The Emergence of High-Level Semantics in a Signaling Game

Timothée Bernard
Université Paris Cité
France
timothee.bernard
@u-paris.fr

Timothee Mickus
University of Helsinki
Finland
timothee.mickus
@helsinki.fi

Hiroya Takamura
AIST, Tokyo
Japan
takamura.hiroya
@aist.go.jp

Abstract

The symbol grounding problem—how to connect a symbolic system to the outer world—is a longstanding question in AI that has recently gained prominence with the progress made in NLP in general and surrounding large language models in particular. In this article, we study the emergence of semantic categories in the communication protocol developed by neural agents involved in a well-established type of signaling game. In its basic form, the game requires one agent to retrieve an image based on a message produced by a second agent. We first show that the agents are able to, and do, learn to communicate high-level semantic concepts rather than low-level features of the images even from very indirect training signal to that end. Second, we demonstrate that the introduction of an adversarial agent in the game fosters the emergence of semantics by producing an appropriate training signal when no other method is available.

1 Introduction

How would it be possible to acquire and represent the meaning of words, not simply their function in language but also their connection to the outer world? A cogent account of this question, known as the problem of *symbol grounding*, is that of Harnad (1990). In the case where all we ever have access to is pure linguistic data, Harnad likens the question of attributing meaning representations to a never-ending chain of dictionary look-ups. Harnad's approach to circumvent this problem is to require agents to deal with *iconic* and *categorical* representations, in addition to manipulating symbols. Iconic representations are nonsymbolic representations of perceptual inputs; categorical representations are nonsymbolic representations of categories or concepts. Together, they form the basis of the interface between the agent's symbolic system and the outer world, and it is this interface that gives meaning to, or grounds, the symbols manipulated by the agent. Since Harnad's article, researchers in AI and NLP have often stressed supplementary requirements beyond perceptual data for the development of meaningful and grounded representations, mentioning embodiment (e.g., Steels, 2008), intent (e.g., Bender and Koller, 2020) or interactions with other agents as well as the environment (Chandu et al., 2021). In effect, there is a growing consensus that meaningful representations can only emerge in goal-driven interactive situations.

The study of *emergent communication* is the study of how interacting agents (human or otherwise) can successfully establish effective communication protocols (Kirby, 2002), and under which conditions this is possible. Recently, much interest has been devoted to emergent communication between *neural* agents involved in signaling games (e.g., Lazaridou et al., 2017), in which the agents have to cooperate through information exchange in order to retrieve some target. Such setups have the advantage that they can provide a very tight control on experimental conditions. In the present paper, we focus on a two-agents single-round signaling game, in which the two agents, playing the role of a *sender* and a *receiver*, are to cooperate by exchanging sequences of arbitrary symbols so that the receiver successfully retrieves an image based on one that was shown to the sender. We propose to study what conditions are necessary to the emergence of semantic categories in neural agents in this setting through two sets of experiments.

It has been shown that under certain circumstances, neural agents trained in similar setups develop "trivial" strategies, describing low-level features of their input (Bouchacourt and Baroni, 2018). Accordingly, we hypothesize that in the absence of any form of pressure towards generalization capabilities, the agents will not tend towards conveying high-level information, but will rather settle on exchanging about low-level image-

*Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 200 - 211

specific information. We test this assumption in our first set of experiments by contrasting emergent communication protocols in three different environments: in the first, agents have a direct training signal towards learning to communicate categorical information; in the second, an indirect signal is given, but categorical information is not necessary to solve the task at hand; in the third, agents have no explicit information about categories. To our surprise, we observe that, given enough (training) time, the agents in the second type of environments reliably pick up the indirect signal about the existence of categories and spontaneously shift from communicating low-level features to high-level information (even though they are equally useful to solve the training task). In the third type of environments, semantic categories might be recovered but to a much lesser extent.

This leads us to our second set of experiments, where we study whether a category-level training signal can be synthesized by introducing an *adversarial* agent. This adversary aims to exploit the message sent by the sender to fool the receiver, and thereby implicitly guides the sender away from communicating information that is too easily falsifiable. We observe that introducing such an adversarial agent in the game can significantly bolster the emergence of high-level semantics in the agents' communication.

2 Related works

Grounding, viz., how to relate the symbols of a symbolic system (e.g., a language) to other aspects of the world, has been a fecund domain of research over the past decades. In particular, Harnad (1990) provides an insightful thought experiment, inspired by Searle's controversial Chinese Room argument, and aimed at showing the necessity of grounding: "Suppose you had to learn Chinese as a *first* language and the only source of information you had was a Chinese/Chinese dictionary! This is more like the actual task faced by a purely symbolic model of the mind" (pp.339–40). He also outlines a cogent program towards practical implementations of grounded hybrid systems, involving trained nonsymbolic input and categorical representations interfacing a symbolic system with the outer world.

