Title

Variability in communication contexts determines the convexity of semantic category systems emerging in neural networks

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Nedelcu, Vlad C
Lassiter, Daniel
Smith, Kenny

Publication Date

2024

Peer reviewed

Variability in communication contexts determines the convexity of semantic category systems emerging in neural networks

Vlad C. Nedelcu (v.c.nedelcu@sms.ed.ac.uk), Daniel Lassiter, Kenny Smith
School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, EH8 9AD

Abstract

Artificial neural networks trained using deep-learning methods to solve a simple reference game by optimizing a task-specific utility develop efficient semantic categorization systems that trade off complexity against informativeness, much like the category systems of human languages do. But what exact type of structures in the semantic space could result in efficient categories, and how are these structures shaped by the contexts of communication? We propose a NN model that moves beyond the minimal dyadic setup and show that the emergence of convexity, a property of semantic systems that facilitates this efficiency, is dependent on the amount of variability in communication contexts across partners. We use a method of input representation based on compositional vector embeddings that is able to achieve a higher level of communication success than regular non-compositional representation methods, and can achieve a better balance between maintaining the structure of the semantic space and optimizing utility.

Introduction

Recent work has attempted to understand whether the regularities and structural properties found in human language are also found in the representations that emerge between neural networks placed in similar communication settings. These properties include compositionality (Guo et al., 2019), word ordering preferences (Chaabouni et al., 2019), and efficiency in the sense of optimizing a trade-off between complexity and informativeness (Chaabouni et al., 2021). Artificial neural networks (NNs) have only recently started being used for understanding and modelling human language, where more explainable learning methods are typically preferred as they allow to easily pick apart individual factors of interest and test how they determine system behavior (L Griffiths et al., 2008). Of particular importance are models based on Bayesian inference, which allow for elegant formulations at the computational level and benefit from the explicit use of prior assumptions. However, one important limitation of Bayesian formulations is that they are often intractable (Van Rooij et al., 2019), which raises a more important scalability problem: do the results of these models hold in larger and more naturalistic settings? State of the art NN architectures trained using deep-learning methods are known to tractably scale to natural languages (Hawkins et al., 2022) and thus could offer a viable way of modelling more complex situations.

In the domain of semantic categorization, work by Chaabouni et al. as well as Kågebäck et al. has demonstrated that emergent category systems in a color-communication task trade off complexity against informativeness (i.e. they are distributed along the optimal frontier of category systems which are as simple as possible for a given level of informativeness, and as informative as possible for a given level of complexity), closely matching the distribution of color naming systems found in human languages. Chaabouni et al. conclude from these results that the root of efficiency in semantic categorization is not related to specific biological constraints of the human mind, and point to the negotiation of signalling conventions on a discrete signal channel as a possible key factor. However, beyond these observations, prior work that uses NNs to study categorization does not address the question of what exact type of structures or regularities in the semantic space could result in efficient categories. At the same time, these studies lack a clear mechanistic account of how categories actually become efficient. Specifically, they only investigate rather minimal convention-formation processes.

These concerns have been addressed somewhat by studies that use Bayesian modelling or artificial language learning experiments. Some of the mechanisms found to be possibly involved in the evolution of efficient semantic categories include cultural transmission (Carstensen et al., 2015), communication (Carr et al., 2020), the interaction of these two (Silvey et al., 2019), as well as rapid adaptation to partners with divergent communicative needs (Nedelcu et al., 2023). In terms of regularities that could facilitate this efficiency, Nedelcu et al. showed using Bayesian simulations that in heterogeneous communities, where communicatively relevant semantic distinctions differ across a focal individual's communicative partners, category systems evolve to reflect the structure of the semantic space, with categories forming convex regions in the meaning space. Meanwhile, in homogeneous communities, languages tend to evolve so as to reflect the specific communicative needs of these communities, possibly compromising on perfect convexity when a non-convex system that uses fewer categories can satisfy these needs. The fact that a social factor like the partner adaptation mechanism could shape the structure of categories would be in line with the hypothesis by Wray & Grace (2007) that some of the linguistic features regarded as universal today (e.g. compositionality) may have only evolved in response to changing social structures. Specifically, they argue that as communities expanded and became more heterogeneous, in the sense that inter-group interactions would have become increasingly fre-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

quent, interlocutors could not consistently rely on common ground to make themselves understood, so they had to express their messages in a more systematic manner, making as few assumptions as possible about their communication partner. These conditions would have thus favoured the evolution of regular, context-independent features.

