Collaborative meaning-making: the emergence of novel languages in humans, machines, and human-machine interactions

Kouwenhoven, T.

5

The Curious Case of Representational Alignment

Natural language has the universal properties of being compositional and grounded in reality. The emergence of linguistic properties is often investigated through simulations of emergent communication using referential games. However, these computational experiments have yielded mixed results compared to similar experiments that address the linguistic properties of human language. Here we address representational alignment as a potential contributing factor to these results. Specifically, we assess the representational alignment between the image representations agents have and between agent representations and input images. By doing so, we confirm that the emergent language does not appear to encode human-like conceptual visual features, as the image representations of agents drift away from their inputs while inter-agent alignment increases. We moreover identify a strong relationship between inter-agent alignment and topographic similarity, a common metric for compositionality, and discuss its consequences. To address these issues, we introduce an alignment penalty that prevents representational drift but interestingly does not improve performance on a compositional discrimination task. Together, our findings emphasise the key role representational alignment plays in simulations of language emergence.

Originally published as: Tom Kouwenhoven, Max Peeperkorn, Bram van Dijk, and Tessa Verhoef. 2024. The Curious Case of Representational Alignment: Unravelling Visio-Linguistic Tasks in Emergent Communication. In Kuribayashi, T., Rambelli, G., Takmaz, E., Wicke, P., Oseki, Y., editors, *In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 57–71, Bangkok, Thailand. Association for Computational Linguistics.

5.1 Introduction

Human language bears unique properties that make it a powerful tool for communication. A well-known property is compositionality: the ability to combine meaningful words into more complex meanings (Hockett, 1959). The emergence of compositionality is studied extensively in the field of language evolution through human experiments (Selten and Warglien, 2007; Kirby et al., 2008, 2015; Raviv et al., 2019a, inter alia). A key finding in this field is that the unique nature of human language can be explained as a consequence of a general preference for simplicity and a pressure to be expressive, both of which are imposed during continuous language learning and use (Smith, 2022). Computational simulations of language emergence have also been used to study the emergence of linguistic properties (e.g. de Boer, 2006; Steels and Loetzsch, 2012), and have seen a rising interest in the field of computational linguistics (Lazaridou and Baroni, 2020). Here, compositionality in the emergent communication protocols is commonly measured through a quantitative proxy for compositionality known as topographic similarity (*TopSim*; Brighton and Kirby, 2006). This metric was first introduced to contemporary computational simulations by Lazaridou et al. (2018) and has been used in a large body of work since. Conceptually, this metric gauges whether similar meanings map to similar messages (see Section 5.4.4). Yet, the interpretation of linguistic properties emerging in simulations remains challenging, since language protocols used among artificial agents often show critical mismatches with known properties of human languages (Galke et al., 2022; Lian et al., 2023b) such as efficiency, word-order vs. case-marking biases, or compositional generalisation (see Section 5.2). Only when human-like biases are introduced artificially, do languages with human-like properties emerge (Galke and Raviv, 2025). Consequently, it is evident that the biases of artificial agents in recent simulations and the signal-meaning mappings they make differ from those of humans. This underscores the critical need to obtain deeper insight into referential games in the language learning setting (Rita et al., 2022b).

A possible explanation for these mismatches could stem from representational alignment—the degree of agreement between the internal representations of two information processing systems (Sucholutsky et al., 2023). To the best of our knowledge, representational alignment in emergent communication was first reported by Bouchacourt and Baroni (2018), who measured the degree to which agents aligned their internal image interpretations (inter-agent alignment) by performing Representational Similarity Analysis (*RSA*; Kriegeskorte et al., 2008). Using *RSA* (see Section 5.3), they showed that agents establish successful communication in an artificial manner by aligning their internal image representations while *losing* any relation to the images presented (image-agent alignment). This enabled them to communicate about noise input even though they were trained on real images. As such, their communication protocol captured not conceptual properties of the objects depicted in pictures, but most likely focused on non-human-like spurious image features (e.g., pixel intensities). While inter-agent alignment is not a problem per se, the loss of image-agent alignment is problematic for two reasons. First, for

emergent communication simulations to provide meaningful insights into the emergence of natural human language, agent image representations must be grounded in the content of the images. Only then can we deduce *what* the agents communicate about and assess linguistic properties or their ability to generalise to novel concepts. Second, emergent communication setups have been proposed to fine-tune pre-trained (vision-)language models, aiming to enhance machine understanding of natural human language (Lazaridou and Baroni, 2020; Lowe et al., 2020; Steinert-Threlkeld et al., 2022; Zheng et al., 2024). In this context, maintaining substantial alignment between representations and images is crucial for preserving mutual understanding between machines and humans.

