Unsupervised Translation of Emergent Communication

Ido Levy[1], Orr Paradise[2], Boaz Carmeli[1], Ron Meir[1], Shafi Goldwasser[2], Yonatan Belinkov[1]

[1]Technion – Israel Institute of Technology, Haifa, Israel
[2]University of California, Berkeley, CA, USA
{idolevy, boaz.carmeli}@campus.technion.ac.il, rmeir@ee.technion.ac.il, belinkov@technion.ac.il,
{orr.paradise, shafi.goldwasser}@berkeley.edu

Abstract

Emergent Communication (EC) provides a unique window into the language systems that emerge autonomously when agents are trained to jointly achieve shared goals. However, it is difficult to interpret EC and evaluate its relationship with natural languages (NL). This study employs unsupervised neural machine translation (UNMT) techniques to decipher ECs formed during referential games with varying task complexities, influenced by the semantic diversity of the environment. Our findings demonstrate UNMT's potential to translate EC, illustrating that task complexity characterized by semantic diversity enhances EC translatability, while higher task complexity with constrained semantic variability exhibits pragmatic EC, which, although challenging to interpret, remains suitable for translation. This research marks the first attempt, to our knowledge, to translate EC without the aid of parallel data.

1 Introduction

Emergent communication (EC) describes the phenomenon in which AI agents develop communication protocols to achieve shared goals (Giles and Jim 2003; Kasai, Tenmoto, and Kamiya 2008; Boldt and Mortensen 2024; Brandizzi 2023). This capacity has garnered attention due to its significant potential for understanding the complexities of language formation and evolution within multi-agent systems. Yet, ECs remain largely opaque, difficult to interpret and translate into human-readable forms.

Although efforts to understand EC have been made, interpretability remains elusive. Traditional approaches, such as *topographic similarity* (Brighton and Kirby 2006), measure the correlation between message and input distances, provide a coarse measure of EC structures, but do not capture the full nuances of EC compositionality. Recent advancements aim to use natural language (NL) to analyze EC (Xu, Niethammer, and Raffel 2022; Carmeli, Belinkov, and Meir 2024; Chaabouni et al. 2020). Another line of research aims to translate EC into NL using parallel data (EC-NL pairs) to gain a more intuitive understanding (Andreas, Dragan, and Klein 2017; Yao et al. 2022). However, this approach may introduce certain biases in the translation process.

Our approach is to utilize *unsupervised neural machine translation* (UNMT) to translate the emergent languages developed during referential games of varying complexity. UNMT is particularly suited for this task as it does not require parallel data (Lample et al. 2018a; Artetxe, Labaka, and Agirre 2018; Conneau et al. 2017; Lample et al. 2018b; Xu et al. 2018), which is typically unavailable for AI-generated languages. To facilitate translation, we generate an EC corpus from each game by compiling the messages exchanged between agents. Concurrently, we employ a separate English caption dataset as a linguistic prior, which provides a resource of image descriptions for images provided to the agents during the game (Naturally, the images provided to the agents contains no caption). By integrating these distinct datasets, we can adapt existing UNMT techniques (Chronopoulou, Stojanovski, and Fraser 2020) to translate the emergent communications into English, showcasing their translatability across varying task complexities.

Leveraging the capabilities of UNMT, our study dives into the intricacies of structured referential games of varying complexity (Lewis 2008; Lazaridou, Peysakhovich, and Baroni 2016; Choi, Lazaridou, and De Freitas 2018; Guo et al. 2019), including *Random*, *Inter-category*, *Supercategory*, and *Category* image discrimination tasks. In these games, agents must identify a target image, such as a *giraffe*, from a set of distractors. Distractors are non-target images selected based on the game type, adding complexity to the task (Figure 1a). For example, in a Supercategory game, a distractor for a *giraffe target* can be *cow*, whereas in a Random game, the distractors could be as unrelated as a *fire hydrant* (Figure 2), for the same *giraffe* target. These games are deliberately designed to simulate diverse communication environments, aimed at constructing ECs that increasingly resemble human languages, and will eventually be useful for interpretability and usability of AI-generated languages in real-world scenarios. We hypothesize that, akin to the evolution of human languages that adapt in response to social and environmental pressures (Kuhl 2004; Kottur et al. 2017), the complexity inherent in these games will catalyze similar transformations in the EC.

Through our research, we have established a new benchmark for translating EC into NL. This benchmark leverages a comprehensive suite of metrics designed to capture the intricacy and diversity of AI-generated languages. These