However, one important limitation of the Bayesian study by Nedelcu et al., echoing the aforementioned critiques of Bayesian models of these processes, is that the explored communication setting is very limiting: the meaning space consists of 12 meanings, agents are allowed to use a maximum of 3 categories, and the contexts in which they communicate are significantly constrained to ensure computational feasibility. Critically, these parameters also do not allow the placement of each meaning in a separate category, forcing the agents to make generalizations in order to solve the task effectively. In this paper, we set to explore the partner adaptation mechanism and the conditions in which convex categories emerge using deep-learning methods, and to test more complex settings. To do so, we move beyond the minimal dyadic setup and study reference games with multiple agents.

Another way in which we explore convexity is in the context of a trade-off faced by NN agents trying to find a solution to a communicative task: the tension between maintaining the structure of the input space (a form of sticking to a prior "bias"), on the one hand, and optimizing utility (with the risk of overfitting and not being able to generalize effectively), on the other. In models of categorization the meanings to be communicated are typically placed in a geometric space and the metric over such a space is related to similarity between meanings (Mollica & Zaslavsky, under review). In such a semantic space, the emergence of a categorization system with perfect convexity is a reflection of agents strongly maintaining the structure of the input space. We are interested in exploring whether the method of input representation will bias the agents in one of the two directions. We argue that in their model of color categorization, Chaabouni et al. use a type of representation which strongly favors categorizations that reflect the input space. At the other extreme are representations that use one-hot encodings, where there is no similarity between meanings (beyond identity), yielding a completely unstructured input space. We use a form of compositional vector embeddings as a middle ground between the two approaches: binding the two features directly using a literal composition instead of sending them separately means that the composed structure of the object space is more inherent in the input representation given to our agents; while the larger number of parameters resulting from the larger input representation allows agents to make more complex and flexible predictions that can suppress some of the power of the geometric space.

Model

Semantic space

We consider a semantic space that consists of 2 distinct features, each having 20 possible values, for a total of 400 distinct meanings. The size of this space is similar to that of the World Color Survey (which has 330 color chips) and allows us to compare our results more easily to the work on color categorization of Chaabouni et al. and Kågebäck et al..

We represent meanings as compositional vector embeddings based on the tensor product representation method proposed by Smolensky (1990). We first generate separate vector embeddings for each possible value of each semantic feature, using a strategy that captures the similarity relations in the space. Specifically, the vector embeddings associated with the first value ($v = 0$) of each feature will be filled with 20 random numbers from a normal distribution with mean 0 and variance 1. To generate the vector associated with value $v = n$ we take the embeddings corresponding to $v = n-1$ and randomly resample without replacement one of the 20 numbers. This strategy makes sure that for each of the two features the vector embeddings associated with two neighbouring values only differ in 1 number. We calculate cosine similarity between the resulting vectors to confirm we are preserving the ordered relation. Finally, to represent a meaning characterized by features $[x, y]$ we take the vector embeddings of x and y and combine them using their outer product; the rows of the resulting matrix are concatenated to produce the final representation (see Figure 1).

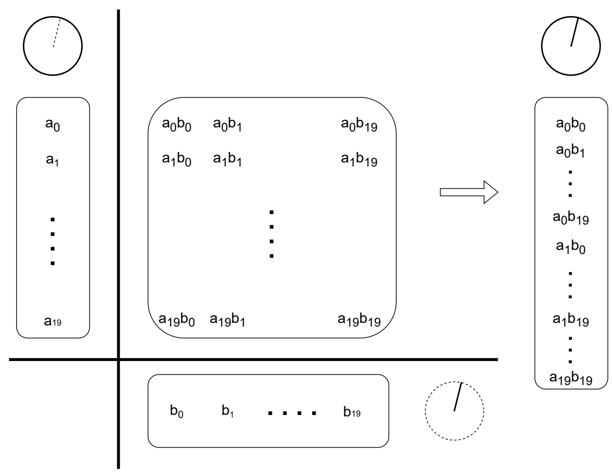


Figure 1: Exemplification of how the compositional input representation of one meaning is calculated. In the figures, we represent meanings using "Shepard circles" (Shepard, 1964) where the two features are the radius of the circle (X axis) and the angle at which the line is oriented (Y axis).