Representational alignment, however, did not receive the necessary attention since a host of papers appeared *after* Bouchacourt and Baroni shared their findings. In these papers, results on referential games were reported without taking *RSA* into account (e.g. Lazaridou et al., 2018; Guo et al., 2019; Li and Bowling, 2019; Ren et al., 2020; Chaabouni et al., 2020; Dagan et al., 2021; Mu and Goodman, 2021; Chaabouni et al., 2022). Admittedly, some use attribute-value objects instead of real images as input. But *importantly*, in nearly all cases, neural agents must map inputs—whether attribute-value objects or image representations—onto agent-specific representations. The problem of inter-agent alignment can, therefore, *always* occur and is *agnostic* to the input type. Although this warrants further analysis of earlier results, the field is already employing referential games in more complex simulations with real images (e.g. Dessi et al., 2021; Chaabouni et al., 2022; Mahaut et al., 2025).

This chapter addresses the understudied alignment problem in standard referential game setups used in emergent communication. We train Reinforcement Learning (RL) agents equipped with a recent vision module (DinoV2; Oquab et al., 2024) to communicate about images. In addition to evaluating the agents on MS COCO (Lin et al., 2014) image pairs, we assess them on noise pairs and image pairs sourced from the Winoground dataset (Thrush et al., 2022). The latter is explicitly created to gauge the visio-linguistic compositional reasoning abilities of vision and language models. We first confirm that effective communication in the referential game relies on inter-agent alignment and then continue with our contributions. First, we find a strong correlation between the degree of inter-agent alignment and the *TopSim* metric. Our second contribution involves a solution to the alignment problem by incorporating an alignment penalty term to the loss, resulting in equivalent communicative success and higher *TopSim* whilst ensuring that the agents communicate about images instead of spurious features. We then argue to start evaluating emergent communication protocols on more stringent tasks that directly target the intuition behind popular metrics to obtain a better understanding of the protocols used. Overall, our results highlight the importance of representational alignment in simulations of language emergence and underscore the need to better understand the divergence in human and artificial language emergence.

5.2 Background

Most research in simulating emergent communication is modelled after the Lewis signalling game (Lewis, 1969) with a speaker and a listener agent. The speaker observes a state (e.g., an image) and sends a signal to the listener, who acts based on this signal. In the case of the referential game, this means selecting a target among a set of distractors. Both agents are rewarded for successful communication, meaning the listener points to the target object. The solution to this game requires the agents to have a shared protocol (i.e., an artificial language), which typically emerges when the agents learn based on trial and error over multiple games. This resembles how, for humans, language learning and use impose constraints such as pressures for learnability and compression that shape our language design (Kirby et al., 2014, 2015). Importantly, the emergent language in the case of simulations with artificial agents is also shaped by biases resulting from, for example, the agent architecture, loss function, and learning protocol (Rita et al., 2022b). The current work uses the referential game: a variant of the Lewis signalling game extensively used to explore language evolution (e.g. Steels and Loetzsch, 2012; Kirby et al., 2015; Lazaridou et al., 2017; Kottur et al., 2017; Lazaridou et al., 2018; Kharitonov et al., 2020; Chaabouni et al., 2022).

An important challenge in emergent communication is that artificial learners often do not behave the same manner as human learners in experimental settings. Some emergent protocols do not follow Zipf's law and thus are anti-efficient unless pressures for brevity are introduced (Chaabouni et al., 2019a), others do not show the word-order vs. case-marking trade-off found in human languages (Chaabouni et al., 2019b; Lian et al., 2021). Additionally, there is an ongoing debate on the degree to which the emergent languages allow for compositional generalisation (Lazaridou and Baroni, 2020; Conklin and Smith, 2023). As such, it has been suggested to introduce communicative (e.g., alternating speaker/listener roles) and cognitive (e.g., memory) constraints (Galke et al., 2022) and use more natural settings to promote more human-like patterns of language emergence with neural agents (Chapter 1). Doing so changes the learning pressures to which the agents need to adapt and can recover initially absent linguistic phenomena of natural language in emergent languages (for a review see Galke and Raviv, 2024). An example of such work, investigating the word-order vs. case-marking trade-off, has successfully replicated this trade-off for neural learners (Lian et al., 2023b). Their setup differs from other work in that agents first learn a miniature language via supervised learning, and then optimise it for communicative success via RL, resulting in emergent languages that share linguistic universals with human language.

To enhance understanding of emergent communication in the Lewis game, Rita et al. (2022b) decomposed the standard objective in Lewis games into two key components: a co-adaptation loss and an information loss. In doing so, they shed light on potential sources of overfitting and how they might hinder the emergence of structured communication protocols. They demonstrated that desired linguistic properties (e.g., compositionality and generalisability)

emerge when they control the listener's ability to converge to the speaker agent (i.e., control for overfitting on the co-adaptation loss). While the co-adaptation loss has parallels to inter-agent alignment, their work does not address the alignment between the agents' image representation and the input features, which we deem crucial in developing grounded communication protocols.