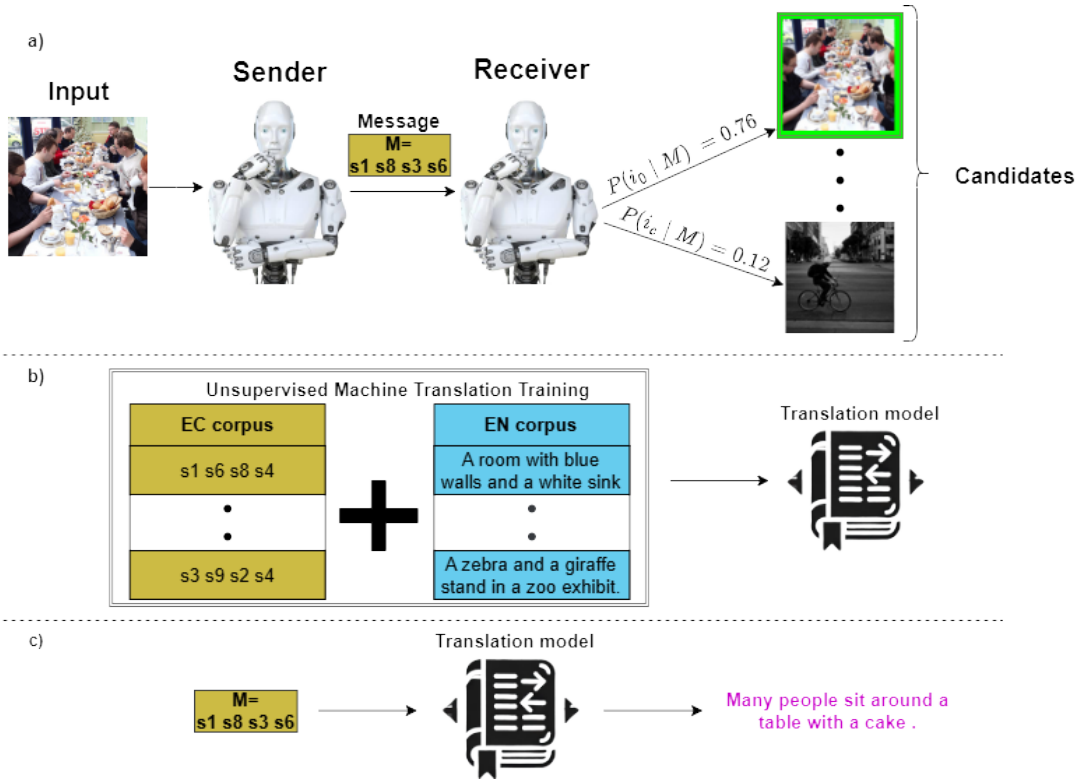

Figure 1: **(a)** Illustration of the referential game setup. The Sender observes an image and sends a message to the Receiver, who must identify the correct image from a set of candidates based on the message received. The exchanged messages are recorded to create the EC corpus. **(b)** Using the monolingual EC corpus and a monolingual English caption corpus to train the UNMT system. **(c)** The UNMT translating an EC message into English.

metrics assess the translation quality in several dimensions: standard machine translation metrics with image captions as ground truth, semantic correlation with images using CLIP scores (Hessel et al. 2021) to assess interpretability, and intrinsic text analysis through metrics like token-type ratio that reflect linguistic richness in the translated text.

Our experiments support the potential of UNMT to translate AI-generated languages into human-readable text. Notably, the *Inter-category* setting showcased superior translation quality, evidenced by higher BLEU and METEOR scores. Furthermore, our analysis indicates that greater vocabulary usage, reflecting a broader range of vocabulary in EC messages, and increased entropy, signaling message unpredictability, may pose challenges to translation accuracy. Qualitative analysis of the resulting translations suggests that UNMT successfully captures the main objects or themes in the described images, but not all the fine-grained details. A notable discovery of our research is the lack of correlation of translation performance across ECs originating from different setups, whether due to varying game complexities or initialization seeds for randomness, when tested on identical test sets. This lack of correlation indicates that each instantiation of a game cultivates distinct communication protocols and not a single universal standard.

Our work makes three key contributions:

- We provide a novel application of UNMT techniques to decipher AI agents' EC without the need for parallel data.
- We investigate how the complexity of referential games influences the ECs development and their translatability.
- We establish a novel benchmark for evaluating EC translations by introducing a comprehensive set of metrics tailored to assess the sophistication and variability of ECs.

2 Related Work

Research in EC has developed various methodologies to understand and evaluate the structure and sophistication of languages that emerge among AI agents. Early efforts to quantify communication among agents focused on traditional metrics like *topographic similarity*, which assesses the correlation between the distances of messages and their corresponding inputs (Brighton and Kirby 2006). To understand EC, Lazaridou et al. (2018) explored the input space effect on EC by demonstrating that AI agents can develop EC that solves the shared task using both symbolic and pixel inputs, while Lee et al. (2018) highlighted how factors like model capacity and communicative bandwidth foster systematic compositional structures in EC. To understand NL properties in EC, several studies have explored aspects such as Zipf's Law of Abbreviation (Ueda and Washio 2021) and Harris's Articulation Scheme (Ueda, Ishii, and Miyao

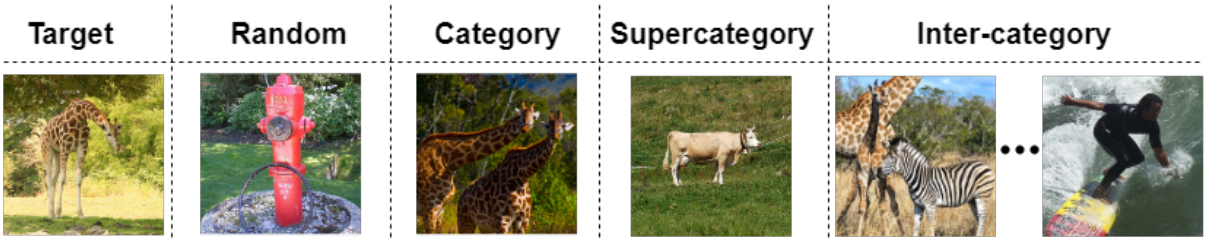

Figure 2: Illustration of various levels of game complexity in referential games. The target image, a giraffe, is shown alongside different levels of distractors: a red fire hydrant representing an random game; another giraffe for category discrimination; cow in a lush field for supercategory discrimination; and a zebra alongside a giraffe and a person wavesurfing for a Inter-category game, where overlapping concepts (giraffes) illustrate the images' inherent complexity in a multi-category setting. Each column exemplifies the escalation of game complexity and the corresponding increase in potential target and distractor candidates.