Communication game

Our setup consists of an asymmetric reference game, where a sender interacts with four receivers over a given number of rounds. The sender is presented with the representation of a target object o_t randomly selected from the set of objects O and sends a word from its vocabulary V (with $|V| = |O| = 400$) to help the current receiver identify o_t from a context containing o_t and a distractor o_d. The sender interacts with its partners in separate blocks, starting with the first receiver, then moving on to the next one, until it has interacted with all four. Critically, each receiver has a different set of associated contexts from which one $[o_t, o_d]$ is randomly drawn for a given round. The extent to which these sets of contexts are similar or different across the receivers will determine if the communication scenario is homogeneous or heterogeneous.

To be able to model the partner adaptation mechanism, we extend the EGG framework (Kharitonov et al., 2019) to support games with more than two agents, and change the method for sampling communication contexts to allow partners to sample from different contexts.

Both the sender and the receiver networks are feed-forward: the sender is comprised of three linear layers of size 250 with leaky ReLU activation functions, while the receivers consist of a single simple layer of size 20. The sizes of the networks were chosen after preliminary simulations with a more comprehensive parameter exploration left for future work. The receivers are optimized using ADAM. As we cannot backpropagate through the discrete channel into the sender, we instead optimize using Gumbel-Softmax (GS) sampling with a temperature of 1.5.

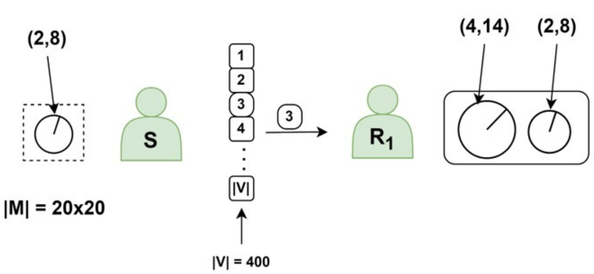


Figure 2: One round of the reference game: a sender is faced with a meaning, selects a signal from its vocabulary that is then sent to a receiver who has to decide which of the two meanings in its context the sender is referring to.

The context sets determine which meanings must be distinguished from each other during communication, and are generated as follows: the 20x20 input space (i.e., 400 objects) is divided into 16 equal sections of size 5x5, so that any two objects situated in the same section are never placed in the same context; the 16 sections are then split into 10 groups, with 12 of the sections randomly chosen to form groups of size two and the remaining 4 sections forming groups of size one, thus creating a semantic configuration (see Figure 3 for an example configuration); the set of contexts for a receiver will finally be generated exhaustively from the semantic configuration associated with that receiver with the added constraint that two objects from the same group cannot be placed in the same context. For a homogeneous audience, all four receivers are attributed the same configuration, whereas for a heterogeneous one, they are each attributed a different configuration.