Another challenge in emergent communication is the disentanglement of the underlying meanings of emergent languages. Earlier studies by Lazaridou et al. (2017) suggested that agents assign symbols to general conceptual properties of objects in images, rather than low-level visual features. However, as previously mentioned, follow-up work from Bouchacourt and Baroni (2018) showed this is not always the case. They found that agents align their agent-specific image representations without developing a language that captures conceptual properties depicted in the images. Moreover, agents lost any sense of meaningful within-category variation where two similar objects in human perception (e.g., two avocados) were observed as maximally dissimilar for the agents. In response to these findings, recent studies have implemented sanity checks testing whether trained agents can communicate about noise (Dessi et al., 2021; Mahaut et al., 2025). However, to the best of our knowledge, there has been little attention to what we consider to be their main result: the alignment problem.

5.3 Representational alignment

Representational alignment is the degree of agreement between the internal representations of two information processing systems, whether biological or artificial. Even though widely recognised in cognitive science, neuroscience, and machine learning (Sucholutsky et al., 2023), representational alignment has not seen much interest in the field of emergent communication, except for the work by Bouchacourt and Baroni who analysed the referential game using *RSA*. This metric measures the alignment between two sets of numerical vectors, for example, image embeddings and agents' representations thereof. In practice, it is calculated by taking the pairwise (cosine) distances between vectors of a set and calculating the Spearman rank correlation between these distances.

In this chapter, we also use *RSA* to operationalise representational alignment. Given the speaker image representations r_s of the DinoV2 input embeddings i and r_l as the same images represented in the listener image representation space, we compute the pairwise cosine similarity between the representations for the speaker s_s and for the listener s_l and calculate Spearman's ρ between s_s and s_l. As such, *RSA* measures the degree of inter-agent alignment (RSA_{sl}) between image representations s_s and s_l, relative to their input. Additionally, we use *RSA* to measure image-agent alignment between the speaker and listener image representations and the DinoV2 embeddings (RSA_{si} and RSA_{li} respectively). It is important to stress that representational alignment is *agnostic to the type of input*—being either images or attribute-value objects—and can always happen when inputs are projected onto agent-specific representations.

Now that representational alignment is formalised, we turn to the question of what it means if agents align their representations. Intuitively, a high inter-agent RSA_{sl} value can be interpreted as agents with *similar* representations for similar images. Importantly, this can have two causes: both agents' image representations either *maintain* a relation to the image input (i.e., have a high RSA_{si} and RSA_{li}), or *lose* this relation (i.e., they have a low RSA_{si} and RSA_{li}). While the former is desirable, the latter means that the agents' image representations diverged from their input, but did so in a similar way. Since the agents' image representations are used to compose a message, low image-agent alignment means that they are not communicating about the same high-level image features that are captured by DinoV2, but are likely communicating about non-human-like spurious features. In the case of a low inter-agent alignment (RSA_{sl}) value, something similar happens. This entails that the agents have developed *different* interpretations for the same image, e.g., the speaker maintains a close relation to the input image while the representation of the listener drifts away. While this may be similar to the question of whether people have different perceptual experiences of colour (Locke, 1847), in the case of emergent communication, agents should develop a referentially grounded vocabulary with overlapping concept-level properties since we wish machines to have a more natural understanding of human language. To unravel how representational alignment plays a role in emergent communication, we use *RSA* 1) as a metric to re-assess findings from Bouchacourt and Baroni and 2) implement it as an auxiliary loss to mitigate the alignment problem and ensure that the agents communicate about image features.

5.4 Methods

The standard Lewis referential game is used as provided by the commonly used EGG framework (Kharitonov et al., 2021). This ensures that our findings are representative of this setup, rather than being influenced by specific design decisions. The game implementation is a multi-agent cooperative RL problem where a speaker and a listener communicate to discriminate a target image from two shuffled distractor images. The speaker receives a target image t and generates a message m of at most length L, using vocabulary V. Using message m, the listener guesses which of the two images is the target image ($\hat{t}$). Communicative success is defined as $\hat{t} = t$, meaning that the listener correctly identified the target image among the candidate images. The speaker, crucially, *only* observes the target image and does not see the distractor images. As such, the speaker constructs messages about the target image only and *cannot* construct messages that entail information about differences or similarities between the target and the distractors. Messages and symbols have no a priori meaning but are assumed to obtain meaning and become grounded during the game. Once meaningful, the symbols are ideally combined in a structured manner to create compositional messages that express more complex meanings.[1]

[1] All code, materials, and data are available on OSF: `https://osf.io/9drb5/`.

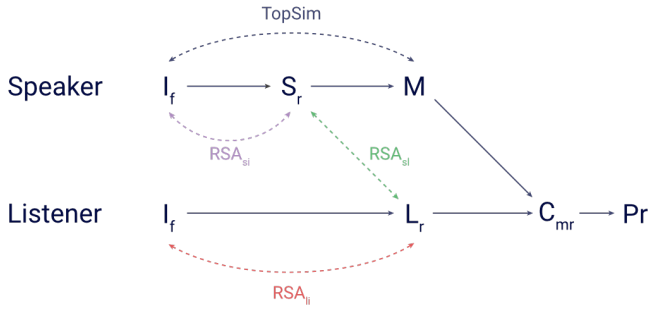


Figure 5.1: An overview of the setup used and the components that are used to calculate our metrics. I_f denotes the image features of DinoV2. S_r and L_r denote the speaker and listener representations of I_f. M is the message, C_{mr} the multimodal representation, and Pr is the probability of an image I_f belonging to message M.