More recent discussions on this concept have been written by Bender and Koller (2020), who emphasize the role of speakers' intent, or Steels (2008), who stresses the importance of embodied usages of symbols. Note, however, that it has been shown that some structures of the outer world can be found in the topology of the embedding space of ungrounded language models (e.g., Abdou et al. 2021 with color terms). There is now sustained interest in establishing if and how symbol grounding can occur within modern large language models, and to what extent their productions match our expectations for situated, intentional and semantically coherent communication (Patel and Pavlick, 2022; Tenney et al., 2019; Hwang et al., 2021; Ghaffari and Krishnaswamy, 2023). Many works focus on harnessing the boons that come with systems handling multiple channels of inputs, be it to create generalist agents (e.g., Reed et al., 2022; Ni et al., 2021), to enrich their inputs (e.g., Jia et al., 2021), or to facilitate human-robot interactions (e.g., Shichman et al., 2023).

However, practitioners of NLP rarely study the multi-agent aspects of grounding (Chandu et al., 2021), despite them being outlined as a crucial component; Steels (2008) goes as far as stating that standard supervised learning alone, possibly involving multiple modalities but without proper agent-agent or agent-environment interaction, cannot solve the symbol grounding problem. At the same time, there is also a growing interest in using multimodal neural networks as models of how perceptual information is used in humans (esp. Khorrami and Räsänen, 2021; Nikolaus and Fourtassi, 2021); this line of work could therefore benefit from developments on multi-agents NLP system.

In that respect, previous works that include simulations of how language and communication can emerge (Kirby, 2002) is especially useful in that they provide data and define a framework to test hypotheses related to symbol grounding. These works generally involve multiple agents negotiating the use of symbols in order to solve a task through the interaction with nonlinguistic data. While prior work has studied multi-turn communication (a.o., Jorge et al., 2016; Evtimova et al., 2018), populations and generations of agents (e.g., Kirby et al., 2014; Foerster et al., 2016; Ren et al., 2020; Chaabouni et al., 2022) or nonsymbolic communication channels (e.g., Mihai and Hare, 2021), we focus on a straightforward signaling game (Lewis, 1969) involving multiple agents communicating through a symbolic channel (Sukhbaatar et al., 2016; Havrylov and Titov, 2017; Lazaridou et al., 2017, 2018). More precisely, our starting point is the setup of Bernard and Mickus (2023), where

we introduced a computer-generated image dataset, studied the impact of many design choices of the learning process (pertaining to the loss function and regularization, the selection of training instances, and pretraining methods) on a two-agent signaling game, and defined several metrics used to study the properties of the emergent languages.

The work of Mu and Goodman (2021) is close to ours in that they study how the choice of training instances in a signaling game can improve the systematicity of the emergent languages. However, they mainly do so by explicitly strengthening the training signal pertaining to semantic classes (through the use of sets of images instantiating these classes), while we try to achieve similar effects without relying on a priori known semantic classes.

One novelty of the present work is the introduction of an adversary agent in the signaling game. Relevant precedents in the literature include non-cooperative language games, such as the competitive setup of Noukhovitch et al. (2021). To our knowledge, the present work is the first to introduce a GAN-like agent (Goodfellow et al., 2014) in an emergent communication setting.

3 Signaling game definition

We start by presenting the basics of the signaling game that we study in this section. We document departures from this base setup where relevant.

Data. Our dataset (see Bernard and Mickus, 2023) consists of images each depicting an object on a gray background (with varying shade); the objects varies in shape (`cube` or `sphere`), size (`large` or `small`), color (`red` or `blue`), and vertical (`top` or `bottom`) and horizontal position (`left` or `right`). Two images are considered to be of the same category if and only if they agree on these five object features.[1]

We use only 22 of the 32 categories during training (*base categories*), the 10 remaining ones are only used during evaluation (*generalization categories*).[2] Evaluation involves only images not seen during training. More precisely, 20% of each base category is reserved for evaluation; these images plus all images from generalization categories are used during evaluation.

Game definition. We study a signaling game involving a sender, who sees one *original image* I_o and then produces a message m_{I_o}; and a receiver, that receives this message m_{I_o} along with a *target image* I_t and a *distractor image* I_D, and must decide which of the two is the target through the production of a probability distribution over these two images. In such a setting, the relation between the three images involved can provide more or less (even no) signal about the categories to the agents.

Both agents are neural networks that contain a convolutional image encoder; in addition, the sender contains an LSTM message decoder while the receiver contains an LSTM message encoder. We use for these sub-networks the same architectures as Bernard and Mickus (2023). The symbols of the message are selected from a vocabulary of size 16.

We train the receiver to assign a higher probability to the target than to the distractor by minimizing its negative log-likelihood. Writing $p_{\text{receiver}}(I_i \mid I_1, \ldots, I_n, m_{I_o})$ for the probability assigned by the receiver to image I_i based on message m_{I_o} when confronted to images I_1, …, I_n, this loss is:

$$-\log(p_{\text{receiver}}(I_t \mid I_t, I_d, m_{I_o})). \qquad (1)$$

In contrast, the sender is trained with REINFORCE (Williams, 1992) by assigning a reward of value $+1$ to each symbol production action when the receiver correctly retrieves the target, and a reward of value -1 when it fails to do so. For each training batch, the sum of the sender's REINFORCE loss and of the receiver's negative log-likelihood loss is minimized (with RMSProp; Hinton et al., 2012).

4 Influence of target and distractor choice

The goal of the present work is to establish what training signal is necessary for categorical information to appear in the communication protocols developed by the agents. This requires modifying

[1] Two images from the same category may not only differ on background color but also on the position of the light source used to render the scene, and the specific shade of blue/red, vertical and horizontal position, and 3D orientation, of the object.