While rather artificial, this context generation method has two major benefits. First, it reflects the observation that real-world categories tend to have what Rosch & Mervis (1975) call a "family resemblance" structure, with meanings that are part of the same category sharing graded similarity on multiple dimensions. This is possible because humans do not always need to distinguish between objects that are slightly different, as otherwise, categories would need to encode very fine-grained distinctions that are not cognitively feasible and, in an extreme case, would need to contain a single meaning to fully satisfy the communicative needs of its users. Thus, our sections could be considered a minimal granularity below which meanings need not be discriminated for successful communication. Second, the random grouping of sections means that, depending on the details of that grouping, agents would not need to distinguish between meanings from some non-neigbouring sections, enabling them to lexicalize semantically distant meanings with the same word. Crucially, this can be done without any negative consequences for communication, but would result in the emergence of non-convex categories. The idea that agents might need to explicitly discriminate between similar concepts but not between very dissimilar concepts is consistent with studies on homophones. Corpus studies show that semantically highly similar homophones are dispreferred by natural languages (Dautriche et al., 2018). Experimental work also shows that children's ability to learn homophones depends on context, with pairs of homophones being more learnable when the two meanings associated with the phonological form are more semantically distant (Dautriche et al., 2018). This is argued to be a consequence of natural contexts already providing enough information to implicitly distinguish very dissimilar concepts.

```
[[ 1.  1.  1.  1.  1.  2.  2.  2.  2.  2.  1.  1.  1.  1.  1.  4.  4.  4.  4.  4.]
 [ 1.  1.  1.  1.  1.  2.  2.  2.  2.  2.  1.  1.  1.  1.  1.  4.  4.  4.  4.  4.]
 [ 1.  1.  1.  1.  1.  2.  2.  2.  2.  2.  1.  1.  1.  1.  1.  4.  4.  4.  4.  4.]
 [ 1.  1.  1.  1.  1.  2.  2.  2.  2.  2.  1.  1.  1.  1.  1.  4.  4.  4.  4.  4.]
 [ 1.  1.  1.  1.  1.  2.  2.  2.  2.  2.  1.  1.  1.  1.  1.  4.  4.  4.  4.  4.]
 [ 5.  5.  5.  5.  5.  6.  6.  6.  6.  6.  7.  7.  7.  7.  7.  2.  2.  2.  2.  2.]
 [ 5.  5.  5.  5.  5.  6.  6.  6.  6.  6.  7.  7.  7.  7.  7.  2.  2.  2.  2.  2.]
 [ 5.  5.  5.  5.  5.  6.  6.  6.  6.  6.  7.  7.  7.  7.  7.  2.  2.  2.  2.  2.]
 [ 5.  5.  5.  5.  5.  6.  6.  6.  6.  6.  7.  7.  7.  7.  7.  2.  2.  2.  2.  2.]
 [ 5.  5.  5.  5.  5.  6.  6.  6.  6.  6.  7.  7.  7.  7.  7.  2.  2.  2.  2.  2.]
 [ 5.  5.  5.  5.  5. 10. 10. 10. 10. 10.  6.  6.  6.  6.  6. 12. 12. 12. 12. 12.]
 [ 5.  5.  5.  5.  5. 10. 10. 10. 10. 10.  6.  6.  6.  6.  6. 12. 12. 12. 12. 12.]
 [ 5.  5.  5.  5.  5. 10. 10. 10. 10. 10.  6.  6.  6.  6.  6. 12. 12. 12. 12. 12.]
 [ 5.  5.  5.  5.  5. 10. 10. 10. 10. 10.  6.  6.  6.  6.  6. 12. 12. 12. 12. 12.]
 [ 5.  5.  5.  5.  5. 10. 10. 10. 10. 10.  6.  6.  6.  6.  6. 12. 12. 12. 12. 12.]
 [13. 13. 13. 13. 13. 14. 14. 14. 14. 14. 10. 10. 10. 10. 10. 13. 13. 13. 13. 13.]
 [13. 13. 13. 13. 13. 14. 14. 14. 14. 14. 10. 10. 10. 10. 10. 13. 13. 13. 13. 13.]
 [13. 13. 13. 13. 13. 14. 14. 14. 14. 14. 10. 10. 10. 10. 10. 13. 13. 13. 13. 13.]
 [13. 13. 13. 13. 13. 14. 14. 14. 14. 14. 10. 10. 10. 10. 10. 13. 13. 13. 13. 13.]
 [13. 13. 13. 13. 13. 14. 14. 14. 14. 14. 10. 10. 10. 10. 10. 13. 13. 13. 13. 13.]]
```

Figure 3: Example configuration, colors indicate grouped sections; to generate a context: sample two numbers from the matrix, then sample one object for each of the two numbers.

Objective function

Since we are aiming to explore what regularities of the semantic space result in efficient category systems rather than just if such systems can naturally emerge, we want to directly control for vocabulary size so as to cover a wider array of possible categorization systems. One option would be to directly limit the maximum number of categories that agents are permitted to use, and run the model with different maximum values. However, this would require retraining the agents from scratch for each maximum vocabulary size. We instead opted for adding an additional term to the regular task-specific utility, term which acts as a penalty on more complex systems and is based on the objective function proposed by Tucker et al.. This measure is grounded in rate-distortion theory, and is equal to the KL divergence between the softmax distribution from which the one-hot vector was sampled using GS and a uniform categorical distribution. While this measure does not exclusively capture vocabulary size, it has one major advantage: it can be computed with very low computational effort using the model parameters, which is crucial for integration into an objective function. At the same time, the additional penalization of other aspects of complexity other than vocabulary size is not a concern in our case, as this penalty will be applied identically in both conditions, and exclusively in the generation process. We thus make sure that it cannot have any impact on the complexity comparison between our two main conditions. To sum up, the sender S and receiver R are trained to jointly maximize the objective:

