabilities on unseen concepts, while the validation split is used to measure performance after every training epoch.

2.4 Metrics

We first measure the training, validation, and test accuracies for all context conditions. We then evaluate the emerging languages based on three entropy-based scores calculated on the set of messages M and the set of concepts C (Ohmer et al., 2022) and take the means over five runs. These are:

- **Normalized Mutual Information (NMI)**, which measures how closely messages and concepts correspond to each other in a one-to-one relationship. Therefore, if the NMI score takes its maximal value 1.0, it is possible to construct a bijective map between the set of messages produced by the agents and the set of target concepts presented to them in the game. The NMI can take values $\in [0, 1]$ and is calculated as follows:

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

[3] We use temperature $\tau = 2$ and a decay rate of 0.99.

- **Effectiveness**, which quantifies the usage of messages that uniquely identify a concept, in other words messages are non-polysemous or not abstract. Effectiveness can take values $\in [0, 1]$ and is calculated as:

$$\text{effectiveness}(C, M) = 1 - \frac{H(C|M)}{H(C)} \quad (3)$$

- **Consistency**, which measures whether agents are consistent in choosing always the same message to communicate the same concept, in other words messages are non-synonymous. Consistency can take values $\in [0, 1]$ and is calculated as:

$$\text{consistency}(C, M) = 1 - \frac{H(M|C)}{H(M)} \quad (4)$$

To quantify differences in scores between the different conditions, we use Bayesian estimation following Kruschke (2013). We report estimated means and 95% Credible Intervals (CrIs) over five runs as well as mean differences and their CrIs.

We then perform a more qualitative analysis of the communication protocols by mapping each message to the concept(s) it was used to refer to.

2.5 Hypotheses

We hypothesize that the communication protocol developed by the agents is influenced by the context condition in which it is learned. We expect that in the fine context condition, agents will develop more one-to-one mappings (i.e., higher NMI) between messages and concepts than in the mixed baseline. In other words, we expect only specific references to be lexicalized, since the context condition requires fine-grained distinctions between concepts. In the coarse context condition, on the other hand, we expect agents to develop fewer one-to-one mappings (i.e., lower NMI) between messages and concepts than in the mixed baseline. In other words, we expect agents to include more abstract references in their communication, as the contextual information can be leveraged by agents to identify the target concept. Following this, we expect higher effectiveness in the fine context condition than in the baseline and lower effectiveness in the coarse context condition than in the baseline. Lower effectiveness in the coarse context condition with the same accuracy would indicate that agents use words with many meanings, or abstract references in the terminology used in Hawkins et al. (2018). These predictions are in line with the results obtained in Hawkins et al. (2018), who found that abstract references were lexicalized only in the coarse context condition. We expect high overall consistency across context conditions, reflecting minimal use of synonymous references. This aligns with Kobrock et al. (2024), who noted that agents use synonyms for the same concept in varying contexts—a scenario not applicable here since context conditions remain constant across trials.

3 Results

3.1 Performance and generalization

Agents achieve very good performance on training and validation sets in both fine and coarse context conditions. Mean train and validation accuracies across runs are ≥ 0.94 for the fine context condition and ≥ 0.99 for the coarse context condition for all datasets. These results are comparable to the ones obtained in (Kobrock et al., 2024) for the baseline mixed condition (mean train and validation accuracies across runs for all datasets ≥ 0.96). These results are an indication that agents are able to learn to communicate about concepts in the setting of the concept-level reference game also when trained only in fine or coarse contexts.

Accuracies on the test split differ more across context conditions. The mean test accuracies across runs are 0.98 (SD=0.02) for the coarse context condition, and 0.75 (SD=0.14) for the fine context condition indicating that while agents can also reasonably well generalize in the fine context condition, they are much better at generalizing when they have been trained in the coarse context condition. The mean test accuracy in the baseline condition is 0.87 (SD=0.11) which is in the middle between the fine and coarse condition. This shows that the generalization abilities of the agents depends on the context condition in which they have been trained: When being trained to make fine distinctions, agents come up with a mapping that does not generalize well. On the other and, when being trained to make coarse distinctions only (or a mix), then agents come up with a mapping that generalizes better.