5.4.1 Agents

Agents contain a language and a vision module. The latter consists of a frozen pre-trained visual network (DinoV2) and a learned agent-specific representation layer. While it is difficult to know what conceptual image features are present in DinoV2 embeddings, they have demonstrated capability in semantic segmentation tasks (Oquab et al., 2024), which is similar to the agents' objective. In contrast to the hybrid structure of the vision module, the language module is entirely trained from scratch.

The speaker agent processes images by applying a linear transformation to the image embeddings i_f, followed by batch normalisation, to create its agent-specific image representation r_s (S_r in Figure 5.1). Its language module embeds this representation and passes it through a single-layer Gated Recurrent Unit (GRU; Cho et al., 2014) that spells out messages to describe the target image.

The listener receives the message and the distractor images. It encodes the message into an embedding using another single-cell GRU layer. To obtain an image representation r_l (L_r in Figure 5.1) for each image, the listener agent, like the speaker, applies a linear transformation and batch normalisation on the image embeddings. Finally, temperature-weighted (with a default temperature of 0.1) cosine scores construct a multi-modal representation C_{mr} between the image and message representation (Dessi et al., 2021), where a higher probability (Pr) should be assigned to the target image. The listeners' target distribution comprises the probability for each possible image. Figure 5.1 illustrates the communicative setup and the components used to calculate our metrics.

5.4.2 Optimisation

Communicative success ($\hat{t} = t$) is used to optimise the trainable parameters of both agents. The listener minimises cross-entropy (ce) loss using stochastic gradient descent, amounting to supervised learning. The ce loss is calculated over the listeners' target distribution and thereby provides a direct pressure for communicative success. During inference, the candidate image with the highest probability is chosen as the target $\hat{t}$. The gradients required to optimise the speaker are calculated using the `REINFORCE` (Williams, 1992) update rule as each generated symbol must be assigned a loss. Following standard practice (Rita et al., 2024), entropy regularisation (Mnih et al., 2016) is added to the loss to maintain exploration in message generation.

In addition to the conventional ce loss, we introduce an alignment loss ($ce + RSA$) that includes an alignment penalty term to enforce high inter-agent and image-agent alignment. The term

$$L_{RSA} = (1 - RSA_{sl}) + (1 - RSA_{si}) + (1 - RSA_{li})$$

is added to the ce loss with equal importance. We use TorchSort (Blondel et al., 2020) to calculate L_{RSA}, ensuring that the entire loss term is differentiable. Importantly, L_{RSA} is not influenced by communicative success and does not interact with the ce loss (Section B.2). Only adding RSA_{sl} to the ce loss is not sufficient as high inter-agent alignment can be achieved while *losing* image-agent alignment (see Section 5.3). As such, we also include RSA_{si} and RSA_{li} to ensure that the agents communicate about the content displayed in the images. Including RSA_{sl} entails that representational information is shared between the agents, thus differing from how humans interact. Yet, ranking the speaker and listener representations in calculating RSA_{sl} bears *some* resemblance to projecting beliefs upon the interpretations of the other communicative partner. The current solution should be seen as a step towards more grounded vocabularies prone to refinements such as cognitive plausibility. We train for 30 epochs regardless of the loss used. The hyperparameters (Table B.1) that yielded the best validation accuracy across 42 different communication channel capacities (Section B.1) were used for our findings.

5.4.3 Data

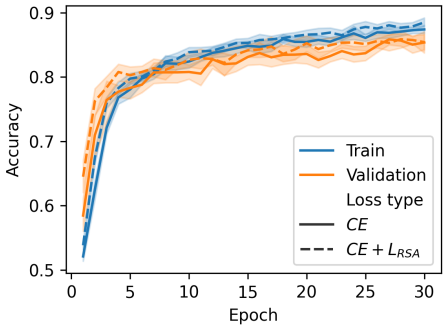
Agents are trained to discriminate MS COCO images but tested on three different datasets (Figure 5.2) to assess out-of-distribution (o.o.d.) performance.

MS COCO – We use a subset of 1200 images from the MS COCO 2017 validation set to train and test the agents using an 80/20 split. To obtain this subset, we first select the categories that contain more than 100 images (resulting in 12 categories) and subsequently sample 100 images for each supercategory present in the resulting set of images. Distractor images are sampled from the same category to ensure that there is *some* relevance to the target image. Sampling these images is done for each batch, meaning targets have different distractors at each epoch.