2022). Additional work focused on compositionality, examining disentanglement-based measures (Chaabouni et al. 2020), Adjusted Mutual Information (AMI) (Mu and Goodman 2021), and rigorous methods to verify compositional structures (Vani et al. 2021; Andreas 2019). More recent advances have introduced more nuanced metrics that aim to understand EC using NL. For instance, Xu, Niethammer, and Raffel (2022) evaluated learned EC on unseen test data to assess generalization, providing insights into NL aspects. Carmeli, Belinkov, and Meir (2024) proposed mapping EC symbols to NL concepts, assessing EC's compositionality. However, their approach forms a global mapping of atomic symbols, rather than full translation of individual messages. Most closely related to our approach is work translating EC into human-understandable language. Andreas, Dragan, and Klein (2017) translated continuous communication channels into NL by collecting agents' messages and corresponding NL strings within the same games. Yao et al. (2022) trained image discrimination referential game and translated the EC to NL captions grounded on the same images. While promising, these studies crucially used parallel EC-NL pairs for training the translation system. In contrast, our approach does not rely on any parallel data. Several works leverage EC as a tool to enhance NL translation in unsupervised settings (Downey et al. 2023; Lee et al. 2018). However, unlike our focus on directly translating EC, these studies utilized EC to facilitate future NL translation tasks.

Guo et al. (2021) termed EC expressivity as the amount of input information encoded in EC, defining theoretically and demonstrating empirically the impact of unpredictability and contextual similarity on EC expressivity. In this paper, we extend the concept of contextual complexity by introducing a broader spectrum of complexity, including low, moderate, and high complexity levels. This allows us to systematically explore how varying degrees of similarity among distractors influence the translatability of EC.

3 Complexity of Referential Games

Complexity refers to the degree of similarity between distractors and the target in referential games. This property can range from being semantically related to completely unrelated. In this section, we define four levels of complexities that represent a wide range of game complexity levels.

Preliminaries

We focus on a two-agent setup, consisting of a *Sender* and a *Receiver*, operating under predefined rules within simulated environments. These agents interact through a discrete communication channel (Lazaridou and Baroni 2020; Denamganaï and Walker 2020) to successfully complete shared tasks by conveying information about objects or concepts represented in the environment.

In the framework of referential games, the *Sender* observes an image $i \in \mathcal{D}$ and sends a message $m \in \mathcal{M}$ to describe the image to the *Receiver*. The *Receiver* uses this message to identify the correct image from a set of candidates, where the target's position is randomized. This setup is defined by the tuple $(A, \mathcal{M}, \mathcal{D}, \mathcal{C}, \mathcal{S})$, where:

- $A = \{A_S, A_R\}$ are the Sender and Receiver agents.
- $\mathcal{M}$ represents the message space.
- $\mathcal{D}$ is the dataset used to sample images.
- $\mathcal{C}$ is the set of categories in the dataset.
- $\mathcal{S}$ encompasses the set of supercategories, grouping similar categories together.

Game Complexity Levels

We craft the environment for EC games around a target image i_T classified under a category c_T and a supercategory s_T. The distractor environment is composed of d images chosen based on the following criteria:

1. **Random:** In this simplest form of the game, agents operate in a completely random environment, where each distractor image i is sampled uniformly from the dataset: $i \sim \mathcal{D}$. For example, the task could involve distinguishing between a car and a zebra. This basic level challenges the agents to develop fundamental communication protocols from scratch, testing their ability to generate and understand rudimentary language. In Lazaridou et al. (2018), it is described as a *uniform* game.
2. **Category Discrimination:** The most contextually complex level involves agents discriminating between images containing objects of the same category, such as distinguishing between different images of giraffes. Here each distractor i is sampled from the set of images that share a category with that of the target images: $i \sim \mathcal{D}_{c_T}$, where

$\mathcal{D}_{c_T} = \{i \in \mathcal{D} \mid c_T \in \text{categories}(i)\}$. This level of complexity is designed to encourage agents to develop highly sophisticated and detailed communication strategies, which means the agents must use nuanced and precise EC to describe subtle differences. In Guo et al. (2021), it is described as a *high-complexity source*.

3. **Supercategory Discrimination:** At this level, agents must distinguish between semantically close images within the same broad category. It is defined analogously to the previous game: $i \sim \mathcal{D}_{s_T}$, where $\mathcal{D}_{s_T} = \{i \in \mathcal{D} \mid s_T \in \text{supercategories}(i)\}$. For instance, they might need to differentiate between a giraffe and a cow in the 'animal' supercategory. This scenario emphasizes the importance of context-aware communication, encouraging agents to address and articulate subtle differences in features and attributes. It is similar to the *context-dependent* game described by Lazaridou et al. (2018).
4. **Inter-category:** In this game setup, d categories are sampled *without replacement*: $\{c_1, \ldots, c_d\} \subseteq \mathcal{C} \setminus c_T$. Then d distractors are sampled, one from each category: $\forall j \in \{1, \ldots, d\}, \quad i_j \sim \mathcal{D}_{c_j}$. Agents discriminate between images that, while representing different categories, may contain overlapping features due to the dataset's inherent complexity. An example task might involve distinguishing between bike and tennis scenes, both featuring a person. This level introduces a layer of ambiguity, requiring agents to refine their communication to highlight distinguishing features in overlapping contexts.

4 Methodology

The training regime of our methodology consists of two phases. First, we train the agents on image discrimination games (Figure 1a) of varying complexity (Section 3) and record the exchanged messages. Each game serves as a unique setup to provoke distinct communication strategies among agents, depending on task complexity. Our next phase is to use EC messages from each game as a monolingual EC corpus for training UNMT system (Figure 1b).

