Concept-Best-Matching: Evaluating Compositionality in Emergent Communication

Boaz Carmeli **Yonatan Belinkov** **Ron Meir**
Technion – Israel Institute of Technology
boaz.carmeli@campus.technion.ac.il,
belinkov@technion.ac.il
rmeir@ee.technion.ac.il

Abstract

Artificial agents that learn to communicate in order to accomplish a given task acquire communication protocols that are typically opaque to a human. A large body of work has attempted to evaluate the emergent communication via various evaluation measures, with *compositionality* featuring as a prominent desired trait. However, current evaluation procedures do not directly expose the compositionality of the emergent communication. We propose a procedure to assess the compositionality of emergent communication by finding the best-match between emerged words and natural language concepts. The best-match algorithm provides both a global score and a translation-map from emergent words to natural language concepts. To the best of our knowledge, it is the first time that such direct and interpretable mapping between emergent words and human concepts is provided.

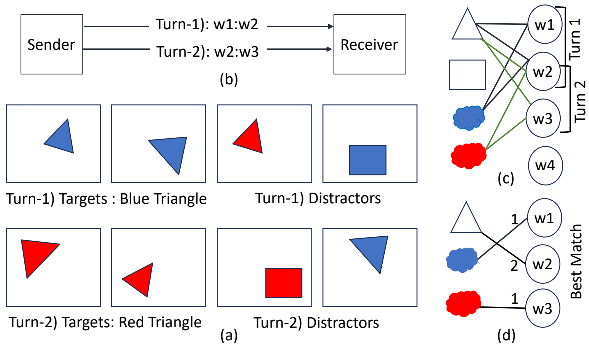

Figure 1: The multi-target shape game played in turns (showing two turns). (a) At each turn, the sender is given a set of images, a subset of them marked as targets by an Oracle. (b) Sender generates messages $[w_1, w_2]$ for the `blue-triangle` (turn 1) and $[w_2, w_3]$ for the `red-triangle` (turn 2). (c) During evaluation, we construct a bipartite graph of words generated by the sender and concepts provided by the Oracle for each turn. (d) The best-match algorithm matches EC words to NL concepts and provides the CBM score. In this example, all EC words are matched with NL concepts, resulting in a CBM score of 1.0.

1 Introduction

Artificial agents that learn to communicate for accomplishing a given task acquire communication protocols. In the common setting, a sender observes a set of objects and sends a message to a receiver, which then needs to identify the correct target objects out of a set of distractors. The two agents learn a communication protocol over discrete vocabulary of atoms, termed “words”. The emergent communication (EC) is typically opaque to a human. As a result, a large body of work has attempted to characterize the emergent communication in light of natural language (NL) traits, such as compositionality (Hupkes et al., 2020; Chaabouni et al., 2020), systematic generalization (Vani et al., 2021), pragmatism (Andreas and Klein, 2016; Zaslavsky et al., 2021), and more.

Chief among these is compositionality, which enables the construction of complex meanings from the meaning of parts (Li and Bowling, 2019). However, quantifying compositionality in EC is notoriously difficult (Lazaridou et al., 2018). In fact, recent studies found that common compositionality measures in EC, e.g., topographic similarity (Brighton and Kirby, 2006; Li and Bowling, 2019), do not correlate well with success in the task (Chaabouni et al., 2021; Yao et al., 2022). Moreover, while seeking to assess an opaque protocol, common evaluation measures in EC are opaque themselves—they do not provide a human-interpretable characterization of the compositionality of the communication.

This work develops a method that characterizes how compositional an EC is. Our method is founded on the key insight that, in an EC setting, a communication is compositional if agents communicate successfully via complex messages formed of simple atoms *and* these atoms are mapped to NL concepts. Our evaluation is based on the classical

Findings of the Association for Computational Linguistics: ACL 2024, pages 3186–3194
August 11-16, 2024 ©2024 Association for Computational Linguistics

best-match algorithm (Hopcroft and Karp, 1973). Given a set of EC atoms and a set of NL concepts, we construct a bi-partite graph and seek the optimal one-to-one mapping between words and concepts. A perfectly compositional EC would yield a perfect match, where every EC atom is mapped to exactly one NL concept. See Figure 1 for an illustration.