$$\mathcal{L}(S,R) = \lambda_U \, \mathcal{E}_{S,R}[U(x,y)] - \lambda_C C_S \tag{1}$$

where $U(x,y)$ represents the utility associated with sender input x and receiver action y, C_S represents the complexity of the system, and λ_U and λ_C control the relative weights of utility and complexity, respectively.

We set $\lambda_U = 1$ and vary λ_C using deterministic annealing. We first trained agents for 80 epochs with a negligible value for λ_C (set to 0.05 rather than 0 to prevent numerical instability) so as to obtain communicatively-optimal systems before beginning to anneal the penalty term for the next 40 epochs, increasing λ_C by 0.1 during each epoch. This will result in systems with increasingly smaller vocabularies that nonetheless maintain communicative optimality, which, as mentioned earlier, allows us to compare convexity between systems emerging from the two context conditions at multiple vocabulary sizes. We chose not to anneal for more than 40 epochs as we found that systems suffer a significant drop in success rates around this point, as they become too simple to permit encoding all communicatively relevant distinctions.

Convexity measures

While convexity understood as the property of category systems that form convex regions in meaning space is binary, we are also interested in assessing to what extent a system leans towards convexity. Thus, by degree of convexity of a system we will refer to the compactness and tightness of the clusters formed in meaning space by the system's categories. In order to quantify the degree of convexity in the agents' category systems we first need to estimate the probability distribution over signals $P(s|o)$ associated with each object. To capture the potential uncertainty of the sender, we sample 20 signals s with replacement for each object o from the sender network after convergence, each time using a different context. To get a representation that is not context dependent we would need to use maximally sized contexts in this process. However, due to technical limitations, we instead use contexts of size 40 with the target being the object for which we are estimating the distribution and the 39 distractors being randomly generated again for each of our 20 samples. Thus, if the sender has no uncertainty about the signal associated with object o, the same signal will be obtained from all 20 samples. Finally we associate a single most sampled signal for each object to obtain our estimation of the sender's category system.

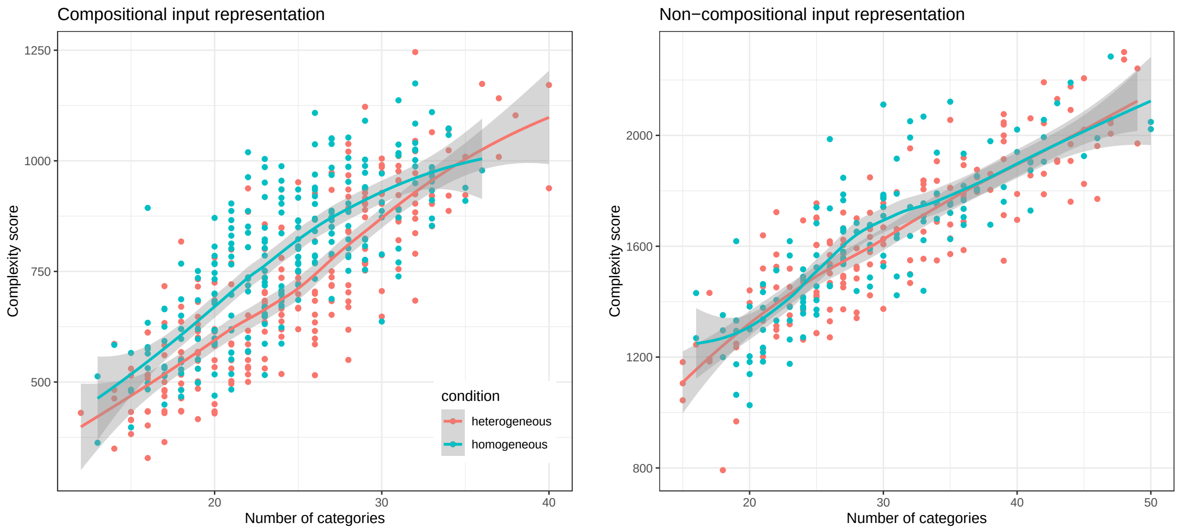
To compare the systems emerging from the homogeneous and heterogeneous conditions, we use a complexity measures from Fass & Feldman (2002) as adapted by Carr et al. (2020). In contrast to the vocabulary size measure used in the objective function, we do not care as much about the computational cost of the measure, but are critically interested that it incorporates some cognitively-grounded assumptions about the properties that make a semantic category system simple, specifically in our case, convexity. Consequently, we define complexity as the minimum description length in bits, with descriptions being constructed using a predefined set of rectangular sections that together fully partition the similarity space in various ways. To obtain the complexity of a system, the minimum description lengths of all contained categories are summed. The minimal description of a category is given by the minimal set of rectangular sections that losslessly encapsulate the region covered by that category in similarity space. While this measure can capture whether categories tend more or less towards convexity, it is also influenced by the number of categories in the system, with systems featuring more categories having higher complexity. We will thus use this measure exclusively to compare systems of the same number of categories to eliminate this factor, effectively transforming the complexity measure into one of convexity.