UNMT Architecture For details on UNMT techniques foundational to our study, see Appendix B. Our work employed the UNMT system by Chronopoulou, Stojanovski, and Fraser (2020) that is implemented in three steps:

1. **Pre-training:** Utilizing a high-resource English corpus to train our model, ensuring a rich linguistic foundation.
2. **Fine-tuning:** Adapting the pre-trained model using EC data collected from AI agents, enhancing its ability to handle the specific linguistic features of EC. This phase creates a shared embedding space with the EC.
3. **Back-translation and Denoising:** Employing back-translation and denoising techniques to refine the model's output and improve translation between EC and English.

5 Experimental Setup

Data Specifications

For this study, we employed the MSCOCO dataset (Lin et al. 2014), a diverse collection of 117K complex images that are annotated with various NL concepts. Each image in the dataset is paired with five distinct captions, providing high-quality captions to inject prior knowledge of image descriptions while training our UNMT models, and to serve as valuable reference points for evaluating translation performance. More information about the dataset appears in Appendix A.

AI Agents Architecture

AI agents were configured with a hybrid architecture combining elements of LSTMs (Hochreiter and Schmidhuber 1997) and ResNets (Koonce and Koonce 2021), to process images and to generate ECs effectively, representing a typical architectural standard in the EC domain. Each agent was initialized with different seeds to ensure the robustness and generalizability of the results. The Sender and the Receiver are sharing the ResNet weights, to ensure that both agents process visual information consistently, supporting coherent communication across the system by minimizing discrepancies in visual perception. The ResNet was initialized with pre-trained weights from ImageNet (Deng et al. 2009). The joint training objective is infoNCE (van den Oord, Li, and Vinyals 2018) which is often used in discrimination setups.

Implementation Details

We used the EGG framework (Kharitonov et al. 2019) to train the models with a batch size of 1024 and an initial learning rate of 0.001, using the Adam optimizer. Training epochs were set to 50, with early stopping based on validation loss. In each game, nine distractors are sampled based on the complexity policy. The same target images were used for each complexity's test, but a new set of distractors was sampled according to the complexity of the game. The agents' communication channel is quantized (Carmeli, Meir, and Belinkov 2023), consisting of 64 symbols, represented as binary vectors of length 6, with each message composed of 6 symbols and *EOS* symbol. To ensure robustness, each referential game was run with 5 different random seeds.

The UNMT model utilized a pre-trained XLM, which was fine-tuned on both the EC corpus and MSCOCO captions. More details on the EN corpus are provided in Appendix A. See Appendix G for full hyperparameters.

Evaluation Metrics

To comprehensively assess the effectiveness of EC games and the UNMT, we employed a diverse set of evaluation metrics. See Appendix E for more details on the metrics.

EC Games Metrics We used the following metrics to evaluate the performance and sophistication of the EC emerged by AI agents during the referential games:

- **Game Accuracy**: Measures the percentage of correct identifications made by the agents (Lewis 2008).
- **Vocabulary Usage (VU)**: Assesses vocabulary diversity by measuring the range of unique symbols utilized during game sessions (Graesser et al. 2003).
- **Message Entropy**: Calculates the uncertainty in the agents' messages (Shannon 1948).
- **Message Novelty**: Determines whether test messages differ from training messages (Chomsky 2002).

Metric	Category	Supercategory	Random	Inter-category	Baseline
ACC 2 (%)	93.62 ± 0.20	96.35 ± 0.56	96.89 ± 0.50	96.90 ± 0.46	54.25 ± 0.81
ACC 10 (%)	70.51 ± 0.88	77.35 ± 2.96	80.20 ± 2.25	78.40 ± 2.62	13.55 ± 1.23
VU (%)	55.95 ± 2.04	73.10 ± 9.77	68.26 ± 9.03	66.34 ± 7.73	100.0 ± 0.0
Entropy	6.78 ± 0.41	9.27 ± 1.38	7.79 ± 0.80	7.71 ± 2.00	14.34 ± 1.02
Novelty (%)	0.84 ± 0.42	6.85 ± 2.28	1.35 ± 0.66	3.20 ± 1.46	0.0 ± 0.0

Table 1: EC results, segmented by complexity and including a baseline of random agent communication. Metrics reported as mean ± standard error from 5 seeds. ACC 2 and ACC 10 represent discrimination accuracy with 1 and 9 distractors, respectively.

Metric	Category	Supercategory	Random	Inter-category	Baseline	Chinese
Novelty (%)	58.74 ± 7.81	**70.00** ± 1.68	60.54 ± 4.25	57.36 ± 5.83	100.0	88.19
BLEU Score	7.41 ± 0.47	6.08 ± 0.31	6.85 ± 0.34	**9.21** ± 0.45	0.071	5.65
BERTScore	**0.734** ± 0.001	0.730 ± 0.001	0.729 ± 0.001	0.730 ± 0.001	0.543	0.74
METEOR Score	0.295 ± 0.06	0.276 ± 0.06	0.289 ± 0.06	**0.310** ± 0.07	0.115	0.234
ROUGE-L	0.361 ± 0.001	0.343 ± 0.006	0.352 ± 0.003	**0.370** ± 0.002	0.173	0.356
Jaro Similarity	0.678 ± 0.02	0.673 ± 0.02	0.676 ± 0.02	**0.682** ± 0.02	0.601	0.674
CLIP Score	0.180 ± 0.018	0.176 ± 0.019	0.183 ± 0.020	**0.191** ± 0.019	0.151	-
TTR (%)	0.42 ± 0.05	**0.71** ± 0.14	0.58 ± 0.11	0.59 ± 0.15	0.19	0.83

Table 2: UNMT Performance Across Different Game Complexities. Each metric is reported as mean ± standard error, derived from 3 different ECs that emerged from the same complexity.