We experiment with the proposed procedure in two EC settings and compare it to two popular evaluation measures – topographic similarity and adjusted mutual information (AMI). We demonstrate that our approach provides a fine-grained characterization of emergent protocols, exposing their strengths and weaknesses, while other measures only provide coarse, opaque scores. According to our evaluation, state-of-the-art communication methods do not achieve satisfactory results.

2 Background

2.1 Emergent Communication Setup

We follow a recent setup suggested by Mu and Goodman (2021), where a sender needs to communicate with a receiver about a *set of target objects* out of a larger set of candidate objects. The sender sends a message to the receiver, which uses it to distinguish targets from distractor objects. For example, to identify all red triangles out of objects with different shapes and colors. This setup is more conducive to emergence of compositional communication than the classical referential game (Lazaridou et al., 2017). Senders in this setup need to form a generalization rather than merely transmit the identity of a single target.

Formally, assume a world $\mathbb{X}$, where each object is characterised by n feature–value pairs (FVPs), $\langle f_1 : v_1, \ldots, f_n : v_n \rangle$, with feature i having k_i possible values, $v_i \in \{f_i[1], \ldots, f_i[k_i]\}$. For instance, the Shape world (Kuhnle and Copestake, 2017) has objects like $\langle$shape:triangle,color:red$\rangle$. Labeling rules are boolean expressions over these FVPs.[1] Each rule $l : \mathbb{X} \mapsto \{0, 1\}$ labels each object as 0 or 1. For instance, the rule `Red Triangle` labels all red triangles as positive and all other objects as negative. We identify the rule `Red Triangle` with the NL *phrase* "red triangle", which comprises the *concepts* "red" and "triangle".

At each turn of the game, we draw a set of candidate objects $\tilde{\mathbb{X}} \subset \mathbb{X}$ that is made of target objects, $\mathbb{T}$, which obey the rule, and distractors, $\mathbb{D}$, which do not: $l(x) = 1$ if $x \in \mathbb{T}$ and $l(x) = 0$ if $x \in \mathbb{D}$.

[1] In this work we only use conjunctive expressions.

The sender encodes the set of target objects $\mathbb{T}$ into a dense representation $\boldsymbol{u}^s \in \mathbb{R}^d$, and generates a message m, a sequence of *words* from some vocabulary. The receiver encodes each candidate object $x \in \mathbb{X}$ into a representation $\boldsymbol{u}^r_x \in \mathbb{R}^d$. It decodes the message m into a representation $\boldsymbol{z} \in \mathbb{R}^d$ and computes a matching score between each encoded candidate and the message representation, $g(\boldsymbol{z}, \boldsymbol{u}^r_x)$. A candidate is predicted as a target if its score is > 0.5. The entire system—sender and receiver networks—is trained jointly with a binary cross entropy on correctly identifying each target.

2.2 Compositionality in EC

Various measures have been proposed to evaluate if an EC protocol is compositional. We mention a few prominent ones below. However, while trying to capture the idea that a complex meaning uses the meaning of its parts, previous measures fail to provide a concrete *mapping of parts*—EC words and NL concepts. We say an emergent communication is compositional if the following conditions hold:

1. EC words are mapped to NL concepts.
2. A complex EC message is composed of simple EC words.
3. Agents communicate successfully via complex messages.
4. An EC message composed of words has the same meaning as an NL phrase composed of the respective concepts.

Being able to compose EC words while preserving their NL meaning allows agents to generalize to new data while using interpretable communication.

2.3 Compositionality Evaluations in EC

We briefly describe two notable measures of compositionality in EC. Appendix A gives more details.

Topographic Similarity (TopSim) (Brighton and Kirby, 2006) measures how well messages align with object representations. Given a set of objects and their EC messages, calculate a matrix of distances between every two messages and a separate matrix of distances between every two object representations. TopSim is the Spearman correlation of the two distance matrices.

TopSim is a global metric that does not require a reference language and can be applied to any EC setup. However, recent studies found that TopSim does not correlate well with success in the task being played (Chaabouni et al., 2021; Yao et al., 2022). More importantly, it does not provide a mapping between EC words and NL concepts, and thus cannot directly assess compositionality.