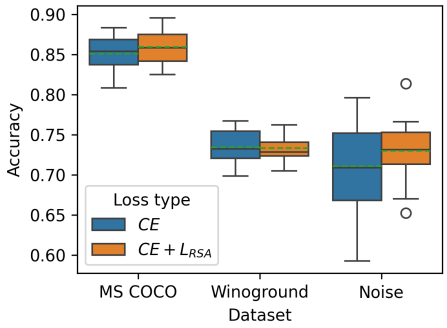
Figure 5.2: Exemplar pairs of each dataset used for evaluation. Left column: an image pair from MS COCO. Middle column: A Winoground example. Right column: A Gaussian noise pair. All images are cropped for display purposes.

Winoground – The Winoground dataset (Thrush et al., 2022) was created to assess the visio-linguistic compositional reasoning abilities of vision and language models. Here, we repurpose it as a proxy for the agents' ability to endow in compositional reasoning for image-based settings. The dataset contains 800 image-caption tuples, comprising 400 Winoground pairs. Image-caption pairs were included when the captions share the same words but are of different *compositions*, implying completely different semantics (e.g., "a tree smashed into a car" versus "a car smashed into a tree" in Figure 5.2 (middle)). As such, discriminating image-caption pairs requires the ability of vision and language models to use compositional language and to understand how language is manifested in the visual modality. Hence, it is posited that to successfully address this task, grounding in images and comprehension of compositional language is imperative. Here, we only use the image pairs, not the captions, and thus test whether RL agents can establish a communicative system that can describe concepts and their compositions. Crucially, this task differs from MS COCO in that the image pairs are *fixed*, *conceptually similar* and meant to be discriminative if the agents' language allows for compositional reasoning and is grounded in the visual modality.

Noise – Following Bouchacourt and Baroni (2018), we test whether agents can communicate about Gaussian noise ($\mu = 0, \sigma = 1$) image pairs when they are trained on real images. If this is the case, it would imply that messages relate to spurious instead of high-level concept features.



(a) The learning curves for the MS COCO dataset on train and validation data.



(b) Communicative performance (Accuracy) during inference on discriminating between two images of different datasets.

Figure 5.3: In (a) we see that the agents learn to communicate successfully without overfitting on the training data. In (b) we see that agents can discriminate MS COCO images but struggle with discriminating Winoground images. Line style indicates the loss type. Results are averaged over 15 seeds, areas indicate the 95% confidence intervals. Green dashed lines indicate averages.

5

5.4.4 Metrics

The performance of our agents is assessed through communicative success (*accuracy*) and the degree of representational alignment is measured using *RSA* (Section 5.3). The degree of compositionality in the emergent language is assessed through the commonly used *TopSim* metric. Formally, *TopSim* is the Spearman correlation between pairwise input distances and the corresponding message distances. As such, it is agnostic to which distance function is used. Input distance can, for example, be computed as attribute-value overlap (when the input space contains categorical attribute-value pairs), or as cosine distance (for continuous input vectors, as is the case in this chapter). The distance between messages is typically calculated as the minimum edit distance. The correlation between these sets of distances is taken as a tendency for messages with similar meanings to have a similar form. However, *TopSim* is relatively agnostic about how these messages are similar, as long as a minimum edit distance captures it. Other metrics for compositionality, such as positional disentanglement and bag-of-symbols disentanglement (Chaabouni et al., 2020), are not straightforward in this chapter due to the continuous nature of the input, i.e., the image embeddings.

5.5 Results

We now present our results, starting with the performance on three datasets, after which we revisit the alignment problem and investigate the relationship between alignment and *TopSim*.

We then show how the alignment penalty term affects communicative success, alignment, and *TopSim*.

5.5.1 Communicative success

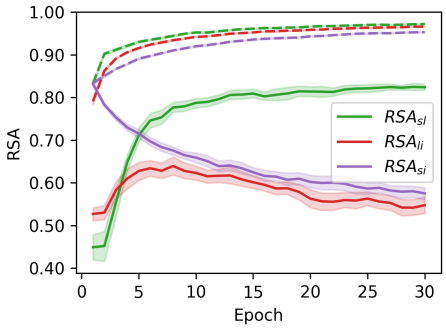
Unsurprisingly, results show that agents can successfully disambiguate between image pairs from MS COCO using an emergent language (Figure 5.3a). Notably, we also confirm previous observations by (Bouchacourt and Baroni, 2018) that agents trained on real images can communicate relatively well about Gaussian noise (Figure 5.3b). Since the speaker must construct its messages purely based on the target image, this suggests that the speaker uses spurious image features to do so. This finding, therefore, again suggests that the emerged languages convey information about spurious features rather than concept-level information. Interestingly, their performance on Gaussian noise is comparable to the performance on Winoground pairs, which requires the messages to capture concept-level properties. This reveals the difficulty of discriminating between strict pairs of conceptually similar images. The observed decrease in out-of-distribution performance aligns with findings from other studies, such as those presented by Lazaridou et al. (2018) and Conklin and Smith (2023) and highlights that generalisation to novel meanings is still difficult for our agents.