UNMT Metrics In the absence of parallel data, we used the captions of the images to evaluate the translation quality of UNMT. Since each caption captures different nuances and writing styles, and we aim to translate EC messages generated by the agents to a specific image, a good translation would be similar to at least one of the captions. Therefore, we reported the max score over 5 captions to account for this variability. Metrics were calculated using the `string2string` package (Suzgun, Shieber, and Jurafsky 2023). We employed a variety of metrics that capture both exact match accuracy and semantic alignment:

- **BLEU**: Measures the n-gram overlap between translated text and reference captions (Papineni et al. 2002). We employed the default SacreBLEU (Post 2018), which provides a standardized way to evaluate translations quality by ensuring consistent and reproducible scores.
- **METEOR**: Considers precision, recall, synonymy, stemming, and word order (Banerjee and Lavie 2005).
- **ROUGE-L**: Focuses the longest common subsequence between the translated text and reference (Lin 2004).
- **BERTScore**: Utilizes contextual embeddings from BERT to evaluate semantic similarity between the translated EC and reference captions (Zhang et al. 2019).
- **CLIP Score**: Employs the CLIP model to assess the semantic alignment between the translated text and the corresponding image. This metric introduces a novel approach for evaluating how well the translation reflects the content and context of the image that prompted the original EC message (Hessel et al. 2021).
- **Jaro Similarity**: Measures similarity and character transpositions between texts (Jaro 1989).
- **Text-Type Ratio (TTR)**: Analyzes the lexical diversity within the message translations (Richards 1987).
- **Novelty Score**: Evaluates how many unique n-grams are produced in translations compared to the training corpus.

6 Results

Emergent Communication Games The baseline performs marginally above random, which is suggestive of Buzaglo et al. (2024)'s findings on limited generalization in randomly sampled networks. As shown in Table 1, it significantly underperforms across all metrics, with a VU of 100%, indicating an arbitrary use of the EC vocabulary without solving the task. These results demonstrate the effectiveness of the EC strategies developed in the games.

As expected, in the ten-candidates (ACC 10) scenario, the game accuracy scores degrade with increasing task difficulty, with *Category* game complexity presenting significantly lower results with 70.5%. Despite differences in the communication channel, as opposed to Guo et al. , who claim that more contextual similarity leads to higher expressivity, we observe that the Category game complexity, which emphasizes fine-grained distinctions, has the lowest VU (55.95%), a metric indicating EC richness. We attribute this finding to the complexity level, as more specificity is required, agents might use what we hypothesize as "low-level features", like angle, background, and object size, which require fewer symbols for the task, leading to the development of simplistic communication strategies.

The low standard errors across the accuracy metrics indicate that EC assessment metrics such as *VU* and entropy do not correlate well with classic referential game accuracy metrics. This indicates that the emerged protocols successfully solve tasks using diverse strategies. The overall low novelty scores across games, combined with low standard errors, imply that agents largely rely on memorized messages generated during training. The table shows some am-

biguity in the results due to interval overlap in some metrics across different complexities. However, there is a high correlation between Entropy and *VU* ($\rho = 0.85$) across all complexities.[1] Among the *Supercategory*, *Random*, and *Inter-category* complexities, we observe that *Supercategory* possesses the highest scores in both Entropy and *VU*, indicating a need for a more extensive vocabulary to discriminate between semantically related concepts. Despite the overall low novelty scores, *Supercategory*'s is significantly higher at 6.85%, which is directly correlated with Entropy ($\rho = 0.8$) and *VU* ($\rho = 0.7$). These correlations reveal that rich and diverse EC impacts the novelty of the agents' messages.

Unsupervised Machine Translation

Benchmark UNMT system's performance We conducted two experiments to understand the UNMT capabilities. The first experiment aimed to calibrate the performance of our UNMT approach on a real, high-resource language by translating Chinese captions to English. For this experiment, we used MSCOCO-CN (Li et al. 2019) composed of $\sim$ 30K annotated images. The results, detailed in Appendix F , showed fluent English translations with an overall BLEU score of 5.65 points. The second experiment, serves as a baseline, evaluates the performance of the UNMT on EC generated by random agents that did not undergo the training phase designed to optimize EC. The results, presented in Table 2, indicate significantly lower performance metrics, with BLEU scores near zero. This confirmed that random agent communication did not form meaningful linguistic patterns, strengthening our translation reliability on EC that emerged from AI agents' collaboration towards a shared goal.

UNMT of EC Our findings demonstrate the potential of UNMT for translating EC into English without the use of parallel data. This is reflected in all translation metrics: While modest compared to full-fledged supervised translation of natural languages, these metrics are significantly higher than the baseline, indicating the formation of meaningful EC during collaborative training. For illustrative translation examples, refer to Figure 3 and Appendix I. Interestingly, these scores are similar or slightly better than the Chinese-to-English translation results, which could be attributed to the richness differences between EC and Chinese or to the relatively small number of Chinese captions compared to EC messages. The translations show a considerable level of coherence and meaning. Notably, the *Inter-category* complexity achieved the highest BLEU score (9.21) and ROUGE-L score (0.37). We speculate this is due to the fact that in this setup the image discrimination game involves different categories, shaping the information conveyed by the EC to be specific to a category among others that may share mutual features, making it more concept-related, and therefore aligning the EC more towards NL, facilitating translations compared to other complexity levels. This alignment is supported by empirical evidence in Appendix D, demonstrating successful mapping of EC messages to NL concepts. Evaluating translation-image alignment, the CLIP score reaches 0.19, compared to 0.29 for MSCOCO's actual captions and 0.15 for the random baseline, indicating acceptable performance given CLIP's tendency toward lower scores with diverse, complex images. Thus, while our translations are better than translating random messages, they fall short of aligning real captions.