Adjusted Mutual Information (AMI) (Vinh et al., 2009) measures the MIT between a set of EC messages and a corresponding set of NL phrases, adjusted for chance. Mu and Goodman (2021) showed AMI is a better compositionality measure than TopSim, as it directly assesses the MI between a message and its NL phrases. Still, AMI operates at the level of messages and phrases, rather than atomic words and concepts, respectively.

3 Concept Best Matching

Our key insight is that compositionality requires a mapping between words (atomic parts of EC messages) and concepts (a set of FVPs) that compose NL phrases. Given an evaluation set of examples D, consider their corresponding EC messages $\mathbb{M}$ and NL phrases $\mathbb{L}$. Let $\mathbb{W}$ denote the set of unique words in $\mathbb{M}$, and $\mathbb{V}$ the set of unique FVPs in $\mathbb{L}$. Let $\{w\}_i$, $\{v\}_i$ be the sets of words and FVPs in sample i, and $m_i = |\{w\}_i|$, $l_i = |\{v\}_i|$ their sizes. We construct a weighted bipartite graph $\mathcal{G} = ((\mathbb{W}, \mathbb{V}), \mathbb{E})$ with words on one side and FVPs on the other side. Edges $\mathbb{E}$ are defined by the evaluation set. The weight q_e of edge e_{ij} is the number of times word i appeared in a message that the sender transmitted for a labeling rule with FVP j. This construction reflects the intuition that we do not know the correct mapping between words and FVPs (concepts). See Figure 1 for an example.

3.1 Best Match Algorithm

Given the graph $\mathcal{G}$, we seek an optimal pairing between words and concepts (FVPs), such that no two edges share the same word nor the same concept. The score of a match $\mathbb{B} \subseteq \mathbb{E}$ is the sum of its edge weights, $\sum_{e \in \mathbb{B}} q_e$. The best match, $\mathrm{BM} = \mathrm{BM}(\mathcal{G})$, maximizes this score:

$$\mathrm{BM} = \operatorname*{argmax}_{\mathbb{B}} \sum_{e \in \mathbb{B}} q_e \tag{1}$$

Such an ideal mapping is fully interpretable. A high score indicates that the agents learned to generate a unique EC word for each NL concept.

Normalizing the BM score by the number of symbols and words, $Q = \sum_{i \in D} \max(|m_i|, |l_i|)$, guarantees that $\mathrm{BM} \in [0, 1]$.[2] The best match for a weighted bipartite graph can be found efficiently by the Hungarian algorithm (Kuhn, 1955; Hopcroft and Karp, 1973). Implementation is provided here.[3]

[2] In practice, the lowest bound is $q_{\tilde{e}}/Q$, where $\tilde{e}$ is the edge with the highest weight in $\mathcal{G}$.

4 Experimental Setup

We experiments with agents that learn to play multi-target referential games, introduced by Mu and Goodman (2021) and described in Section 2.1. All experiments reported in Table 1 have the number of possible words $\geq$ the number of concepts, allowing perfect match of words to concepts. In Appendix D we report more experiments with sub-optimal configurations.

Datasets. We experiment with two datasets, described briefly here; Appendix B has more details. **(1)** SHAPE: A visual reasoning dataset (Kuhnle and Copestake, 2017) of objects over a black background. Our version of the dataset contains four attributes: shape, color, and horizontal and vertical positions. Overall, the SHAPE dataset contains 17 concepts, and a maximal labeling rule of 4 FVPs. **(2)** THING: A synthetic dataset of 100,000 objects. Each object has five attributes, each with 10 possible values. Overall, the THING datasets contains 50 concepts and a maximal labeling rule of 5 FVPs.

Communication Channel. We train the system with two types of communication channels: the popular Gumbel-Softmax (GS) (Havrylov and Titov, 2017; Jang et al., 2017) and quantized communication (QT) (Carmeli et al., 2023). In GS, each word is a d-dimensional one-hot vector. In QT, each word is a d-dimensional binary vector. QT was shown to be superior and easier to optimzie compared to GS in EC games. In both cases we use a recurrent network to generate multi-word messages. Appendix C has implementation details.