3.2 Contextual pressures shape the emerging language

To obtain information on the emerging language at its final stage of development, information-theoretic scores are calculated on the interactions

of the last training epoch (see Kobrock et al., 2024; Ohmer et al., 2022). First, we look at the NMI scores. The NMI scores averaged across runs for all datasets and context conditions are summarized in Figure 3.[4] For the mixed context condition, the mean NMI is estimated at M=0.87 [0.85, 0.89]. For the fine context condition, the mean NMI is M=0.92 [0.90, 0.93]. There seems to be a small difference between the fine and mixed context condition, where the NMI scores are higher for the fine context condition. To quantify this difference statistically, we used Bayesian estimation following Kruschke (2013). We find a substantial, though very small, difference in NMI between the fine condition and our mixed baseline (M=0.05, CrI[5]=[0.02, 0.07], pd=100%, 1% in ROPE). For the coarse context condition, the mean NMI is M=0.59 [0.58, 0.61]. The difference between the coarse context condition and our mixed baseline is quite large and substantial (M=-0.28, CrI=[-0.30, -0.25], pd=100%, 0% in ROPE). In summary, we find that NMI scores in the fine and coarse context condition differ significantly from the mixed condition baseline with NMIs in the fine context condition being slightly higher than the baseline and NMIs in the coarse context condition being much lower than the baseline. This suggests that agents tend to create one-to-one mappings between concepts and messages, assigning to each concept its own message in the fine context condition. Agents likely adopt this communication strategy because they are required to draw very fine-grained distinctions to correctly discriminate the targets from the distractors in the fine context condition. In the coarse context condition, on the other hand, the low NMI scores indicate that the communication protocol emerging from the coarse context condition is not based on a strict one-to-one correspondence between concepts and messages.

Second, we look at the effectiveness score. For the mixed context baseline, the estimated mean effectiveness is M=0.88 [0.85, 0.91]. For the fine context condition, the mean effectiveness is M=0.90 [0.87, 0.94]. The difference in effectiveness between the fine and mixed context conditions is not substantial (M=0.02, CrI=[-0.02, 0.07], pd=85%, 46% in ROPE). For the coarse context condition, the mean effectiveness is M=0.44 [0.42, 0.46]. The difference between the coarse and the mixed context condition is large and substantial (M=-0.44, CrI=[-0.47, -0.40], pd=100%, 0% in ROPE). The low effectiveness scores in the coarse context condition indicate that agents use the same message to identify multiple target concepts, i.e. they use abstract references. One possible cause for these results is that agents might rely on context to clarify which concept they are referring to, rather than using one message for each concept regardless of the context. This is possible in the coarse context condition: As targets and distractors are very different from each other and share no attributes, a message that does not encode all information about the target concept (i.e. an abstract message) can still be informative enough to correctly identify the target objects.

Third, we look at the consistency scores. For the mixed context condition, the estimated mean consistency is M=0.87 [0.86, 0.88]. For the fine context condition, consistency is estimated at M=0.94 [0.93, 0.95]. The difference in consistency between the fine and mixed context condition is very small but substantial (M=0.07, CrI=[0.06, 0.09], pd=100%, 0% in ROPE). For the coarse context condition, mean consistency is M=0.92 [0.90, 0.94]. The difference in consistency between the coarse and mixed context conditions is small and substantial (M=0.05, CrI=[0.03, 0.08], pd=100%, 0% in ROPE). The values for consistency are comparable to the ones of the fine context condition, indicating that agents make very limited use of synonymy in the coarse context condition, too. In both the fine and the coarse context condition, we observe higher consistency scores than in the mixed baseline which suggests that agents come up with non-synonymous mappings, or, in other words, there are no concepts that can be referred to with more than one distinct message.[6]

3.3 Coarse contexts drive the emergence of smaller and more efficient lexica

To further investigate the communicative strategy adopted by agents in the coarse and fine context condition, we reconstruct the emergent lexica (i.e. the mappings between messages and concepts) from the interactions of the last training epoch. A description of the methodology we used, along with

[4]Plots for effectiveness and consistency scores can be found in appendix B.

[5]The Credible Interval (CrI) was estimated as the 95% Highest Density Interval.

[6]The observed differences might also vary with respect to the conceptual hierarchy. For this reason, in Appendix C, we analyze the entropy scores depending on the level of specificity of the concepts.