[2] In Bernard and Mickus (2023), we partitioned the set of categories in such a way that two distinct base categories never differ on just a single feature. This makes it possible for the agents to achieve perfect performance during training while ignoring entirely one of the five features; a possibility they do take advantage of to some extent. In contrast, we are here interested in the agents communicating about category-level information as much as possible, and thus partition the categories differently: we take as generalization categories (`cube`, `small`, `blue`, `down`, `left`)—chosen arbitrarily—and all 10 categories differing from it on exactly three features.

the environment in which the agents interact so that its categorical structure is more or less obvious.

4.1 Three types of environments

If we provide the sender with an image I_o from some category C, then we can either provide as target image I_t for the receiver either this very same image ($I_o = I_t$) or a different image of the same category; the latter option provides a clear signal that different images should be construed as part of the same category. When using the former option ($I_o = I_t$), selecting the distractor image I_d such that it always belongs to a category $C' \neq C$ still induces a training signal pertaining to the categorization of images, albeit a much more indirect one. These two choices compound to three types of environments for our agents, which are illustrated (along with a variant introduced in Section 5.1) in Figure 1.

Direct signal environments. In *direct signal environments*, we provide as the receiver's target image I_t an image randomly sampled from the category of the original image, and select the distractor image I_d from a different category ($I_o, I_t \in C$, $I_d \in C'$ and $C \neq C'$). As a result, the message produced by the sender cannot focus solely on low-level, image-specific, features of the original image (e.g., the average brightness of the image), as they might not match with the target. In other words, the selection of a target image I_t that differs from the original image I_o but shares the same category provides these models with an explicit signal towards learning high-level semantic information. Hence, the performance of these models indicates what sort of communication protocol emerges under optimal conditions for retrieving categorical information.

While a successful game in this environment requires that the receiver be able to derive category-level information in its messages, this does not prevent the sender from describing its input image. Indeed, the sender could go as far as to purely convey enough image-specific information and let the receiver infer the relevant category. This would however arguably lead to a remarkably complex communication protocol, whereas having the sender infer and describe the category ought to lead to a much simpler solution.

Indirect signal environments. Models trained in *indirect signal environments* only differ from those trained in direct signal environments in that the target is exactly the original image ($I_t = I_o$). In this setting, describing low-level, image-specific, features of the original image, such as the background color, is a perfectly viable strategy. We expect this strategy to be favored by the sender as low-level features are intuitively easier to recognize than high-level (category-level) ones (e.g., shape or size of the object depicted).

Remark that in such environment, we still sample the distractor image I_d from a category C' distinct from that of the target image ($C \neq C'$). As such, this environment does provide some means by which categorical information can be recovered: Implicitly, receivers are only ever presented pairs of images that belong to different categories, and may very well learn to segregate them along their categories. This could in turn provide a weak, indirect training signal for the sender. We however expect image-specific information to be more straightforward, although inductive biases in the agents' neural architectures could also shape the emergent communication towards category-level descriptions.

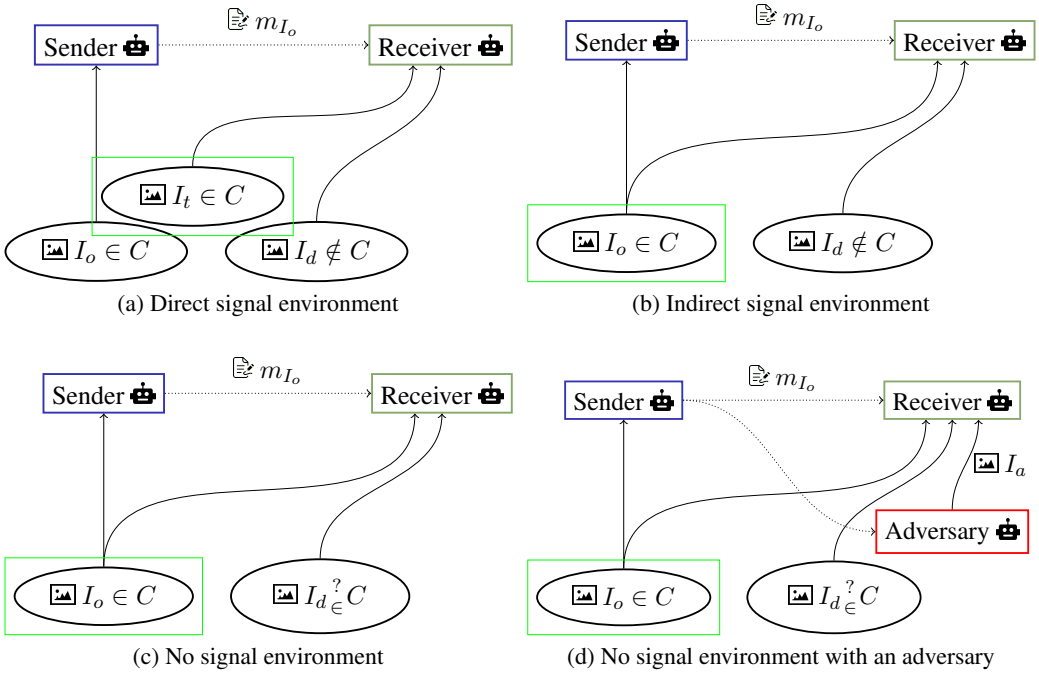
No signal environments. Our ability to train models in direct signal and, to a lesser extent, indirect signal environments hinges on the existence of well-defined semantic categories in our dataset. However, natural pictures of everyday scenes, for instance, do not readily come with such annotations. We therefore also study models that can be trained without such information, so as to determine what are the minimal requirements for non-trivial semantics to emerge. Accordingly, in *no signal environments*, we use the sender's original input image as the target image for the receiver to retrieve ($I_o = I_t$) and select a distractor image at random, regardless of which category it comes from.[3]

In this last type of environment, no training signal about the categories in the dataset is given to the agents. If category-specific information does emerge in the communication protocol, this would have to be pinned on inductive biases present in our architectures.