[1]Reported correlations are significant at $p < 0.05$.

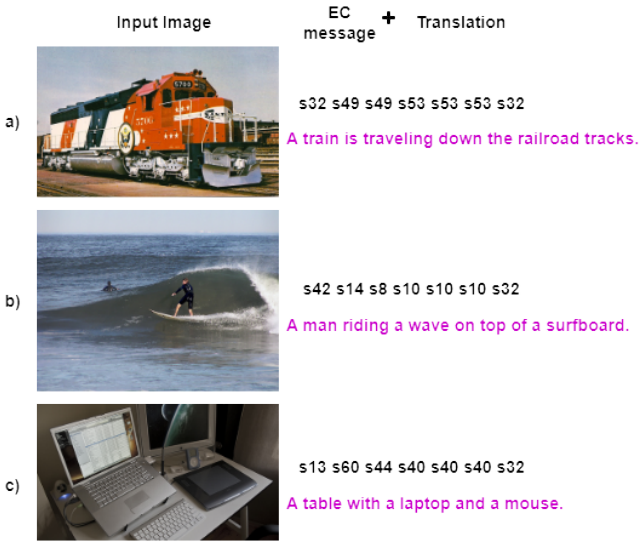

Figure 3: Selected translation examples. Each panel shows an EC message, composed of six symbols followed by an *EOS* symbol, paired with its corresponding translation beneath. The translations capture nuanced details from the visuals, producing coherent and contextually appropriate text. Notably, in panel c), the MT model extends its response beyond the visible elements in the image, by adding "*and a mouse*", suggesting a tendency to "hallucinate" details potentially influenced by prior context knowledge and previous training on similar caption lengths.

While the translations are coherent, we observed an interesting phenomenon where translations sometimes capture several features from the image but end with unrelated ones. We attribute this to the model "hallucinating" based on the narrow distribution of caption lengths it encountered during training; for an example refer to Figure 3c. This suggests the model captures multiple concepts from the messages, but occasionally extends the translation with irrelevant details. This observation indicates that while the translations demonstrate potential and capture the main theme, they are not perfect and warrant further investigation in future research.

Game complexity vs. Translatability Several key insights emerge from our analysis. Translation performance metrics, including BLEU and ROUGE-L, are highest for *Inter-category*. This indicates that partial supervision and candidates from broad categories facilitate better translation, making emerging languages in these contexts more translatable. In contrast, *Supercategory*, with high contextual complexity, presents the highest novelty score (70.00%) while having the lowest BLEU (6.08) and ROUGE-L (0.352). This highlights the difficulty of translating ECs that were developed based on semantically similar image discrimina-

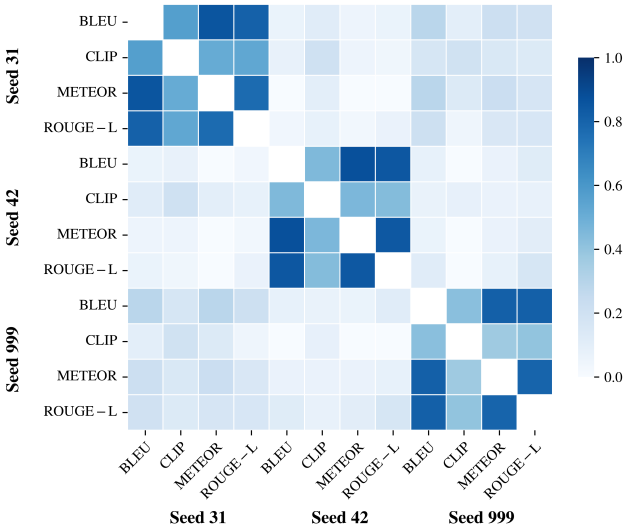

Figure 4: Correlation matrix for Inter-Category complexity across different seeds. The near-zero correlation between seeds indicates that translation model performs uniquely on the same examples. Additionally, the strong correlation between exact match metrics, which positively correlate with the semantic metric within each seed, highlights the significant text-image alignment achieved by our translations.

tion. As illustrated by Table 1, this complexity type is associated with the highest Entropy and VU, indicating that agents communicated diverse messages, which directly affected translation performance and encouraged more novel translations. Conversely, the *Category* complexity, designed for fine-grained distinctions, achieved moderate translation metrics with BLEU (7.41) and ROUGE-L (0.361), significantly surpassing *Supercategory* and *Random*, despite the lower scores among the EC metrics. We attribute this result to the high specificity leading to more simplistic communication strategies, characterized by a more limited and predictable set of symbols, making it easier to translate.

In addition, the exact match metrics (BLEU, METEOR, ROUGE-L) are strongly correlated with each other ($\rho = 0.72$), while they are slightly correlated with the CLIP score ($\rho = 0.26$), which provides semantic similarity between the candidate translation and the image itself. We attribute this to the inductive bias inherent in the experiment, as the ground truth are the images' captions written by annotators, and each image can be described in many ways. In spite of this, we observe a stronger correlation with *Random* ($\rho = 0.37$) than with any of the other complexities.

Finally, Figure 4's correlation matrix offers insights into how translation metrics interrelate across different setups and seeds. The matrix reveals that results on the same test set and game type are not correlated among all ECs, suggesting the translation system captures different nuances from each EC. This lack of correlation indicates that each EC produces unique linguistic structures, highlighting the diversity in the communication protocols developed by the agents.