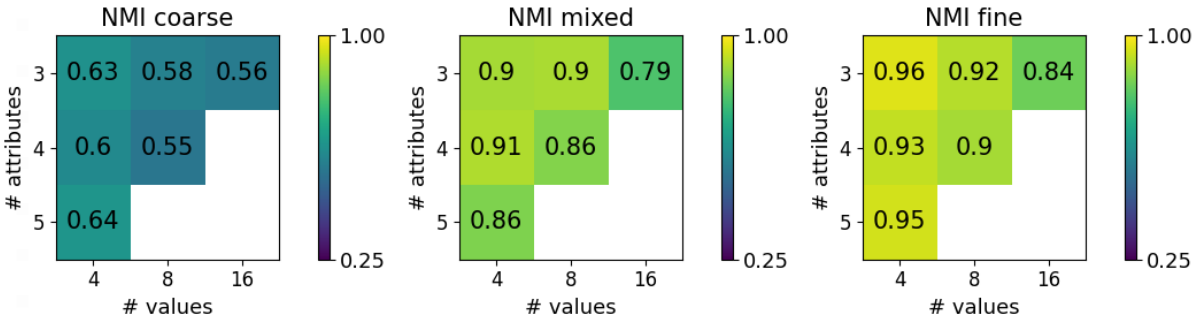

Figure 3: Mean NMI scores for coarse, mixed and fine context condition for each dataset. The different datasets are identified by the number for attributes and the number of possible values each of those attributes can take.

some examples, can be found in Appendix D.

Dataset	# Concepts	Context Condition	# Messages
D(3,4)	99	Coarse	33
		Fine	133
D(3,8)	582	Coarse	41
		Fine	438
D(3,16)	3929	Coarse	133
		Fine	1308
D(4,4)	499	Coarse	204
		Fine	1071
D(4,8)	5248	Coarse	788
		Fine	6636
D(5,4)	2499	Coarse	1443
		Fine	4337

Table 1: Lexicon sizes for all datasets in coarse and fine context conditions. We present the number of unique concepts and the number of unique messages.

A comparison of the sizes of the lexica can be found in Table 1. Across all datasets, agents converge to smaller lexica in the coarse context condition, where agents consistently lexicalize fewer messages than the total number of target concepts. By doing so, they appear to develop an efficient communication protocol that avoids encoding each concept into a unique message, but likely leverages contextual information to disambiguate the target concept. Conversely, in the fine context condition, agents make use of a bigger set of messages. A possible explanation for this is that agents in the fine context condition create specific messages to be able to unambiguously identify the target concept among similar distractors.

How do agents in the coarse context condition achieve this efficient mapping? Figure 4 shows for each message the number of concepts that it refers to, normalized by the total number of concepts in each dataset and condition. A high concept ratio means that this message refers to many different concepts in a dataset, and a ratio close to 0 means that this message refers to only one concept in the dataset. Agents in the fine context condition use messages that refer to one or only very few concepts, while in the coarse context condition agents incorporate abstract messages and use them to refer to a bigger set of concepts. Those messages are the ones mapped to most generic concepts. This is particularly interesting, as it indicates that agents in the coarse context condition rely on messages that encode minimal information (i.e. one attribute value corresponding to a generic concept) also to identify more specific concepts in certain contexts.

4 Discussion

This work employs an agent-based simulation of a concept-level reference game to investigate the influence of context granularity on the emergent communication protocol. We compare our findings to the results obtained in the artificial language learning study with human participants and comparable context conditions (Hawkins et al., 2018).

First, we observe higher accuracies for the coarse context condition than the fine context condition with the mixed baseline in between in both performance (i.e. on the train and validation sets) and generalization (on the test set). The superior performance of agents in the coarse context and their lower performance in the fine context align with the findings of Hawkins et al. (2018), who observed more correct responses by the listeners in the coarse than in the fine context condition. They explain this by the fine condition being hardest and the coarse condition being easiest, with the mixed condition in between. Similarly, the difficulty of the task may explain the differences in performance we observe in the agent-based simulation. While exhibiting similar performance on the train and validation sets, agents trained in the coarse context