4.2 Automatic evaluation metrics

To assess whether settings are conducive to the emergence of semantic categories, we use two automated metrics as our primary means of evaluation: *abstractness* and *category communication*

[3] As a result, for some training instances, the target and the distractor belong to the same category.



(a) Direct signal environment

(b) Indirect signal environment

(c) No signal environment

(d) No signal environment with an adversary

Figure 1: The four training setups. C is an image category sampled randomly and uniformly. The green frame indicates which image is the target for the receiver agent. (a)-(c) are introduced in Section 4.1; (d) is introduced in Section 5.1.

efficiency.[4]

Abstractness (abs.). We define this measure as

$$2 \cdot p_{\text{receiver}}(I_t \mid I_o, I_t, m_{I_o}), \tag{2}$$

where I_o and I_t are two images from the same category. This measure quantifies the use of image-specific information by the sender-receiver system: Abstractness scores near 0 indicate that the message m_{I_o} contains image-specific information that the receiver uses to accurately distinguish I_o from I_t, whereas scores near 1 suggest that the message does not include such information.

Category communication efficiency (c.c.e.). We define this measure as

$$p_{\text{receiver}}(I_t \mid I_t, I_d, m_{I_o}), \tag{3}$$

where I_o and I_t are two images of the same category, and I_d is an image of a different category. This measure corresponds exactly to the objective maximized in direct signal environment. It is relevant to make a distinction between *category* communication efficiency and a notion of *image* communication efficiency (i.c.e.), defined as $p_{\text{receiver}}(I_o \mid I_o, I_d, m_{I_o})$, which corresponds to the objective maximized in indirect signal environment.

A sender-receiver system with both low abstractness and low c.c.e. only communicates image-level information (low abstractness) that does not generalize to other images of the same category (low c.c.e.). A system with low abstractness but high c.c.e. communicates at least image-specific information; nothing, however, can be concluded a priori about category-level information because, as two images of the same category tend to be more similar than two images of different categories, image-specific information may be enough to achieve high c.c.e. A system with high abstractness but low c.c.e. does not communicate about image-specific neither category-level information (such a system is not properly trained). Only for a system with both high abstractness and high c.c.e. can we conclude about the emergence of high-level semantics: The system does not communicate image-specific information (high abstractness) but must then communicate category-level information (high c.c.e.).

For finer-grained analyses, we consider other metrics: meaning-form correlation (Brighton and Kirby, 2006), as well as scrambling resistance and

[4]The definitions below are given based on a single evaluation instance; the values reported later are averaged over a large number of such instances.

semantic probes accuracy (Bernard and Mickus, 2023); see Appendix A for further details.

4.3 Experimental results

Training & evaluation procedure Models are used with a baseline term in the sender's loss and no entropy term; we pretrain all image encoders and decoders on an auto-encoding task (without freezing their parameters afterwards).[5] For each of the three environment types, we select the learning rate through a grid search, ran on 10 runs per settings (trained for 200 epochs each; 1000 batch updates per epoch; batches of 128 instances) so as to maximize c.c.e. We then use these optimal learning rates to train 40 models in each environment for 1000 epochs.[6] Each run is evaluated once every 1000 batch updates. Unless otherwise stated, we keep the values of the metrics obtained when the c.c.e. is maximal so as to focus our observations on effective communication protocols, and report medians over the 40 runs for any given setup.

Direct signal environments. We first begin by looking at models trained in direct signal environments (first row of Table 1). We observe very high c.c.e. and abstractness scores; in other words, messages produced by the senders tend to contain only category-level information, and no image-specific information. This is expected, since the receivers in these models are tasked with retrieving a target that is not the original image. We can also point out that these models often develop protocols that appear compositional, even though they likely remain simplistic: They achieve a high scrambling resistance of 82.2% (suggesting that the information carried by a symbol is independent of its position in the message), as well as a relatively high MFC score of $\rho = 0.39$. In line with this analysis, we observe perfect probing accuracy for all features except shape (64.2% accuracy): This suggests that most relevant categorical information is robustly encoded in senders' messages. In short, there is reasonably strong evidence that direct signal environments allow models to learn to link symbols to the values of the five features.