We predict that for a sender interacting with a homogeneous set of partners, given that the same meaning distinctions are communicatively relevant for all partners, the emerging systems will score lower on convexity, and this effect will be stronger for systems resulting from later epochs (i.e., with a higher pressure for a low number of categories). This is because the agents will be able to leverage the invariable set of contexts to come up with a communicatively optimal system that also uses fewer categories. Conversely, with a heterogeneous set of partners, as the contexts are more variable across partners, the best solution will be a categorization system that leverages on the structure of the semantic space.

Results

Heterogeneous vs. homogeneous context sets

We ran 20 simulations for each condition, each simulation with a different set of contexts generated by a different random seed, and obtained the category systems that are used in our analysis as follows: once we start to anneal the complexity term after the 80 burn-in epochs, we sample a system every two epochs over the next 40 using the method described in the **Model** section, resulting in 20 systems per simulation.



(a) (b)

Figure 4: Each point represents one category system emerging in the heterogeneous (red) and homogeneous (blue) conditions on the complexity/vocabulary-size plane, for the compositional (a) and non-compositional (b) input representations. Lines are calculated using Loess regression while standard error bars are computed using a t-based approximation. Observe that in (a), for any given number of categories below 34, blue points tend to be situated above red points.

Table 1: Success rates for each context condition and both types of input representation methods.

Communication success rates				
Representation	Homo avg.	Hetero avg.	Homo best	Hetero best
Comp.	98.04%	95.22%	99.48%	97.53%
Non-comp.	95.12%	92.86%	97.31%	95.88%

In the first row of Table 1 we report the average accuracy on the test set for the 20 systems per simulation obtained after the burn-in period, as well as the average accuracy for the systems generated in the last burn-in epoch of each simulation (i.e., with no vocabulary size penalty). For both sets of figures, the systems emerging in the homogeneous condition achieve higher success compared to the ones in the heterogeneous condition. This can be explained by the relatively smaller number of meaning distinctions that need to be made across partners in the homogeneous condition, and by the larger consistency of such distinctions across partners.

As our complexity measure is influenced by the number of categories in a system, we assess convexity by directly comparing systems of equal number of categories. Figure 4a shows that systems in the heterogeneous condition of equal number of categories to those in the homogeneous condition have noticeably lower levels of complexity; a visual inspection of the category systems confirms that there are more non-convex and non-compact categories in the heterogeneous condition. We can explain this result by considering that due to our randomized context generation method a system that leverages the common ground (i.e., the distinctions that are known to be needed in communication) between partners has to go against the structure observed in the input space, as it would require meanings from adjacent sections to be part of separate categories and meanings from far-away sections to be part of the same category. The sender in the homogeneous condition is aided by a coherent common ground across partners that it can leverage, as object-category associations are being reinforced over interactions with successive partners. However, the sender in a heterogeneous condition encountering receivers with highly varying contexts is required to continuously adapt its categories to support communication of different sets of meaning distinctions. This sender would thus have to overcome the additional hurdle of an incoherent common ground across partners. Instead, it will favor leveraging on the input space structure, which is shared by all agents, leading to categories that form tight clusters in similarity space. We also notice two distinct patterns, with differences in complexity between the two conditions being evident for systems with fewer categories, but less apparent in systems with the largest number of categories. This is in line with our predictions, as there is little incentive to try to overcome the biases originating from the input space structure and form less convex categories if larger systems that would adhere to these biases are not additionally penalized.