5.5.2 The alignment problem

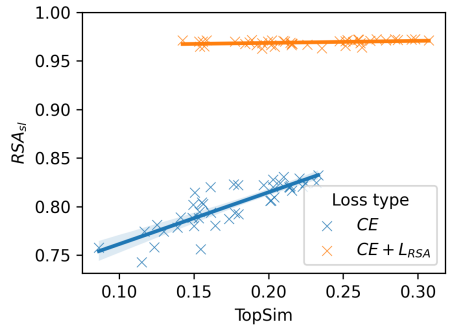
Considering the metrics used to assess representational alignment, the solid lines in Figure 5.4a clearly show that inter-agent alignment increases while alignment sensitivity to image features decreases for both agents. Again, it is in principle not a problem that the agents' image representations align, but it becomes problematic when the alignment between the image embeddings and the image representations declines. Ablations across different channel capacities (Section B.1) and with different pre-trained vision modules (Section B.3) showed that these trends appear consistently and are not influenced by the capacity or type of vision model. In addition to the communicative success on Gaussian noise, this re-confirms that the agents do not learn to extract concept-level information from the image embeddings but instead use the embeddings to solve this task differently.

5.5.3 *TopSim* and representational alignment

Earlier findings show mixed results on the relationship between *TopSim* and generalisation in image-based settings, *TopSim* was either positively related to generalisation (Chaabouni et al., 2022) or not (Rita et al., 2022b). Our results indicate that generalisation and *TopSim* are correlated with both ce ($r = .856$, $p < .001$) and $ce + RSA$ ($r = .767$, $p < .001$) losses. This suggests that more structured languages, as measured using *TopSim*, enable better communication on unseen validation pairs. Moreover, we find a strong positive relationship between RSA_{sl} and *TopSim*



(a) The representational alignment curves for inter-agent image representations (green) and between the image and the sender/listener representations (purple, red).



(b) The relationship between *TopSim* and inter-agent alignment (RSA_{sl}) for both loss types.

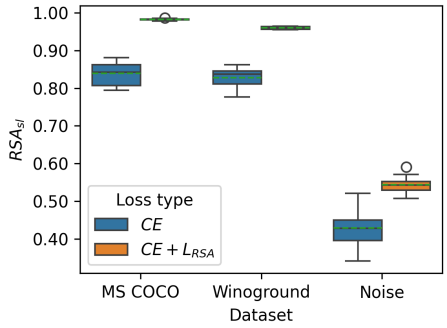
Figure 5.4: In (a) we see that the alignment problem occurs with the ce (solid lines) but not the $ce + RSA$ (dashed lines) loss. In (b) we see that *TopSim* and RSA_{sl} are correlated when the ce loss is used ($r = .838$, $p < .001$). This is also the case with the $ce + RSA$ loss ($r = .408$, $p = .001$) but the effect is decoupled from *TopSim*. Results are averaged over 15 seeds, areas indicate the 95% confidence intervals.

($r = .838$, $p < .001$) in the ce setup (Figure 5.4b). While this relation is also present in the ce+RSA setup ($r = .408$, $p = .001$), it is decoupled from *TopSim* given the (very) small spread ($\sigma = .003$) of RSA_{sl}. Although representational alignment may alleviate the need for discriminative messages, we do not observe an influence of inter-agent alignment on the number of uniquely produced messages.

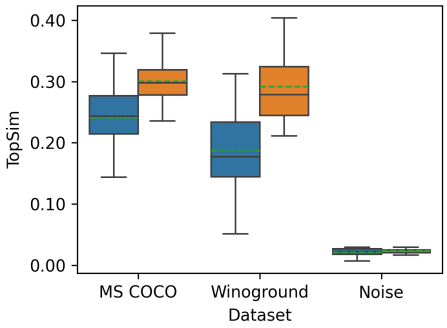
5.5.4 Mitigating the alignment problem

We now focus on the $ce + RSA$ setup, which was introduced to ensure that the agents maintain alignment with the image embeddings. Figure 5.4a and Figure 5.5a show that this indeed happens: inter-agent alignment *and* agent-image alignment increase during training and remain high during inference. Yet, there does not seem to be a benefit for communicative success at inference time as *accuracy* across the datasets remains relatively similar (Figure 5.3b). This is likely because the alignment penalty only forces agents to represent images similarly to the image embeddings and acts independently from the cross-entropy loss used to assess the success of communication (Section B.2). In the case of images containing Gaussian noise, we still observe above-chance performance, which suggests that communication between the agents still occurs in an artificial manner.

In addition to increased representational alignment between agents, the alignment penalty also leads to increased *TopSim*, which suggests that the messages used during communica-



(a) Inter-agent representational alignment (RSA_{sl}) across different datasets.



(b) Topographic similarity (*TopSim*) between the images and the messages displayed for different datasets.