Indirect signal environments. Turning to models trained in indirect signal environments (second row of Table 1), we observe both a very high median c.c.e. score and a high median abstractness score. As pointed out earlier, a high c.c.e. score could be due to the presence of category-level information in the message, but also to enough image-specific information—as two images from the same category resemble each other more than two images from different categories. As for the high abstractness score, it shows that the receiver assigns a similar probability mass to the image based on which the sender produces the message, and to another image of the same category. More precisely, 0.853 corresponds to assigning a probability of $p_{\text{receiver}}(I_t \mid I_o, I_t, m_{I_o})) = \frac{0.853}{2} = 0.4265$ to the target image, and $1 - 0.4265 = 0.5735$ to the original image, i.e., roughly a 4-to-5 odds. Even if two images of the same category resemble each other, they are however clearly distinct from a low-level perspective, and if the sender were sending enough low-level information, it would not be hard for the receiver to confidently distinguish between the original image and another from the same category. Furthermore, the high performance of the semantic probes does confirm that all five high-level features of the images are reliably encoded in the sender's messages. This suggests that the sender mainly conveys category-level information. Our hypothesis—that the sender does not communicate category-level information if other strategies are available—appears thus to be disproved. Furthermore, the fact that the sender only conveys little image-specific information on top of the category-level information it communicates is surprising, as nothing in this setting seems to prevent the sender from communicating more image-specific information (e.g., background color).[7]

Figure 2 shows the evolution of abstractness, c(ategory).c.e and i(mage).c.e. (see Section 4.2) during training in indirect signal environments. We observe that i.c.e converges much more rapidly than c.c.e. and abstractness; the agents learn fairly quickly to communicate about specific images but also gradually shift to communicating about image categories themselves.

Interestingly, even though the messages do contain some image-specific information that ought

[5] Using the notation suggested in (Bernard and Mickus, 2023), the setups considered here correspond to $\langle +P_{AE}, -F, -A, -H, -C, +B \rangle$.

[6] For no signal environments, we report results after 200 epochs as preliminary results indicate further training to have very limited impact.

[7] Somewhat paradoxically, models trained in indirect signal environments obtain a higher median c.c.e. than those trained in direct signal environments, despite the latter being directly trained to maximize c.c.e. scores. This is likely due to the little bit of image-specific information included in the messages along with category-level information, reinforcing the ability of the receiver to recognize the target.

Env.	c.c.e.	abs.	s.r.	semantic probes					MFC
				shape	size	color	h. pos.	v. pos.	
Direct signal	0.986	0.992	0.822	0.642	0.996	0.998	0.999	0.999	0.387
Indirect signal	0.992	0.853	0.949	0.818	0.993	0.993	0.999	0.999	0.439
No signal	0.771	0.511	0.898	0.624	0.869	0.677	0.812	0.754	0.265
No signal + adv.	0.768	0.594	0.859	0.609	0.901	0.601	0.867	0.838	0.243

Table 1: Summary of performances observed at maximal c.c.e., according to the training environment. Direct/Indirect/No signal environments are introduced in Section 4.1; the adversary agent is introduced in Section 5.1.

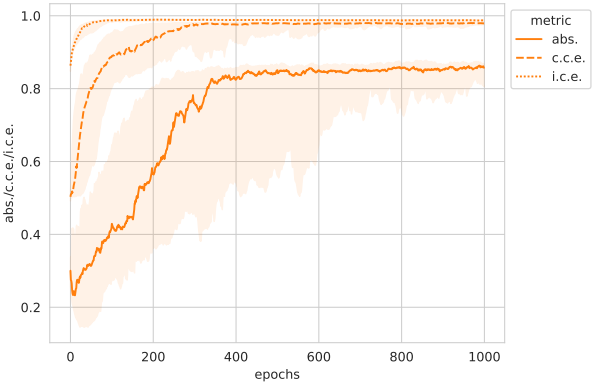


Figure 2: Evolution of the abstractness, c.c.e. and i.c.e. scores over 1000 epochs of training in indirect signal environments. Median over all runs, interquartile intervals shaded; exponential moving average with $\alpha = 0.1$.

to deteriorate MFC scores (as evidenced by the lower than 1 abstractness), the MFC is higher than what we observe for models in direct signal environments ($\rho = 0.439$). This is probably explained by communication protocols in this setting having very high scrambling resistance (94.9%), suggesting that receivers treat messages as orderless bags-of-symbols. Indeed, we compute MFC based on Jaccard indices; therefore, distances between messages are not sensitive to symbol order.

No signal environments. If we now study models trained in no signal environments (third row of Table 1), we can observe a sharp decrease in abstractness, although performances remain non-trivial (an abstractness of 0.511 corresponds to assigning a fourth of the probability mass on a target image of the same category).

Likewise, while it remains firmly above a random chance threshold of 0.5, c.c.e. drops to 0.771. This shows that the sender not only communicates more about image-specific information, but also communicates less about category-level features. Looking at semantic probes accuracy, we find more evidence of the same trend—all probes perform worse than what we saw thus far; shape and color appear especially unreliably encoded.

5 Fostering the emergence of high-level semantics

As we just saw, encoding category-level information systematically seems to require the agents to have access (directly or indirectly) to category-level information. We now turn to whether we can dispense from including this explicit information while retaining category-level information in the messages.

Spike (2017, §.5) suggests that noisy inputs can foster more robust and effective communication channels: Adding noise to input images would prevent agents from communicating about very low-level information (e.g., specific pixel brightness), since this information may not match with what the receiver would perceive. Such a procedure is therefore a natural candidate to explore. However, preliminary experiments involving the addition of normal noise to the images showed this technique to only make the training process less reliable, without any observable benefit.