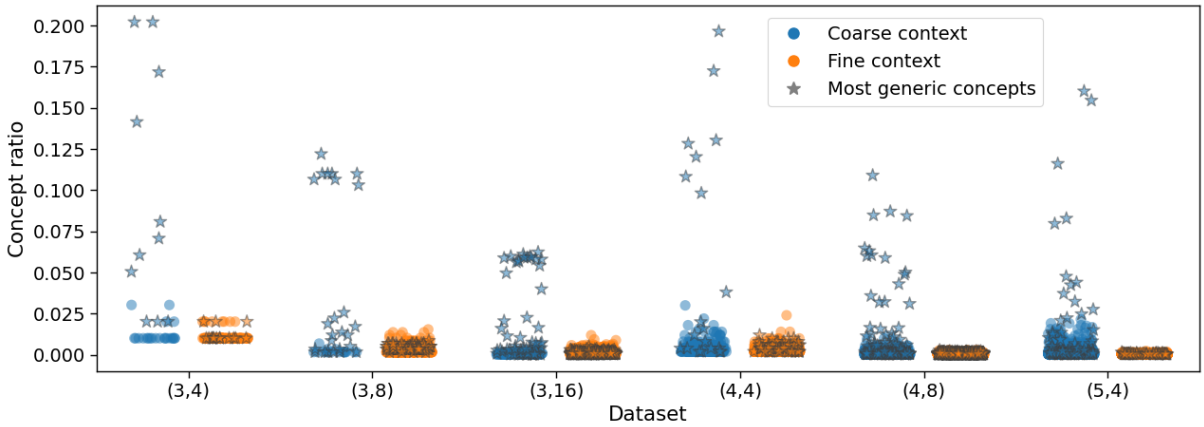

Figure 4: Concept coverage in the messages: Number of concepts referred to by each message normalized by the total number of concepts across datasets and conditions. Messages referring to most generic concepts (i.e. the ones with only one fixed attribute) are identified by the star marks.

condition generalize better than agents trained in the fine context condition or in the mixed baseline condition. As Hawkins et al. (2018) did not test generalization, this is a novel finding and suggests that languages that have been shaped by coarse contexts serve better for generalization. Second, the granularity of the context plays an important role in shaping the emerging language. In line with our hypotheses, agents trained in the fine context condition tend to associate each concept with a unique message, and we do not find many synonymous or polysemous messages in their protocols, whereas agents in the coarse context condition exhibit fewer one-to-one mappings and tend to reuse the same message for multiple target concepts. The use of more specific references, i.e. references with a single meaning, in the fine context condition and more abstract references, i.e. references with more than one meaning or polysemous references, is in line with the results obtained in Hawkins et al. (2018). These abstract references seem to be the reason for the better generalization abilities in the coarse context condition due to the fact that polysemous messages foster flexibility and hence allow for a better adaption to newly encountered concepts. Polysemy in human language has been argued to make human languages particularly efficient (e.g. Piantadosi et al., 2012). Third, we have shown that these abstract references that emerge only in the coarse context condition are used to refer to a very high number of concepts in the datasets. Agents in the coarse context condition seem to develop a communication protocol that extensively employs abstract messages, allowing them to correctly identify target concepts, including specific ones, by relying on contextual information. This suggests that agents can develop a more efficient communication strategy, which goes beyond mere one-to-one mappings between messages and concepts, when the context allows it. Importantly, agents did not have any external incentive such as a regularization cost or efficiency pressure to modify their communication based on the granularity of the context. Therefore, the adaptation to the granularity of the context appears to be an emergent feature: The manipulation of the context alone drives the agents to a more efficient communication strategy.

In conclusion, the results of this study suggest that the granularity of the context significantly influences the development of emerging languages. Specifically, agents can leverage contextual information to create more effective and efficient communication protocols that utilize abstract references. These results are in line with but go beyond previous work on human language by scaling the number of possible referents (dataset size) and the number of possible messages (vocabulary size and message length, Hawkins et al., 2018). The effectiveness of abstract references increases with scaling: In our setup, agents use abstract references to refer to up to one fifth of the entire dataset. Viewed through the lens of pragmatic inference and cost-efficiency, our results reveal how context-driven conventions can balance communicative precision with pressures for lexical economy. Our findings also connect to ongoing theoretical discussions in emergent communication research about efficiency principles and information bottlenecks, highlighting the role of context-based conventions on efficient lexical choices (Gualdoni et al., 2024).

5 Limitations

This work has been carried out on symbolic datasets with hierarchical concepts. Future work could investigate whether these findings generalize to more natural datasets. One possible limitation of our setup is that dataset sizes vary between conditions. Future work could use a sampling approach to make sure that dataset sizes are comparable between conditions. Future work could explore a similar concept-level reference game and manipulate other dynamics of context additional to fine and coarse only. For example, our setup could be used to investigate what happens if the contextual complexity is not a binary of extremes (fine vs. coarse) but rather continuously varied. Another idea would be to look at what happens if the training environment's distribution of context types is skewed. Additionally, constraints that encourage the use of fewer messages or shorter messages could be implemented to try to promote the emergence of an even more efficient language.