Effects of the type of input representation

We compare the systems that emerge using our compositional input representation method with the systems emerging using the non-compositional representation method used by Chaabouni et al.. To encode a meaning composed of mul-

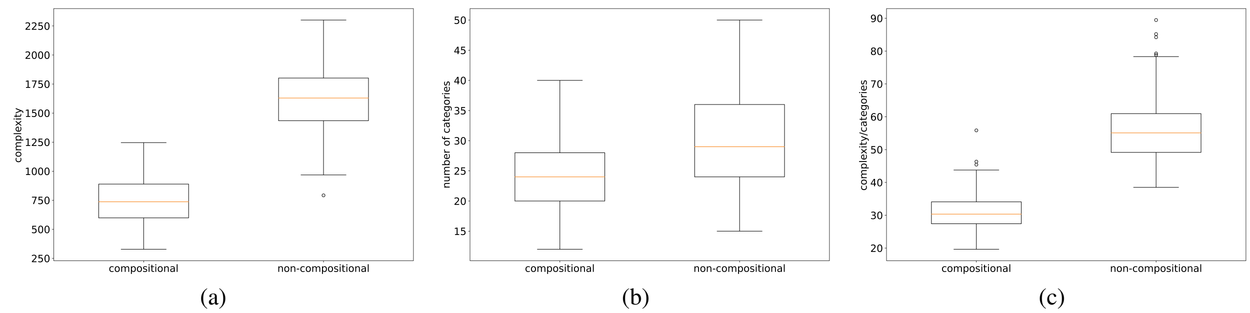


Figure 5: Comparison of the category systems emerging from the compositional and non-compositional input representation methods in terms of complexity (a), number of categories (b), and complexity divided by number of categories (c).

tiple semantic features using this non-compositional method, each feature value is simply encoded as a single numerical value and the values that make up the meaning are placed in a vector (e.g., in our case, a 2D vector). We first show in Figure 3 and Table 1 (rows 1 and 2) that compositional representations result in systems that are less complex, have a smaller number of categories, and achieve higher success than non-compositional representations. The average complexity of a category is also lower, which would suggest that the systems emerging with the compositional representation are more aligned with properties of the semantic space that we look for with our complexity measure (i.e., convexity). A visual inspection of the outputs of both models suggests that systems emerging with compositional embeddings do tend to be much more rectangular and less spread out, being better at capturing the square-like semantic configurations used in our context sets. We infer that the differences are a consequence of the composition of the two features being made explicit with our representation. Second, if we tease apart the category systems emerging in the two context conditions, we also observe (Figure 4, a vs. b) that the compositional method results in a much larger difference in complexity between these conditions for systems with the same number of categories. We suspect this is due to a higher number of non-convex categories in the systems resulting from our method. Specifically, the dimensionality of our embeddings is much higher than that of the vectors used in the non-compositional representation (i.e., 400 vs. 2 elements), which results in more parameters, allowing agents to encode more complex patterns. We leave the testing of this hypothesis to future work.

Bayesian vs. NN category systems

We can conclude that the results obtained by Nedelcu et al. using Bayesian agents also hold for NN agents playing the same type of communication game, with more convexity emerging in the heterogeneous condition compared to the homogeneous condition. Using neural networks allowed us to explore a much larger vocabulary (400 words for 400 meanings vs. 3 words for 16 meanings), which means that the agents are not required to make any generalizations in order to solve the task accurately, as they could in principle use one word for each meaning (although the number of categories in the emerging systems is always much smaller than this even when no penalty for vocabulary size is applied; this result was also obtained by Chaabouni et al. who explained it by the discreteness of the signal channel). Additionally, given that the bias of adhering to the semantic similarities of the input space in strong enough, the smaller the penalty for large vocabulary sizes, the less the incentive to form non-convex categories. As such, we would expect that, without added constraints, our NN agents are less inclined to form non-convex categories compared to the Bayesian agents. While we have no way of comparing convexity between the two types of architectures, this effect seems to be confirmed by our simulations, as smaller penalties do result in less non-convexity. Second, we use a much more extensive set of contexts (4 vs. over 50 000 contexts per agent), which is also necessary given that NN agents can potentially memorize the contexts if diverse enough learning examples are not provided. However, this also means that the incentive is even higher to stick to the input representation because agents might require even more evidence to learn how to leverage on a very large set of contexts, even if that set is consistent across partners.