5 Results

Table 1 shows results on several configurations of the two games. As expected, accuracy degrades as the task becomes more difficult, with longer labeling rules. QT communication yields better task accuracy than GS, consistent with Carmeli et al. (2023). TopSim is not well correlated with accuracy, as found in prior work (Section 2.3). AMI and CBM are more correlated with accuracy. With long messages and labeling rules ($l = 4$), AMI is less correlated with accuracy. This may be explained by AMI operating at the level of whole messages and labeling rules—AMI struggles with large numbers of unique messages and NL phrases. In contrast,

[3] https://github.com/bcarmeli/cbm

	Com	l	d	NL Cons	NL Phrs	EC $\#w$	EC $\#m$	Acc	Top Sim	AMI	Best Match CBM	Best Match Amb	Best Match Para	Best Match Unm
Shape	GS	1	17	17	17	11	11	0.84	0.44	0.76	0.62	0.37	0.01	0.25
	QT	1	5	17	17	26	26	0.87	0.39	0.85	0.52	0.41	0.08	0
	GS	4	17	17	34	16	226	0.78	0.34	0.38	0.52	0.48	0	0.01
	QT	4	5	17	34	26	187	0.83	0.23	0.37	0.55	0.37	0.08	0
Thing	GS	1	50	50	50	18	18	0.86	0.45	0.72	0.38	0.62	0	0.55
	QT	1	6	50	50	50	50	0.92	0.07	0.74	0.73	0.25	0.02	0
	GS	4	50	50	901	36	198	0.63	0.33	0.03	0.24	0.76	0	0.28
	QT	4	6	50	901	64	450	0.77	0.14	0.05	0.23	0.69	0.08	0

Table 1: Results for SHAPE and THING games with Gumbel-Softmax (GS) or Quantized (QT) communication using rules of length l and words of length d, showing number of unique NL concepts (Cons) and phrases (Phrs) and EC words ($\#w$) and messages ($\#m$). TopSim and AMI give only coarse scores. Our method gives the best match score (CBM) and rates of Ambiguities (Amb), Paraphrases (Para), and Unmatched concepts (Unm).

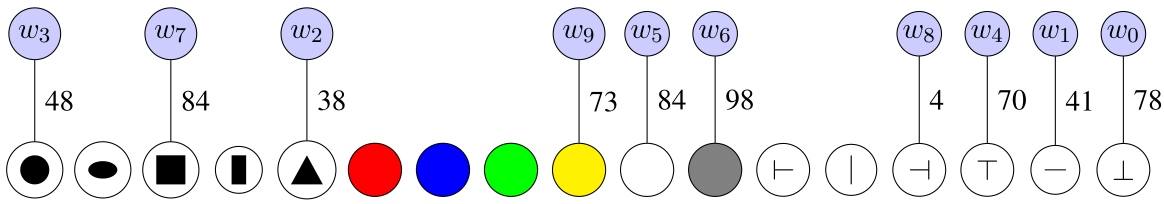

Figure 2: The word $\leftrightarrow$ FVP best-match graph for the SHAPE game (GS communication, $l = 1$). The algorithm matched just 10 concepts to EC words out of 17 possible concepts.

CBM assesses word-to-concept matching, so it is less affected by long messages and rules. However, AMI and CBM are well correlated (Pearson 0.7 and 0.79 on the SHAPE and THING games).

The table also shows the rate of ambiguities and paraphrases, exposed by our method. **Ambiguities** happen when the same EC word is mapped to different concepts. Table 1 shows higher rates of ambiguities in GS compared to QT, and in longer rules. **Paraphrases** are EC words that do not have a best-matched concept node. Paraphrases occur more with QT, which enables 32 and 64 unique EC words for the SHAPE and THING games, while these games have only 17 and 50 NL concepts, respectively. GS allows setting the number of unique words equal to number of concepts, thus is less exposed to this sub-optimal phenomenon.

Finally, **Unmatched concepts** have no matching EC word. They result from a too narrow or underutilized channel. As Table 1 shows, GS communication suffers from a high rate of unmatched concepts (up to 0.5). This phenomenon can be observed by the low # of unique words (18) compared to # of unique concepts (50) for the THING game with GS and $l = 1$. QT has a sufficient number of words and thus no unmatched concepts.