Acknowledgments

References

Daniel Akkerman, Phong Le, and Raquel G Alhama. 2024. The Emergence of Compositional Languages in Multi-entity Referential Games: from Image to Graph Representations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18713–18723. Association for Computational Linguistics.

Timothée Bernard, Timothee Mickus, and Hiroya Takamura. 2024. The emergence of high-level semantics in a signaling game. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 200–211, Mexico City, Mexico. Association for Computational Linguistics.

Gautier Dagan, Dieuwke Hupkes, and Elia Bruni. 2021. Co-evolution of language and agents in referential games. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2993–3004, Online. Association for Computational Linguistics.

Eleonora Gualdoni, Mycal Tucker, Roger Levy, and Noga Zaslavsky. 2024. Bridging semantics and pragmatics in information-theoretic emergent communication. *Society for Computation in Linguistics*, 7(1).

Robert D. Hawkins, Michael Franke, Kenny Smith, and Noah D. Goodman. 2018. Emerging abstractions: Lexical conventions are shaped by communicative context. *Cognitive Science*.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax.

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

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

Kristina Kobrock, Xenia Isabel Ohmer, Elia Bruni, and Nicole Gotzner. 2024. 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.

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

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 (ICLR)*.

Jesse Mu and Noah Goodman. 2021. Emergent communication of generalizations. *Advances in neural information processing systems*, 34:17994–18007.

Jonas Nölle, Riccardo Fusaroli, Gregory J. Mills, and Kristian Tylén. 2020. Language as shaped by the environment: linguistic construal in a collaborative spatial task. *Palgrave Communications*, 6(1):1–10.

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, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Catriona Silvey, Simon Kirby, and Kenny Smith. 2015. Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive Science*, 39(1):212–226.

Kenny Smith and Elizabeth Wonnacott. 2010. Eliminating unpredictable variation through iterated learning. *Cognition*, 116(3):444–449.

Peeter Tinits, Jonas Nölle, and Stefan Hartmann. 2017. Usage context influences the evolution of overspecification in iterated learning. *Journal of Language Evolution*, 2(2):148–159.

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.

A Computational resources and dataset sizes

We ran the simulations for the fine and coarse context condition using an NVIDIA Tesla T4 GPU. The whole training process lasted about 4 days.

The dataset sizes can be inspected in Table 2. We report the number of samples for each dataset and condition for train, validation and test data split. Datasets for the fine and coarse context condition have the same sizes. They present each concept in either a fine or coarse context condition, respectively. In the mixed baseline, however, the concepts were presented in each possible context condition, leading to larger datasets (Kobrock et al., 2024).

B Effectiveness and consistency heatmaps

For purposes of visualization, Figures 5 and 6 display the effectiveness and consistency scores for different context conditions and datasets. They visually support the main findings reported in section 3.2.

Dataset	# Concepts	Context Condition	# Train	# Validation	# Test
D(3,4)	99	Fine/Coarse	742	248	250
		Mixed	1852	618	530
D(3,8)	582	Fine/Coarse	4364	1456	1460
		Mixed	11654	3886	3900
D(3,16)	3929	Fine/Coarse	29467	9823	9830
		Mixed	83294	27766	27660
D(4,4)	499	Fine/Coarse	3742	1248	1250
		Mixed	11917	3973	4110
D(4,8)	5248	Fine/Coarse	39359	13121	13120
		Mixed	139957	46653	46670
D(5,4)	2499	Fine/Coarse	18742	6248	6250
		Mixed	74962	24988	25050

Table 2: Dataset sizes for all datasets in coarse and fine context conditions. We present the number of unique concepts and the number of samples in the train, validation and test splits.