Conclusion

Artificial neural networks trained to solve a simple reference game by optimizing a task-specific utility develop efficient semantic categorization systems that trade off complexity against informativeness, much like the category systems of human languages do. But what exact type of structures in the semantic space could result in efficient categories, and how are these structures shaped by the contexts of communication? We use a population-level NN model to show that agents develop category systems that are more convex when encountering partners with highly varying contexts than when encountering partners with invariable contexts. We also propose a method of input representation based on compositional vector embeddings, which, to the best of our knowledge, has not been previously applied in emergent communication. We show that this method results in a higher level of communication success than regular non-compositional methods, and can also achieve a better balance between maintaining the structure of the semantic space and optimizing utility.

Acknowledgements

This research received funding from a University of Edinburgh College of Arts, Humanities and Social Sciences Research Award (held by V. Nedelcu) and from the European Research Council under the European Union's Horizon 2020 research and innovation program (Grant 681942, held by K. Smith).

References

Carr, J. W., Smith, K., Culbertson, J., & Kirby, S. (2020). Simplicity and informativeness in semantic category systems. *Cognition*, *202*, 104289.

Carstensen, A., Xu, J., Smith, C., & Regier, T. (2015). Language evolution in the lab tends toward informative communication. In *Cogsci.*

Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2021). Communicating artificial neural networks develop efficient color-naming systems. *Proceedings of the National Academy of Sciences*, *118*(12), e2016569118.

Chaabouni, R., Kharitonov, E., Lazaric, A., Dupoux, E., & Baroni, M. (2019). Word-order biases in deep-agent emergent communication. *arXiv preprint arXiv:1905.12330*.

Dautriche, I., Fibla, L., Fievet, A.-C., & Christophe, A. (2018). Learning homophones in context: Easy cases are favored in the lexicon of natural languages. *Cognitive psychology*, *104*, 83–105.

Fass, D., & Feldman, J. (2002). Categorization under complexity: A unified mdl account of human learning of regular and irregular categories. *Advances in neural information processing systems*, *15*.

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

Hawkins, R. D., Franke, M., Frank, M. C., Goldberg, A. E., Smith, K., Griffiths, T. L., & Goodman, N. D. (2022). From partners to populations: A hierarchical bayesian account of coordination and convention. *Psychological Review*.

Kågebäck, M., Carlsson, E., Dubhashi, D., & Sayeed, A. (2020). A reinforcement-learning approach to efficient communication. *Plos one*, *15*(7), e0234894.

Kharitonov, E., Chaabouni, R., Bouchacourt, D., & Baroni, M. (2019). Egg: a toolkit for research on emergence of language in games. *arXiv preprint arXiv:1907.00852*.

L Griffiths, T., Kemp, C., & B Tenenbaum, J. (2008). Bayesian models of cognition.

Mollica, F., & Zaslavsky, N. (under review). Information-theoretic and machine learning methods for semantic categorization. In *The oxford handbook of approaches to language evolution.* Oxford University Press.

Nedelcu, V. C., Smith, K., & Lassiter, D. (2023). Degree of heterogeneity in the contexts of language users mediates the cognitive-communicative trade-off in semantic categorization. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, *7*(4), 573–605.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of mathematical psychology*, *1*(1), 54–87.

Silvey, C., Kirby, S., & Smith, K. (2019). Communication increases category structure and alignment only when combined with cultural transmission. *Journal of Memory and Language*, *109*, 104051.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, *46*(1-2), 159–216.

Tucker, M., Levy, R., Shah, J. A., & Zaslavsky, N. (2022). Trading off utility, informativeness, and complexity in emergent communication. *Advances in neural information processing systems*, *35*, 22214–22228.

Van Rooij, I., Blokpoel, M., Kwisthout, J., & Wareham, T. (2019). *Cognition and intractability: A guide to classical and parameterized complexity analysis*. Cambridge University Press.

Wray, A., & Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, *117*(3), 543–578.