Example Match. Beyond global scores, the CBM provides an interpretable *translation graph*, which maps EC words to NL concepts, facilitating insights on reasons for sub-optimal communication. Figure 2 shows an example graph for the SHAPE game with GS communication.[4] As seen, the algorithm successfully matched 10 EC words to 10 concepts. The sender generated 11 unique words in this experiment, indicating that one word is a paraphrase of an already matched concept, even within this narrow-channel setup.

6 Conclusion

We proposed a new procedure for assessing compositionality of emergent communication. In contrast to other evaluations, our procedure provides a human-interpretable translation map of emerged words to natural concepts. Moreover, our approach provides detailed insights into the reasons for sub-optimal translation. We demonstrated it on two EC games with two different communication types. Our evaluation reveals even that quantized communication performs better than Gumbel-softmax, yet none exhibits compositionality in a level similar to natural language.

[4]The exact configuration is given in Table 1, first row.

Limitations

Our analysis is limited to datasets where gold label phrases exist. The evaluated dataset should be composable and the gold language should use finite set of concepts. Preferably, these concepts can be classified into several categories. We further assume an Oracle function that can divide the objects to targets an distractors during training. Importantly, our approach does not require labels during training.

Acknowledgements

This research was supported by grant no. 2022330 from the United States - Israel Binational Science Foundation (BSF), Jerusalem, Israel. The research of YB was also supported by the Israel Science Foundation (grant 448/20), an Azrieli Foundation Early Career Faculty Fellowship, and an AI Alignment grant from Open Philanthropy. The work of RM was partially supported by the Skillman chair in biomedical sciences and by the Ollendorf center.

References

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *EMNLP*.

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

Boaz Carmeli, Ron Meir, and Yonatan Belinkov. 2023. Emergent quantized communication. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11533–11541.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. 2020. Compositionality and generalization in emergent languages. In *ACL*, pages 4427–4442.

Rahma Chaabouni, Florian Strub, Florent Altché, Eugene Tarassov, Corentin Tallec, Elnaz Davoodi, Kory Wallace Mathewson, Olivier Tieleman, Angeliki Lazaridou, and Bilal Piot. 2021. Emergent communication at scale. In *ICLR*.

Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *NeurIPS*, 30.

John E Hopcroft and Richard M Karp. 1973. An n^5/2 algorithm for maximum matchings in bipartite graphs. *SIAM Journal on computing*, 2(4):225–231.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795.

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

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 2019. EGG: a toolkit for research on emergence of lanGuage in games. In *EMNLP-IJCNLP: System Demonstrations*, pages 55–60.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Alexander Kuhnle and Ann Copestake. 2017. Shapeworld – a new test methodology for multimodal language understanding. *arXiv preprint arXiv:1704.04517*.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. 2018. Emergence of linguistic communication from referential games with symbolic and pixel input. In *ICLR*.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. In *ICLR*.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Fushan Li and Michael Bowling. 2019. Ease-of-teaching and language structure from emergent communication. *NeurIPS*, 32.

Jesse Mu and Noah Goodman. 2021. Emergent communication of generalizations. *Advances in Neural Information Processing Systems*, 34:17994–18007.

Mycal Tucker, Roger P Levy, Julie Shah, and Noga Zaslavsky. 2022. Generalization and translatability in emergent communication via informational constraints. In *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems*.

Ankit Vani, Max Schwarzer, Yuchen Lu, Eeshan Dhekane, and Aaron Courville. 2021. Iterated learning for emergent systematicity in VQA. In *ICLR*.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pages 1073–1080.

Shunyu Yao, Mo Yu, Yang Zhang, Karthik R Narasimhan, Joshua B. Tenenbaum, and Chuang Gan. 2022. Linking emergent and natural languages via corpus transfer. In *ICLR*.

Noga Zaslavsky, Jennifer Hu, and Roger P. Levy. 2021. A Rate–Distortion view of human pragmatic reasoning? In *Proceedings of the Society for Computation in Linguistics*, pages 347–348.

Appendices

A Details on Evaluation Measures

We provide here information on how to calculate the topographic similarity (TopSim) (Brighton and Kirby, 2006; Lazaridou et al., 2018; Yao et al., 2022) and adjusted mutual information (AMI) measures (Vinh et al., 2009; Mu and Goodman, 2021), in the context of EC. Refer to the original papers for full details.