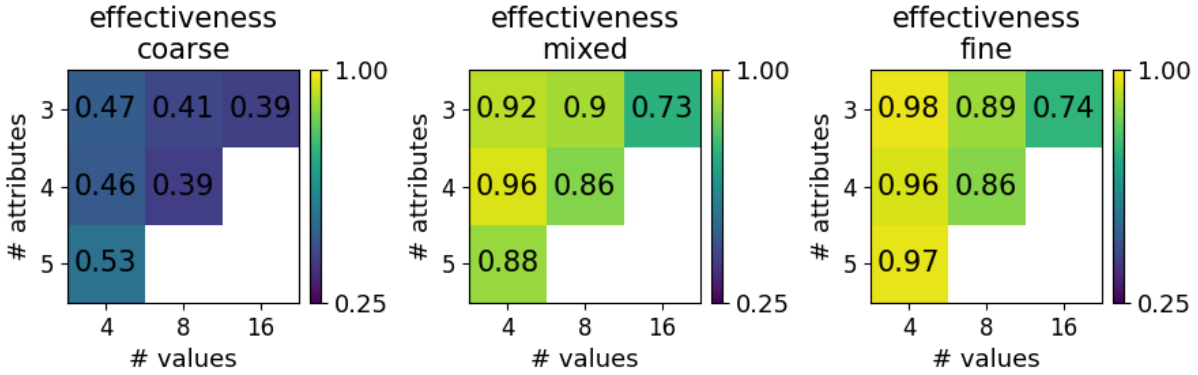

Figure 5: Mean effectiveness scores for coarse, mixed and fine context condition for each dataset. The different datasets are identified by the number for attributes and the number of possible values each of those attributes can take.

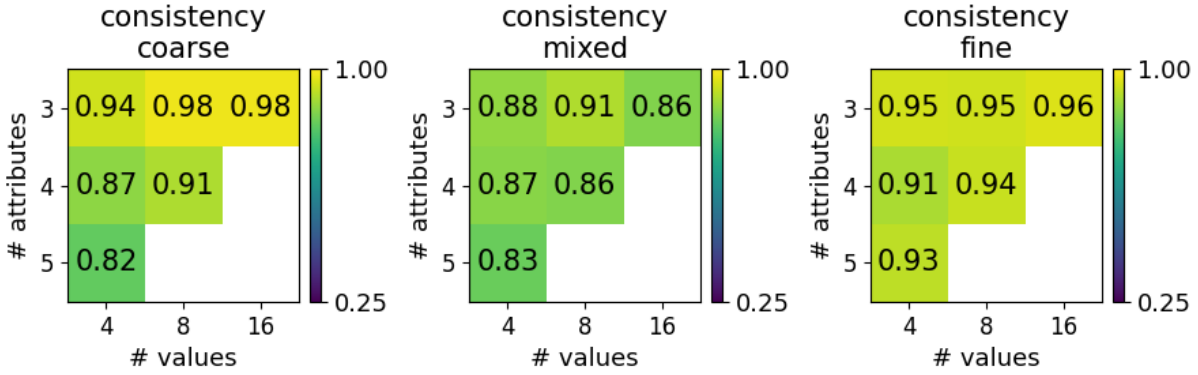

Figure 6: Mean consistency scores for coarse, mixed and fine context condition for each dataset. The different datasets are identified by the number for attributes and the number of possible values each of those attributes can take.

C Analysis of entropy scores by concept level

We investigate how the specificity of the target concept influences the entropy-based scores. Figure 7 reports mean entropy scores across all datasets plotted against concept specificity.

In the mixed baseline, we have not enough evidence for a substantial difference in NMI scores between specific and generic concepts (M=-0.03, CrI=[-0.07, 0.02], pd=91.5%, 19% in ROPE).

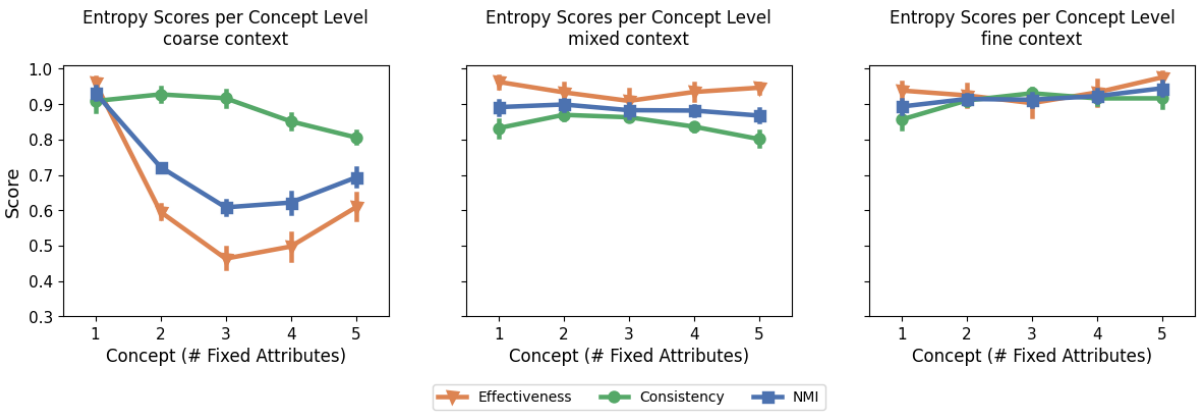

Figure 7: Mean entropy scores across all datasets for different concept levels in coarse and fine context conditions. The specificity of the concepts is indicated by the number of fixed attributes. The more attributes are fixed, the more specific are the concepts. Error bars indicate bootstrapped 95% confidence intervals.