Instead, we focus on a more involved approach: incorporating an agent playing an adversarial role to discourage the sender and receiver to exchange image-specific information.

5.1 An adversary agent

In this section, we introduce a third agent in the signaling game. This *adversary* agent is implemented with an LSTM message encoder (like the receiver) and a convolutional image decoder. In this setting, the message produced by the sender is also passed to the adversary, which outputs an *adversary image* $I_a = \text{adversary}(m_{I_o})$ intended to fool the receiver. Our intuition is that messages that convey low-level information can easily be counterfeited by this adversary, and therefore should be disfavored by the receiver, and therefore by the

sender—thereby creating an implicit training signal towards communicating high-level semantic information.[8]

As in previous settings, the sender is trained with REINFORCE using rewards determined by the ability of the receiver to distinguish between the target and the distractor only. Unlike in previous settings, however, receivers in adversary settings are trained to distinguish the target from both the distractor and the adversary image by minimizing the negative log-likelihood of the target image considering the three images:

$$-\log(p_{\text{receiver}}(I_t \mid I_t, I_d, I_a, m_{I_o})). \quad (4)$$

We use an adversarial scheme (Goodfellow et al., 2014) to train the adversary to generate an image that the receiver cannot distinguish from the target; i.e., the adversary is trained to minimize the negative log-likelihood of the adversary image:

$$-\log(p_{\text{receiver}}(I_a \mid I_d, I_a, m_{I_o})). \quad (5)$$

To foster the diversity of adversary images, we add Gaussian noise to the output of the adversary's message encoder before feeding it into the image decoder.

To perform the optimization, each agent's loss is scaled by a factor that depends on the agent' performance. Let us define

$$\begin{aligned} s_{\text{sender}} &= p_{\text{receiver}}(I_t \mid I_t, I_d, m_{I_o}), \\ s_{\text{receiver}} &= p_{\text{receiver}}(I_t \mid I_t, I_d, I_a, m_{I_o}), \\ s_{\text{adversary}} &= p_{\text{receiver}}(I_a \mid I_t, I_a, m_{I_o}). \end{aligned}$$

Over the course of training, we compute moving averages of these values, noted "$\hat{s}_a$" for "s_a". Now consider the following values:

$$\begin{aligned} w_{\text{sender}} &= 2 \cdot \hat{s}_{\text{sender}} - 1, \\ w_{\text{receiver}} &= 3 \cdot \hat{s}_{\text{receiver}} - 1, \\ w_{\text{adversary}} &= 2 \cdot \hat{s}_{\text{adversary}}. \end{aligned}$$

Except in pathological situations (that we have not observed), each of these values is nonnegative. These weights are normalized using the softmax function and a "temperature" hyperparameter τ, and then used to scale each of the three losses:

$$\frac{\exp(-w_a/\tau)}{\sum_{a' \in \text{ agents}} \exp(-w_{a'}/\tau)} \cdot \mathcal{L}_a.$$

[8] This adversary agent can also be seen as an auxiliary module of the receiver: one devoted to formulating plausible alternative targets that the receiver has yet to learn to discriminate.

This scaling of the losses (and therefore of the gradients) entails that training focuses on the agents that perform the worst at their task. Note that to avoid updating agents with gradients derived from their adversaries' loss, the losses are not summed: Each agent's loss is minimized by a distinct optimizer that only updates this agent's parameters.

Finally, because image-generation is a particularly challenging task, when the adversary is present, we send the target and distractor images through a pretrained auto-encoder before showing them to the receiver. Indeed, convolution image decoders like the one used to produce the adversary images are very likely to generate visual artifacts that a receiver can easily use to distinguish between neurally generated images and images from the dataset (which lack such artifacts). If the adversary images were to be spotted in this trivial manner, the additional agent would be rendered entirely ineffective. Using auto-encoded versions of the target and distractor images, then exhibiting similar artefacts, we make it technically possible (though still quite challenging) for the adversary to fool the receiver. We implement this auto-encoder using the same image encoding architecture as the sender and receiver agents, and the same image decoding architecture as the distractor agent. This network is trained beforehand and its parameters are frozen during the signaling game.

5.2 Experimental results

Training & evaluation procedure Unless otherwise specified, we rely on the same implementation choices as in Section 4.3. As previously (but with the temperature τ as an additional hyperparameter), we employ a grid search with 10 runs per settings over 200 epochs to maximize c.c.e. We then use these optimal learning rates to train 40 models in each environment (still on 200 epochs as preliminary experiments show that further training brings no improvements).