Topographic Similarity. The topographic similarity measures how messages align with the object representations. Concretely, let $\cos_{ij} = \cos(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i \cdot \boldsymbol{x}_j / (||\boldsymbol{x}_i||_2 ||\boldsymbol{x}_j||_2)$ be the cosine similarity of object representations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and $\text{edit}_{ij} = \text{edit}(m_i, m_j)$ be the Levenshtein distance (Levenshtein et al., 1966) between messages m_i and m_j. Let $\text{ncos} = -\{\cos_{ij}\}_{ij}$ be the list of negative cosine similarities, $\text{edit} = \{\text{edit}_{ij}\}_{ij}$ the list of Levenshtein distances, and $R(\cdot)$ the ranking function. Then the topographic similarity is the Spearman rank correlation of the two matrices:

$$\text{Topsim} = \rho(R(\text{edit}), R(\text{ncos})) \quad (2)$$

where ρ is the standard Pearson correlation.

Topographic similarity is a global metric that does not require a reference language and can be easily applied to every EC setup. However, despite its popularity, recent studies found that topographic similarity does not correlate well with success in the task being played (Chaabouni et al., 2021; Yao et al., 2022). More important, it does not provide a mapping between EC atoms and NL concepts, and thus cannot directly assess compositionality.

Adjusted Mutual Information. Given a test set D, let $\mathbb{M}$ denote the set of messages generated by the sender and $\mathbb{L}$ denote the set of NL phrases (e.g., `Red Triangle`, `Blue Square`) exist as labels in the data. The adjusted mutual information (AMI) (Vinh et al., 2009) measures the mutual information between messages and labels, adjusted for chance:

$$\text{AMI}(M, L) = \frac{I(M, L) - \mathbb{E}(I(M, L)}{\max(H(M), H(L)) - \mathbb{E}(I(M, L))} \quad (3)$$

where $I(M, L)$ is the mutual information between M and L, $H(\cdot)$ is the entropy, and $\mathbb{E}(I(M, L))$ is computed with respect to a hypergeometric model of randomness.

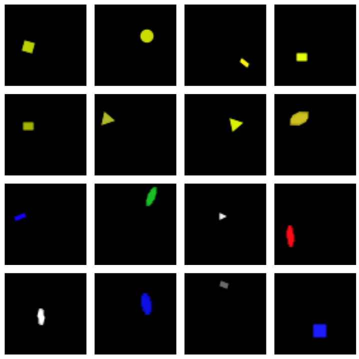

Figure 3: The SHAPE dataset, presenting one turn. Top 8 images are `Yellow` targets. Bottom 8 images are distractors.

B Datasets

We use two datasets in our experiments. Both will be made publicly available.

Shape. A visual reasoning dataset (Kuhnle and Copestake, 2017) of objects over a black background. Our version of the dataset contains four attributes: the shape attribute has five shapes, the color attribute has six colors. To these we added horizontal and vertical position attributes each with three values. Overall, the SHAPE dataset contains 17 concepts, and a maximal labeling rule of 4 FVPs which sum to 270 unique 4-concept labels. The dataset contains many images that obey the same label. See Figure 3 for an example.

Thing. A synthetic dataset of 100,000 objects which we design for controlled experiments. Each object in the dataset has five attributes, each with 10 possible values. Overall, the THING datasets contains 50 concepts and a maximal labeling rule of 5 FVPs. We experiments with label lengths varied from 1, describing a single concept, to 5, which completely describes an object in the dataset.

C Implementation Details

We run all experiments over a modified version of the Egg framework (Kharitonov et al., 2019).[5] In our version the communication modules used by the sender and the receiver are totally separated from their perceptual modules, which we term 'agents'. The Quantized (QT) and Gumbel-Softmax (GS) protocols are implemented at the

[5] Our modification is available at https://github.com/bcarmeli/egg_qtc.

Batch Size	lr	Num Targets	Num Distr	Cell Type	Sender Hidden	Sender Embed	Receiver Hidden	Receiver Embed
10	0.0005	8/20	8/20	LSTM	60	60	40	40

Table 2: Hyper-parameters for the agents in the SHAPE game.