Rather, the NMI score seems to be relatively constant throughout the different levels of concept specificity (see Figure 7, in line with Kobrock et al., 2024). In the fine context condition, we observe that the NMI increases with increasing concept specificity. The mean NMI score for the most general concepts is 0.89 [0.88, 0.91], while the mean NMI for the most specific concepts rises to 0.95 [0.92, 0.96] for concepts with five fixed attributes. The difference in NMI between specific and generic concepts is small but substantial (M=0.05, CrI=[0.01, 0.09], pd=97.6%, 3% in ROPE). This indicates that high concept specificity enforces the tendency of agents to build one-to-one mappings between concepts and messages when the context is fine. In the coarse context condition, we observe the opposite trend: With increasing concept specificity, NMI drops and we find a substantial difference in NMI between specific and generic concepts (M=-0.23, CrI=[-0.29, -0.18], pd=100%, 0% in ROPE). The lowest mean NMI score is observed for concepts with three fixed attributes, at 0.61 [0.59, 0.63]. A possible explanation for these fluctuations in entropy scores in the coarse context condition is that agents might initially create a single message for each of the most general concepts (with only one fixed attribute value). Subsequently, they may reuse these messages to refer to more specific concepts and rely on contextual cues to disambiguate the targets. The slight drop in consistency for concepts with four and five attribute values suggests that agents utilize multiple messages to refer to the same concept. This behavior could be attributed to the flexibility agents have in specifying any attribute to identify a concept in the coarse context condition. As concepts accumulate more fixed attribute values, agents have a broader range of messages at their disposal to identify these concepts.

D Example Lexica

We chose to reconstruct the lexica from the agents' interactions of the last traning epoch for the coarse and fine context conditions, which are the main objects of interest in this study. We proceeded as follows for each dataset. We began by retrieving the message used by the sender in each trial and associating it with the target concept for that trial. For each message, this process yielded a corresponding list of target concepts. Next, we compared the concepts within each list (i.e., the concepts referred to by the same message) to determine whether they satisfied a more generic concept. If this was the case for 90% or more of the concepts in the list, we assigned that more generic concept as the meaning of the message. For instance, imagine a message *m* referring to the concepts BLUE TRIANGLE and BLUE SQUARE: in this case we would establish *m* to be an ending for the concept BLUE. Table 3 displays an example of the complete lexicon for $D(3, 4)$ in the coarse condition, while Table 4 presents a sample (for space reasons) of the lexica emerged in the fine context condition for the same dataset.

The first column of each lexicon table (Tables 3 and 4) contains the messages sent by the agents, while the second column lists the symbols used in each message (excluding the EOS symbol 0). The third column indicates the number of unique concepts referenced by the corresponding message. The fourth column shows the meaning of the message, i.e., the concept encoded by the message, which was reconstructed as explained above. The question marks represent unfixed attributes. Concepts with all but one question mark (i.e., those with only one fixed attribute value) are the most generic concepts in that dataset.