Adversary agents. In the last (fourth) row of Table 1, we list the performances of models in no-signal environments that involve an adversary agent. Compared with similar environments but without an adversary, we notice a boost in terms of abstractness (from 0.511 to 0.594). This boost is unlikely to be due to random variation only, as indicated by a Pitman permutation test targeting the difference of abstractness scores (p-value $\simeq 0.02$). C.c.e. scores are comparable (the difference is

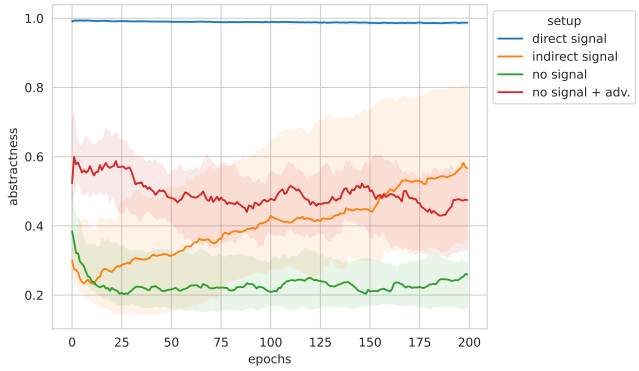


Figure 3: Evolution of the abstractness scores over 200 epochs of training in the four setups studied. For each setup, median over all runs, interquartile intervals shaded; exponential moving average with $\alpha = 0.1$.

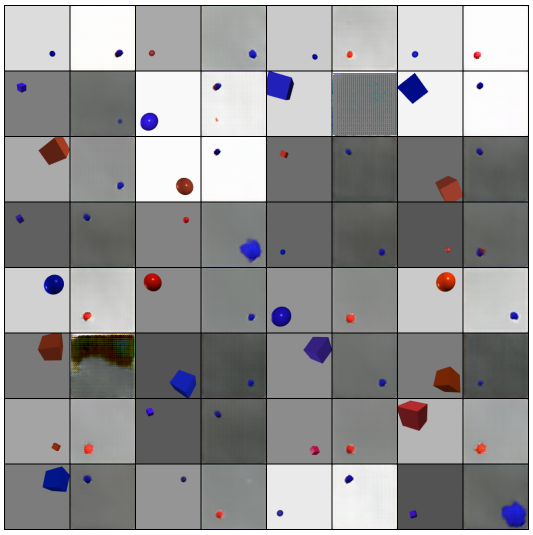


Figure 4: Original and adversary images (no signal environment). Each image in an even column is an adversary image crafted from the sender's message for the original image immediately on its left.

not statistically significant, p-value $\simeq 0.5$), which demonstrates that the presence of an adversary agent tends to remove image-specific information in the sender's messages with no impact on the receiver's ability to retrieve a target selected from the original image category. The accuracy of the semantic probes suggests that the sender and receiver rely less on the color and shape of the object when an adversary is present, and more on its size and its position (both horizontal and vertical).

Figure 3 shows the evolution of abstractness during training in all four setups. The information about categories provided in indirect signal environments has a very progressive effect on abstractness, which starts low and raises gradually. In contrast, the presence of an adversary immediately limits the reliance of the sender and receiver agents on image-specific information. Additional experiments not presented here in details due to space constraints show that in indirect signal environments, while the presence of an adversary agent does not lead to an increase in abstractness in the long run, it clearly fosters higher abstractness scores in the early stages of training.

Adversary images. We include a grid of selected examples from one model in Figure 4. We can observe many images with severe defects, but also that in most images, the background color and even some higher-level features are properly recreated. It is important to keep in mind that the rationale behind introducing the adversary was not to produce high quality images, but to drive the sender and receiver away from communicating only about low-level features of the image, such as the background color. As indicated by the increase in abstractness, this goal has been achieved. These images contribute to explain how: The fact that adversary images often faithfully reproduce the original images' background indicates that the sender and the receiver used to rely on this feature to retrieve the target; the adversary then prevents them from relying only on this feature.

6 Conclusions

Do agents learning to identify images through symbolic communication develop a language able to describe category-level features of these images? Interestingly, indirect signal environments provide evidence that models are able to develop high-level semantics even when the only relevant training signal is extremely tenuous.

The results of models in no signal environments suggest, however, that one cannot expect the sender to encode category-level information systematically without an appropriate training signal.

Our last experiment shows that even without relying on the availability of semantic categories—as is often the case with natural images—, fostering the emergence of high-level semantics is possible via the introduction of an adversarial agent.

In the future, we would be interested in studying whether this technique is effective on other datasets that the one used here, and in whether improvements of the (delicate) training procedure of the adversary may lead to a stronger impact on the emergent languages.

Acknowledgments

We thank Emmanuel Chemla for fruitful discussions on this work.

This work was supported by an Émergence 2021 grant (SYSNEULING project) from IdEx Université Paris Cité.

This work is part of the FoTran project, funded by the European Research Council (ERC) under the EU's Horizon 2020 research and innovation program (agreement № 771113). We also thank the CSC-IT Center for Science Ltd., for computational resources. Preliminary results were obtained from project JPNP15009, commissioned by the New Energy and Industrial Technology Development Organization (NEDO), using the computational resources of the AI Bridging Cloud Infrastructure (ABCI), provided by the National Institute of Advanced Industrial Science and Technology (AIST), Japan.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Timothée Bernard and Timothee Mickus. 2023. So many design choices: Improving and interpreting neural agent communication in signaling games. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8399–8413, Toronto, Canada. Association for Computational Linguistics.

Diane Bouchacourt and Marco Baroni. 2018. How agents see things: On visual representations in an emergent language game. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 981–985, Brussels, Belgium. Association for Computational Linguistics.

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

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. 2020. Compositionality and generalization in emergent languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4427–4442, Online. Association for Computational Linguistics.