7 Conclusion

This research introduces a novel application of UNMT to translate AI agents' EC emerged in referential games into natural language. Our study demonstrates UNMT's potential to translate EC without parallel data and examines the impact of game complexity on EC translatability. Our findings indicate that different complexities produce distinct communication patterns. For instance, the *Inter-Category* game generated EC that achieved significantly higher translation accuracy than the *Supercategory* game, which often had lower scores despite its complexity. This suggests that not all complexities foster equally translatable communications.

Contrary to the initial hypothesis that more complex setups result in more detailed communications that are easier to translate, our empirical evidence shows otherwise. EC emerging from the *Category* game, characterized by high contextual complexity, resulted in more pragmatic communication with significantly lower VU and Entropy while exhibiting high game accuracy. In contrast, the *Supercategory* game, with slightly less contextual complexity, produced richer communication. We attribute this phenomenon to the agents' optimization for efficiency, where fine-grained distinctions encourage the use of low-level features. However, in terms of translatability, the *Category* game achieved higher translation scores than the *Supercategory* game, suggesting that a more pragmatic protocol facilitates translation.

Limitations and Future Work

This study provides valuable insights into translating EC using UNMT. However, several limitations should be acknowledged. First, the number of distractors adds a layer of complexity that was not fully explored. Second, sharing a pre-trained visual module may limit generality. Future work could explore separate or scratch-trained modules to assess their impact on EC translatability. Third, the communication channel, vocabulary size, and messages length were fixed based on preliminary experiments. While these configurations were chosen to optimize performance, they may not capture the full range of possible communication strategies. Before translating, we analyzed many such configurations, ultimately selecting the setups with the best results.

Beyond limitations, the study suggests several future research avenues. EC is a rich field, characterized by a variety of techniques and strategies that form unique ECs. These range from multi-agent collaboration (Michel et al. 2023), to bidirectional communication (Nikolaus 2023), and reconstruction objectives (Chaabouni et al. 2021). Future research should enhance UNMT techniques to more effectively capture the subtleties of these ECs (Chauhan et al. 2022; Amani et al. 2024) and assess faithfulness by having the Receiver evaluate how well the back-translation (EC→EN→EC) captures the original message's nuanced details. Finally, success in unsupervised translation of EC can further motivate similar attempts to translate other communication systems, such as animal communication (Goldwasser et al. 2024).

Acknowledgements

This research was supported by grant no. 2022330 from the United States - Israel Binational Science Foundation (BSF), Jerusalem, Israel. IL, BC, and YB were supported by the Israel Science Foundation (grant no. 448/20), an Azrieli Foundation Early Career Faculty Fellowship, and an AI Alignment grant from Open Philanthropy. OP was funded by Project CETI via grants from Dalio Philanthropies and Ocean X; Sea Grape Foundation; Virgin Unite and Rosamund Zander/Hansjorg Wyss through The Audacious Project: a collaborative funding initiative housed at TED. RM was supported by the Skillman chair.

References

Amani, M. H.; Baldwin, N. M.; Mansouri, A.; Josifoski, M.; Peyrard, M.; and West, R. 2024. Symbolic Autoencoding for Self-Supervised Sequence Learning. *arXiv preprint arXiv:2402.10575*.

Andreas, J. 2019. Measuring compositionality in representation learning. *arXiv preprint arXiv:1902.07181*.

Andreas, J.; Dragan, A.; and Klein, D. 2017. Translating Neuralese. In *Proceedings of the ACL*.

Artetxe, M.; Labaka, G.; and Agirre, E. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv preprint arXiv:1805.06297*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Boldt, B.; and Mortensen, D. 2024. A review of the applications of deep learning-based emergent communication. *arXiv preprint arXiv:2407.03302*.

Brandizzi, N. 2023. Toward More Human-Like AI Communication: A Review of Emergent Communication Research. *IEEE Access*, 11: 142317–142340.

Brighton, H.; and Kirby, S. 2006. Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial life*, 12(2): 229–242.

Buzaglo, G.; Harel, I.; Nacson, M. S.; Brutzkus, A.; Srebro, N.; and Soudry, D. 2024. How Uniform Random Weights Induce Non-uniform Bias: Typical Interpolating Neural Networks Generalize with Narrow Teachers. *arXiv preprint arXiv:2402.06323*.

Carmeli, B.; Belinkov, Y.; and Meir, R. 2024. Concept-Best-Matching: Evaluating Compositionality In Emergent Communication. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Carmeli, B.; Meir, R.; and Belinkov, Y. 2023. Emergent quantized communication. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10, 11533–11541.

Chaabouni, R.; Kharitonov, E.; Bouchacourt, D.; Dupoux, E.; and Baroni, M. 2020. Compositionality and Generalization In Emergent Languages. In *Proceedings of the ACL*.

Chaabouni, R.; Strub, F.; Altché, F.; Tarassov, E.; Tallec, C.; Davoodi, E.; Mathewson, K. W.; Tieleman, O.; Lazaridou, A.; and Piot, B. 2021. Emergent communication at scale. In *International conference on learning representations*.

Chauhan, S.; Daniel, P.; Saxena, S.; and Sharma, A. 2022. Fully unsupervised machine translation using context-aware word translation and denoising autoencoder. *Applied Artificial Intelligence*, 36(1): 2031817.

Choi, E.; Lazaridou, A.; and De Freitas, N. 2018. Compositional obverter communication learning from raw visual input. *arXiv preprint arXiv:1804.02341*.

Chomsky, N. 2002. *Syntactic structures*. Mouton de Gruyter.

Chronopoulou, A.; Stojanovski, D.; and Fraser, A. 2020. Reusing a Pretrained Language Model on Languages with Limited Corpora for Unsupervised NMT. In *Proceedings of EMNLP*.