Figure 5.5: In (a) we see the effect of the loss function on the degree of inter-agent representational alignment. In (b) we see that *TopSim* increases as a result of the $ce + RSA$ loss.

tion have a higher degree of structure (Figure 5.5b). Given the higher values of RSA_{sl}, this strengthens our finding that *TopSim* and inter-agent alignment are related. This suggests that the observed variations in *TopSim*, whether higher or lower, as noted in previous studies (e.g. Kottur et al., 2017; Chaabouni et al., 2020), should not be interpreted without considering representational alignment since they may be attributable to this underlying artefact rather than alterations to the original setup.

When tested on more strict Winoground pairs, communicative success does not improve as a result of using the alignment penalty (Figure 5.3b). Given the correlation between *TopSim* and generalisation that was observed earlier, this is surprising since the higher degree of *TopSim* should imply that the language is more structured. Moreover, both, RSA_{si} and RSA_{li} have not drifted away from the image features (Figure 5.4a). This combination, *in theory*, should be ideal for discriminating image pairs from the Winoground dataset since it was designed to be discriminative with compositional visio-linguistic reasoning. However, in *practice* this is not the case.

5.6 Discussion

In this chapter, we revisited the representational alignment problem in a common setup used in emergent communication and proposed a solution to this underrepresented problem. We corroborated earlier findings by demonstrating that agents align their image representations and rely on spurious image features instead of human-like concept-level information (Bouchacourt and Baroni, 2018). We then showed that inter-agent alignment strongly correlated with the

commonly used *TopSim* metric. Our solution to the alignment problem involves an alignment penalty that forces the agents to remain aligned with the input features, thereby mitigating the alignment problem without compromising communicative success. Finally, when agents are tested on more challenging Winoground pairs, we observed reasonable but lower performance regardless of whether image representations were similar to the image embeddings or not. With this work, we hope that the alignment problem will receive more attention in the field of emergent communication, as is already the case in adjacent fields (Sucholutsky et al., 2023).

5.6.1 Importance of representational alignment

It is common practice in simulations of emergent communication to process (visual) inputs into an agent-specific hidden representation and update their weights simultaneously (e.g. Lazaridou et al., 2017; Bouchacourt and Baroni, 2018; Chaabouni et al., 2019a, 2020; Rita et al., 2022b). As such, inter-agent alignment, *irrespective of the input form*, likely happens in other simulations too. This phenomenon is therefore potentially widespread and can perhaps be the cause for findings that are at odds with experimental findings. This bears much similarity to a concept known as shortcut learning: a form of understanding that is in many ways not human-like, but introduces a new "alien" kind of problem-solving (Schwartz and Stanovsky, 2022; Mitchell and Krakauer, 2023). While it is not always the case that the representation structure we *expect* to help solve a task will do so (e.g. Montero et al., 2021; Xu et al., 2022), such discrepancies may hinder the use of emergent communication models in developing a more natural understanding of human languages and leave them less suitable for directly simulating language evolution phenomena. Especially so if we want machine representations of natural language to align with human representations (Sucholutsky et al., 2023). *RSA* should therefore be used to rule out, or at the bare minimum report about, representational alignment in the future.

5.6.2 Relating *TopSim* and representational alignment

Measuring representational alignment using *RSA* is similar to how *TopSim* measures the structure in messages. While they differ in their inputs, they both calculate the Spearman-ranked correlation between metric-agnostic pairwise distances. Crucially, the input makes all the difference; the inputs for *RSA* are from both agents and are trained independently, whilst *TopSim* only assesses the relation between the fixed inputs and learned output (Figure 5.1). Despite the similarities, the metrics thus describe different phenomena and are rarely reported simultaneously.

We hypothesise that the relationship between *TopSim* and inter-agent representational alignment is a by-product of the setup, which in essence implies that the listener has to align its representation r_l to the speaker representation r_s (Rita et al., 2022b). It has to do so using only the speakers' messages, being a compressed abstraction of r_s. A possible solution to

this problem is to align representations, which eases the listeners' training objective. If the speaker consistently produces structured messages during training, aligning r_l with r_s becomes easier, thereby leading to higher inter-agent alignment. Essentially, this renders *TopSim* to be an *indirect* metric for the rate of alignment, for which RSA_{sl} is a *direct* metric. In the context of learnability, the relationship between *TopSim* and inter-agent alignment and the fact that alignment always occurs can be seen as reasons for why languages with higher *TopSim* are easier to learn (Li and Bowling, 2019; Cheng et al., 2023). This underscores the need to report inter-agent representational alignment to avoid conclusions drawn about the effect of specific interventions on *TopSim* which may be attributable to inter-agent alignment.

5.6.3 Targeted o.o.d. evaluations

An important implication of our findings concerns the standard practice of reporting o.o.d. accuracy where the agents are tested on unseen input after training (e.g. Auersperger and Pecina, 2022; Conklin and Smith, 2023). In essence, doing so should inform us about the agents' ability to generalise from one dataset (e.g., MS COCO) to another dataset (e.g., the Winoground pairs), much like human language allows us to talk about an infinite number of situations. Crucially, this overlooks the representational alignment problem in that we do not know *what* the agents are precisely generalising about. This problem can be mitigated using the alignment penalty term to assess generalisation more directly, or at least should be taken into consideration.