Rahma Chaabouni, Florian Strub, Florent Altché, Eugene Tarassov, Corentin Tallec, Elnaz Davoodi, Kory Wallace Mathewson, Olivier Tieleman, Angeliki Lazaridou, and Bilal Piot. 2022. Emergent communication at scale. In *International Conference on Learning Representations*.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online. Association for Computational Linguistics.

Katrina Evtimova, Andrew Drozdov, Douwe Kiela, and Kyunghyun Cho. 2018. Emergent communication in a multi-modal, multi-step referential game. In *International Conference on Learning Representations*.

Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2137–2145. Curran Associates, Inc.

Sadaf Ghaffari and Nikhil Krishnaswamy. 2023. Grounding and distinguishing conceptual vocabulary through similarity learning in embodied simulations.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335 – 346.

Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. 2012. Coursera lectures slides, lecture 6.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Emilio Jorge, Mikael Kågebäck, and Emil Gustavsson. 2016. Learning to play guess who? and inventing a grounded language as a consequence.

Khazar Khorrami and Okko Räsänen. 2021. Can phones, syllables, and words emerge as side-products of cross-situational audiovisual learning? - a computational investigation. *Language Development Research*, 1(1):123–191.

Simon Kirby. 2002. Natural language from artificial life. *Artif. Life*, 8(2):185–215.

Simon Kirby, Tom Griffiths, and Kenny Smith. 2014. Iterated learning and the evolution of language. *Curr. Opin. Neurobiol.*, 28:108–114.

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*.

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

David Lewis. 1969. *Convention: a philosophical study*. Harvard University Press Cambridge.

Timothee Mickus, Timothée Bernard, and Denis Paperno. 2020. What meaning-form correlation has to compose with: A study of MFC on artificial and natural language. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3737–3749, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Daniela Mihai and Jonathon Hare. 2021. Learning to draw: Emergent communication through sketching. In *Advances in Neural Information Processing Systems*, volume 34, pages 7153–7166. Curran Associates, Inc.

Jesse Mu and Noah Goodman. 2021. Emergent communication of generalizations. In *Advances in Neural Information Processing Systems*, volume 34, pages 17994–18007. Curran Associates, Inc.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. 2021. M3p: Learning universal representations via multitask multilingual multimodal pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3977–3986.

Mitja Nikolaus and Abdellah Fourtassi. 2021. Modeling the interaction between perception-based and production-based learning in children's early acquisition of semantic knowledge. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 391–407.

Michael Noukhovitch, Travis LaCroix, Angeliki Lazaridou, and Aaron Courville. 2021. Emergent communication under competition. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, page 974–982, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Roma Patel and Ellie Pavlick. 2022. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. 2022. A generalist agent. *Transactions on Machine Learning Research*. Featured Certification.

Yi Ren, Shangmin Guo, Matthieu Labeau, Shay B. Cohen, and Simon Kirby. 2020. Compositional languages emerge in a neural iterated learning model. In *International Conference on Learning Representations*.

John R. Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–457. 06894.

Mollie Shichman, Claire Bonial, Austin Blodgett, Taylor Hudson, Francis Ferraro, and Rachel Rudinger. 2023. Use defines possibilities: Reasoning about object function to interpret and execute robot instructions. In *Proceedings of the 15th International Conference on Computational Semantics (IWCS 2023)*.

Matthew Spike. 2017. *Minimal requirements for the cultural evolution of language*. Ph.D. thesis, University of Edinburgh.

Luc Steels. 2008. The symbol grounding problem has been solved, so what's next? In *Symbols and Embodiment: Debates on meaning and cognition*. Oxford University Press.

Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2016. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256.

A Supplementary metrics

To further evaluate the communication protocols that emerge from our various models, we rely on abstractness and c.c.e., as well as three metrics previously proposed in the literature.

Meaning–form correlation (MFC), also known as topographic similarity (Brighton and Kirby, 2006), consists in evaluating whether changes in form are commensurate to changes in meaning. The metric was originally proposed as a means of quantifying compositionality, but see Mickus et al. (2020); Chaabouni et al. (2020) for discussions. In our specific case, we use Jaccard distance (Jaccard, 1912) as a form metric and Hamming distance between categories as a meaning distance. Noting $|m|_x$ for the number of occurrences of symbol x in message m, the Jaccard distance between two messages m and m' is defined as

$$1 - \frac{\sum_{x \in Alphabet} \min(|m|_x, |m'|_x)}{\sum_{x \in Alphabet} \max(|m|_x, |m'|_x)}. \quad (6)$$

For instance, the Jaccard distance between "A A B A C" and "A B C D" is $1 - \frac{1+1+1+0}{3+1+1+1}$, i.e., $\frac{1}{2}$. The Hamming distance between two categories c and c' is simply the number of features (i.e., among color, size, shape, h. pos., v. pos.) on which c and c' disagree.

The two other metrics are borrowed from Bernard and Mickus (2023). *Scrambling resistance* (s.r.), quantifies how sensitive to symbol ordering receivers are: Values close to 1 indicate that each symbol is interpreted independently of its position in the message, whereas values close to 0 indicate that the message is only interpreted as a whole. We also rely on *semantic probes* to detect how much each of the five category-level features is communicated in the sender's messages. In practice, they are implemented as a decision tree per feature, trained to predict the corresponding value for the original image based on a bag-of-symbol representation of the sender's message (i.e., a vector in $\mathbb{N}^{16}$).