Conneau, A.; Lample, G.; Ranzato, M.; Denoyer, L.; and Jégou, H. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Denamganaï, K.; and Walker, J. A. 2020. Referentialgym: A nomenclature and framework for language emergence & grounding in (visual) referential games. *arXiv preprint arXiv:2012.09486*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Downey, C.; Zhou, X.; Liu, Z.; and Steinert-Threlkeld, S. 2023. Learning to translate by learning to communicate. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*.

Giles, C. L.; and Jim, K.-C. 2003. Learning communication for multi-agent systems. In *Innovative Concepts for Agent-Based Systems: First International Workshop on Radical Agent Concepts*.

Goldwasser, S.; Gruber, D.; Kalai, A. T.; and Paradise, O. 2024. A Theory of Unsupervised Translation Motivated by Understanding Animal Communication. *Advances in Neural Information Processing Systems*, 36.

Graesser, A. C.; McNamara, D. S.; Louwerse, M. M.; and Cai, Z. 2003. Automated Evaluation of Discourse Coherence Quality in Essays. In *Proceedings of the Workshop on Educational Data Mining*, 7–14.

Guo, S.; Ren, Y.; Havrylov, S.; Frank, S.; Titov, I.; and Smith, K. 2019. The emergence of compositional languages for numeric concepts through iterated learning in neural agents. *arXiv preprint arXiv:1910.05291*.

Guo, S.; Ren, Y.; Mathewson, K.; Kirby, S.; Albrecht, S. V.; and Smith, K. 2021. Expressivity of emergent language is a trade-off between contextual complexity and unpredictability. *arXiv preprint arXiv:2106.03982*.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings EMNLP*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Jaro, M. A. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406).

Kasai, T.; Tenmoto, H.; and Kamiya, A. 2008. Learning of communication codes in multi-agent reinforcement learning problem. In *2008 IEEE Conference on Soft Computing in Industrial Applications*, 1–6.

Kharitonov, E.; Chaabouni, R.; Bouchacourt, D.; and Baroni, M. 2019. EGG: a toolkit for research on Emergence of lanGuage in Games. In Padó, S.; and Huang, R., eds., *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*, 55–60. Hong Kong, China: Association for Computational Linguistics.

Koonce, B.; and Koonce, B. 2021. ResNet 50. *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, 63–72.

Kottur, S.; Moura, J. M.; Lee, S.; and Batra, D. 2017. Natural language does not emerge 'naturally' in multi-agent dialog. *arXiv preprint arXiv:1706.08502*.

Kuhl, P. K. 2004. Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, 5(11): 831–843.

Lample, G.; Ott, M.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2018a. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Lample, G.; Ott, M.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2018b. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Lazaridou, A.; and Baroni, M. 2020. Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*.

Lazaridou, A.; Hermann, K. M.; Tuyls, K.; and Clark, S. 2018. Emergence of Linguistic Communication from Referential Games with Symbolic and Pixel Input. *ArXiv*, abs/1804.03984.

Lazaridou, A.; Peysakhovich, A.; and Baroni, M. 2016. Multi-agent cooperation and the emergence of (natural) language. *arXiv preprint arXiv:1612.07182*.

Lee, J.; Cho, K.; Weston, J.; and Kiela, D. 2018. Emergent Translation in Multi-Agent Communication. In *International Conference on Learning Representations*.

Lewis, D. 2008. Convention: A Philosophical Study. In *Convention: A Philosophical Study*. Harvard University Press.

Li, X.; Xu, C.; Wang, X.; Lan, W.; Jia, Z.; Yang, G.; and Xu, J. 2019. COCO-CN for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9): 2347–2360.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*.

Lin, T.-Y.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*.

Michel, P.; Rita, M.; Mathewson, K. W.; Tieleman, O.; and Lazaridou, A. 2023. Revisiting Populations in multi-agent Communication. In *Proceedings of ICLR*.

Mu, J.; and Goodman, N. 2021. Emergent communication of generalizations. *Proceedings of NeurIPS*.

Nikolaus, M. 2023. Emergent Communication with Conversational Repair. In *The Twelfth International Conference on Learning Representations*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the ACL*.

Post, M. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Richards, B. 1987. Type/Token Ratios: What do they really tell us? *Journal of Child Language*, 14(2): 411–421.

Shannon, C. E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3): 379–423.

Suzgun, M.; Shieber, S. M.; and Jurafsky, D. 2023. string2string: A Modern Python Library for String-to-String Algorithms. *ArXiv*, abs/2304.14395.

Ueda, R.; Ishii, T.; and Miyao, Y. 2022. On the Word Boundaries of Emergent Languages Based on Harris's Articulation Scheme. In *Proceedings of ICLR*.

Ueda, R.; and Washio, K. 2021. On the relationship between Zipf's law of abbreviation and interfering noise in emergent languages. In *Proceedings of the ACL and IJCNLP: Student Research Workshop*.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. *ArXiv*, abs/1807.03748.

Vani, A.; Schwarzer, M.; Lu, Y.; Dhekane, E.; and Courville, A. 2021. Iterated learning for emergent systematicity in VQA. In *Proceedings of ICLR*.

Xu, R.; Yang, Y.; Otani, N.; and Wu, Y. 2018. Unsupervised cross-lingual transfer of word embedding spaces. *arXiv preprint arXiv:1809.03633*.

Xu, Z.; Niethammer, M.; and Raffel, C. A. 2022. Compositional Generalization in Unsupervised Compositional Representation Learning: A Study on Disentanglement and Emergent Language. In *Proceedings of NeurIPS*.

Yao, S.; Yu, M.; Zhang, Y.; Narasimhan, K.; Tenenbaum, J.; and Gan, C. 2022. Linking Emergent and Natural Languages via Corpus Transfer. In *Proceedings of ICLR*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of ICLR*.