We assessed o.o.d. performance on the more challenging Winoground pairs as a proxy for the agents' ability to endow in compositional reasoning for image-based settings. Good performance on the Winoground dataset requires a grounded language that can be used to create compositional messages since the objects and their underlying relations need to be described. In general, we suggest starting to evaluate simulations of referential games on targeted, strict tasks, such as probing state-of-the-art vision language models on, for example, visio-compositional (Thrush et al., 2022; Diwan et al., 2022; Hsieh et al., 2023; Ray et al., 2023) or spatial (Kamath et al., 2023) reasoning tasks. Re-purposing such datasets can reveal more directly whether agents develop the attested communicative abilities that are trivial to humans without having to rely solely on metrics. Our results illustrate this through a shortcoming of the *TopSim* metric. We observed that agents still struggle with distinguishing pairs of *conceptually similar* Winoground images, even though *TopSim* is higher with the alignment penalty. If the language protocol were to communicate concept-level information *and* compositional messages were created, we should not observe this struggle, meaning that the emerged protocols do not enable human-like communicative success.

Interestingly, the o.o.d. performance remains substantially above chance in the $ce + RSA$ setting. Given that MS COCO is not a dataset for learning to model compositionality, this delineates the limits of what can be achieved qua performance based on MS COCO image features in the Winoground context. Nevertheless, this leaves open the question of the above-

chance performance on Gaussian noise with the $ce + RSA$ loss. A tentative explanation is that the higher inter-agent alignment on noise input ($M_{ce} = .428$, $M_{ce+RSA_{sl}} = .543$, $t = -8.71$, $p < .001$) alleviates part of the problem (Figure 5.5a). To validate this, future experiments should involve controlling the prior distributions of the agents' image encoders by training their vision modules on different data. Doing so ensures that they have to communicate about novel objects and cannot rely on similar representations.

5.7 Conclusion

This chapter revisited the underrepresented alignment problem present in the referential game often used in simulations of emergent communication. Specifically, we focused on the problem of increasing alignment between agent-image representations in combination with a decreasing alignment between the input and agent representations. We first confirmed that agents align their image representations while losing connection to their input, meaning that the emergent languages do not appear to encode human-like visual features. We then showed that, in the common setup, inter-agent alignment is related to topographic similarity, and argued that this renders *TopSim* an *indirect* metric of the rate of inter-agent alignment. To further investigate the effects of alignment, we introduced an alignment penalty to mitigate the alignment problem. We showed that the communicative ability on a strict compositionality benchmark did not improve, leaving the question of inducing compositional generalisation in emergent communication for images unsolved. Our findings underscore the need to better understand the divergence between human and artificial language emergence within the prevalent referential setup and highlight the importance and potential impact of representational alignment. We hope that future work rules out or at least reports about representational alignment.

5.8 Limitations

Our work has a few notable limitations. First, it only involves the referential game. Another popular variant, the reconstruction game (e.g. Chaabouni et al., 2019a, 2020; Lian et al., 2021; Conklin and Smith, 2023), requires the listener to reconstruct the input object based on the speaker's message. Since this setup has a different objective and presents different learning biases, it may have different results. We still expect the results to be similar as there is no pressure to retain alignment between the image input and agent representation. It would, however, be interesting to investigate whether the language protocol in this scenario is more structured than in the referential game.

Another limitation in our setup is that we only consider the scenario with two agents, which may be a requirement for alignment to be possible. Since experiments with human participants show that larger communities create more systematic languages (Raviv et al., 2019b), simulations

on emergent multi-agent communication with populations of agents are also conducted, but these yield mixed results. The emergent communication protocols oftentimes do not evolve to be more structured unless explicit pressures such as population diversity or emulation mechanisms are introduced (Rita et al., 2022b; Chaabouni et al., 2022). However, Michel et al. (2023) showed that population setups can result in more compositional languages if agent pairs are trained in a partitioned manner to prevent co-adaptation. Despite the mixed results, we believe that emergent communication with populations of agents is ecologically more valid and could result in different alignment effects. Much like how Tieleman et al. (2019) showed that autoencoders encode better concept category representations when they learn representations in a community-based setting with multiple encoders and decoders collectively.

The final limitation of our study regards its scale. While simulations of emergent communication are typically conducted on relatively small-scale datasets, human language emergence is accompanied by rich and diverse multimodal experiences. Recent results in the field of computer vision suggest that dataset diversity and scale are the primary drivers of alignment to human representations (Conwell et al., 2023; Muttenthaler et al., 2023). As such, this key difference between the setting of artificial emergent communication and human language emergence can drive the observed differences in representations. Due to the difficulty of interpreting these representations, we see this as another reason to evaluate emergent protocols on more strict datasets with clear pragmatic value for humans.