Exp Num	Game	Comm Type	Label Len	word Len	Msg Len	Best Epoch	ACC	AMI	Best Match
1	SHAPE	gs	1	8	4	97	0.829	0.404	0.405
2	SHAPE	gs	1	64	4	96	0.843	0.390	0.339
3	SHAPE	gs	4	8	1	8	0.742	0.490	0.152
4	SHAPE	gs	4	64	1	10	0.748	0.413	0.130
5	SHAPE	qt	1	3	4	86	0.878	0.540	1.401
6	SHAPE	qt	1	64	4	85	0.991	0.000	0.001
7	SHAPE	qt	4	3	1	9	0.768	0.474	0.152
8	SHAPE	qt	4	64	1	12	0.897	0.021	0.009
9	THING	gs	1	8	4	83	0.793	0.102	0.091
10	THING	gs	1	64	4	94	0.798	0.074	0.151
11	THING	gs	4	8	1	97	0.655	0.016	0.052
12	THING	gs	4	64	1	99	0.626	0.014	0.058
13	THING	qt	1	3	4	96	0.838	0.267	0.119
14	THING	qt	1	64	4	100	0.973	0.000	0.014
15	THING	qt	4	3	1	26	0.622	0.020	0.057
16	THING	qt	4	64	1	96	0.844	0.000	0.013

Table 3: Configurations for more SHAPE and THING game experiments. In these configuration we intentionally define sub-optimal communication parameters in order to demonstrate the usefulness of the evaluation metrics. Experiment results are provided in Table 4

Exp Num	Unq Msgs	Unq Wrds	Unq Prs	Ung Cons	Totl Edgs	Good Edgs	Amb Edgs	Phrs Edgs	Totl Cons	Unm Cons	Par Scor	Prc	Rcl
1	258	7	17	17	4k	1621	2379	0	1k	444	0	0.137	0.546
2	442	55	17	17	4k	1356	1724	920	1k	0	0.23	0.144	0.577
3	8	8	34	17	4k	610	390	0	4k	1457	0	0.61	0.152
4	44	44	34	17	4k	520	326	154	4k	0	0.04	0.52	0.13
5	238	8	17	17	4k	1605	2395	0	1k	385	0	0.144	0.577
6	1000	3899	17	17	4k	46	0	3954	1k	0	0.99	0.011	0.044
7	7	7	34	17	4k	607	393	0	4k	1727	0	0.607	0.152
8	974	974	34	17	4k	38	0	962	4k	0	0.24	0.038	0.009
9	543	7	50	50	4k	366	3634	0	1k	807	0	0.048	0.19
10	806	60	50	50	4k	601	3369	30	1k	0	0.01	0.1	0.399
11	8	8	901	50	4k	210	790	0	4k	3332	0	0.21	0.052
12	21	21	901	50	4k	233	767	0	4k	2267	0	0.233	0.058
13	412	8	50	50	4k	478	3522	0	1k	795	0	0.046	0.184
14	1000	3991	50	50	4k	57	0	3943	1k	0	0.99	0.013	0.052
15	11	6	901	50	4k	229	771	0	4k	3483	0	0.229	0.057
16	1000	1000	901	50	4k	50	0	950	4k	0	0.24	0.05	0.013

Table 4: Results for more SHAPE and THING game configurations described in Table 3. In these experiments we intentionally define sub-optimal communication parameters in order to demonstrate the usefulness of the evaluation metrics.

communication layer. For GS we use an implementation provided by the Egg framework, where we do not allow temperature to be learned, and set the straight-through estimator to False. For QT we followed parameter recommendations from Carmeli et al. (2023) and use a binary quantization in all experiments. Both GS and QT uses a recurrent neural network (RNN) for generating multiple words in a message. See Table 2 for RNN details.

The agents for the THING game use fully connected feed-forward networks with input dimension of $d = 270$. Objects are encoded by concatenating five one-hot vectors, one per attribute. There are 50 unique values and 4 communication words (SOS, EOS, PAD, UNK), thus the total object representation length is $d = 270$ The agents of the SHAPE game are implemented with a convolutional neural network similar to Mu and Goodman (2021). Network hyper-parameters are provided in Table 2. In the experiments, we report results when varying the communication elements (word length and message length). Experiments in Table 1 are done with 20 targets and 20 distractors, while experiments reported in Tables 3 and 4 are done with 8 targets and 8 distractors. Our main interest in this work is to evaluate the emerged communication and not to achieve the best possible performance, thus we did not conduct a thorough hyper-parameter search for the agents themselves.