Message	Symbols	# Referred Concepts	Encoded Concept
(1, 1, 1, 0)	{1}	7	(?, 3, ?)
(1, 1, 8, 0)	{8, 1}	1	(?, 3, 3)
(1, 1, 12, 0)	{1, 12}	1	(3, 3, 1)
(1, 2, 2, 0)	{1, 2}	1	(3, 3, 0)
(1, 7, 1, 0)	{1, 7}	1	(3, 3, 2)
(1, 12, 1, 0)	{1, 12}	1	(3, 3, 1)
(1, 12, 12, 0)	{1, 12}	1	(?, 3, 1)
(2, 1, 2, 0)	{1, 2}	1	(?, 3, 0)
(2, 2, 2, 0)	{2}	5	(?, ?, 0)
(3, 3, 3, 0)	{3}	8	(?, 1, ?)
(3, 6, 6, 0)	{3, 6}	1	(2, 1, ?)
(3, 9, 9, 0)	{9, 3}	1	(0, 1, 3)
(4, 4, 4, 0)	{4}	17	(1, ?, ?)
(6, 6, 6, 0)	{6}	14	(2, ?, ?)
(6, 6, 15, 0)	{6, 15}	2	(2, 0, ?)
(6, 15, 6, 0)	{6, 15}	3	(2, 0, ?)
(6, 15, 15, 0)	{6, 15}	1	(2, 0, 2)
(7, 1, 1, 0)	{1, 7}	1	(3, 3, 2)
(7, 7, 7, 0)	{7}	2	(3, ?, ?)
(8, 1, 8, 0)	{8, 1}	1	(?, 3, 3)
(8, 8, 1, 0)	{8, 1}	1	(?, 3, 3)
(8, 8, 8, 0)	{8}	2	(?, ?, 3)
(9, 9, 3, 0)	{9, 3}	1	(0, 1, 1)
(9, 9, 9, 0)	{9}	20	(0, ?, ?)
(9, 9, 15, 0)	{9, 15}	1	(0, 0, 2)
(9, 15, 9, 0)	{9, 15}	1	(0, 0, 2)
(12, 1, 12, 0)	{1, 12}	1	(?, 3, 1)
(12, 12, 12, 0)	{12}	2	(?, ?, 1)
(14, 14, 14, 0)	{14}	20	(?, 2, ?)
(15, 6, 6, 0)	{6, 15}	3	(2, 0, ?)
(15, 9, 9, 0)	{9, 15}	2	(0, 0, ?)
(15, 9, 15, 0)	{9, 15}	1	(0, 0, ?)
(15, 15, 15, 0)	{15}	6	(?, 0, ?)

Table 3: Example lexicon for the language emerging from the concept-level reference game in the coarse context condition with the $D(3, 4)$ dataset.

Message	Symbols	# Referred Concepts	Encoded Concept
(1, 1, 1, 0)	{1}	1	(2, 0, ?)
(1, 1, 2, 0)	{1, 2}	1	(2, 0, 2)
(1, 1, 4, 0)	{1, 4}	1	(2, 0, 3)
(1, 1, 11, 0)	{1, 11}	1	(2, 0, 3)
(1, 1, 13, 0)	{1, 13}	1	(2, 0, 2)
(1, 1, 14, 0)	{1, 14}	1	(2, 0, ?)
(1, 2, 13, 0)	{1, 2, 13}	1	(1, 0, 2)
(1, 13, 2, 0)	{1, 2, 13}	1	(1, 0, 2)
(1, 13, 13, 0)	{1, 13}	1	(1, 0, 2)
(1, 14, 14, 0)	{1, 14}	2	(1, 0, ?)
(1, 15, 1, 0)	{1, 15}	1	(2, 0, 1)
(1, 15, 14, 0)	{1, 14, 15}	1	(1, 0, 1)
(2, 2, 1, 0)	{1, 2}	1	(1, ?, 2)
(2, 2, 2, 0)	{2}	2	(1, ?, 2)
(2, 2, 10, 0)	{2, 10}	1	(1, ?, 2)
(2, 2, 13, 0)	{2, 13}	1	(1, 1, 2)
(2, 4, 4, 0)	{2, 4}	1	(2, ?, 3)
(2, 10, 4, 0)	{2, 10, 4}	1	(1, ?, 3)
(3, 3, 8, 0)	{8, 3}	1	(3, 1, ?)
(3, 3, 11, 0)	{11, 3}	1	(3, 1, 2)
(3, 3, 12, 0)	{3, 12}	1	(3, 1, 1)
(3, 3, 13, 0)	{3, 13}	1	(3, 1, 2)
(3, 5, 3, 0)	{3, 5}	1	(0, 1, ?)
(3, 5, 5, 0)	{3, 5}	1	(0, 1, ?)
(3, 5, 12, 0)	{3, 12, 5}	1	(0, 1, 1)
(3, 12, 12, 0)	{3, 12}	1	(3, 1, 0)
(3, 13, 4, 0)	{3, 4, 13}	1	(0, 1, 3)
(3, 13, 5, 0)	{5, 3, 13}	1	(0, 1, 2)
(4, 3, 4, 0)	{3, 4}	1	(?, ?, 3)

Table 4: Sample from an example lexicon for the language emerging from the concept-level reference game in the fine context condition with the $D(3, 4)$ dataset.