We run all experiments on a single A100 GPU with 40 GB of RAM. Model size is less than 10M. We usually trained the model for 200 epochs which took about 10 hours to complete. Our code will be made publicly available upon de-anonymization.

D Evaluating Sub-optimal Configurations

Here we provide results from experiments with several sub-optimal configurations and demonstrate how CBM identifies these deficiencies. We report configuration parameters for each experiment together with accuracy, AMI, and CBM Score in Table 3. We report all CBM metrics for these experiments in Table 4. Information from the two tables is aligned by the experiment number (first column).

Sub-optimal communication may stem from a channel that is either too wide or too narrow. Narrow channels are restricted by a small number of words and short messages. Wide channels allow long words and/or long messages. Past studies (Tucker et al., 2022) suggest that creating an information bottleneck by narrowing the channel en-

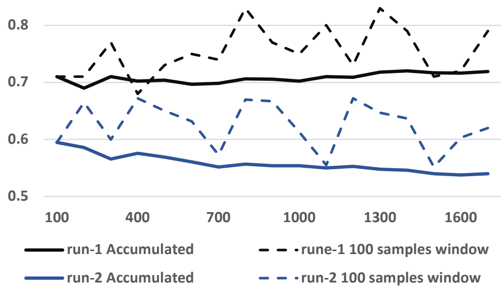

Figure 4: Assessing CBM sensitivity to the size of the evaluation set. We show results for two randomly selected experiments. As seen, the CBM score stabilizes when assessing datasets comprising more than 500 samples.

courages generalization. As seen, CBM scores for most reported sub-optimal configurations are low. Results from experiments 1, 2, and 5 are the only ones for which the CBM score is higher than 0.2. Experiment 6 (SHAPE) achieves almost perfect accuracy (0.991) due to its wide channel. However, the paraphrase score for this experiment is 0.989 showing that the sender uses many different words to refer to the same concept. This phenomenon can also be seen by the large number of unique words (3899) generated for just 17 unique concepts. A similar phenomenon is observed for the THING game (experiment 14).

Experiment 11 (THING) uses a narrow channel. The vocabulary contains only 8 words and the channel allows one-word-long messages. This configuration resulted in 3332 concepts that do not match any word. Interestingly, the number of unmatched concepts remains high even when extending the word length to 64, in experiment 12. We attribute this sub-optimal phenomenon to GS communication deficiencies. The number of unique words generated by the sender in this experiment is only 21 out of the possible 64 words allowed by the configuration. For the SHAPE game, on the other hand, the same wide-channel configuration resulted in 44 unique words generated by the sender for the 17 concepts available in this game and all concepts have matched words.

Precision and recall. After finding the best match, precision and recall between words and concepts can be computed. We define message precision (Prc) and recall (Rcl) to be the number of matched edges in a message, divided by the number of words or concepts, respectively. Precision and recall add insights in situation where the number of words in messages differs significantly from the number of concepts used for labeling the data. Experiment 5 (SHAPE) shows the highest recall, gained partially due to the long message length (four words) allowed by the channel, compared to the label length (one concept) dictated by the data. In contrast, experiment 3 (SHAPE) shows the highest precision. Indeed, this configuration restricts message the length to one word while the data dictates four-concept long phrases.

E CBM Sensitivity to Dataset Size

We evaluated CBM scores on increased data sizes. Figure 4 shows results for two randomly selected experiments. The solid lines indicate the CBM score, computed when accumulating data samples in 100 quintets. The dashed lines indicate CBM score calculated independently for 18 successive data segments of 100 samples each. As seen, scores stabilize for evaluation sets of size larger than 500 samples. Interestingly, the 100-sample scores (dashed) lines are higher than the accumulated scores (solid lines). This is in line with our insight that low evaluation sizes result in higher match scores. Standard deviation for the 18 accumulated measures is 0.009 and 0.016, for the two experiments, respectively. Standard deviation for the 18 segmented measures is 0.044 and 0.040, for the two experiments, respectively.