Universiteit Leiden
The Netherlands

Opleiding Informatica

Structural systematicity under varying meaning spaces

Sayfeddine el Kaddouri

Supervisors:
Tessa Verhoef & Tom Kouwenhoven

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

14/08/2023

Abstract

In this study, the causal relationship between meaning space and various language metrics including topsim, bosdis, posdis, and mutual information was examined. The findings indicate that the pressure to generalize to new inputs may not necessarily drive the development of a compositional language. While meaning space does seem to influence mutual information and generalization to some degree, the relationship between topographic similarity and mutual information wasn't conclusively positive. This leads to a hypothesis that agents might develop holistic languages, which differ fundamentally from compositional languages. Theresults align with prior research, suggesting that group size, rather than meaning space, influences compositionality due to increased input variability. While topsim and posdis displayed correlated behaviors, bosdis deviated, highlighting its unique property. The insights shed light on the nuances of language emergence and the complexities surrounding metrics like bosdis, posdis, and topsim. This work sets a foundation for future exploration into how disentanglement metrics can better measure language properties.

Contents

1 Introduction

The field of language emergence studies and describes how language emerged and came into existence. Both linguistics [Cro08, MM12] and artificial intelligence researchers [MCH+20] are concerned with studying the field of language emergence. On the one hand, artificially reproducing language emergence may help to understand the evolution of human languages. On the other hand, language is known to be structured and compositional [Bic07], and imitating such properties would enhance machine learning representations. Namely, natural language requires constructing hierarchical structures, and artificial neural networks are known to be able to capture hierarchical structures and features [MCH+20]. This can be shown in convolutional neural networks, where features are extracted hierarchically, similarly to the human brain[AAS20]. Research in language emergence could lead to architectures that could capture such properties in language.

By means of language games, experiments are done in experiments with human participants and computational simulations, such as Lewis' referential game [Ste12]. Participants or agents are instructed to complete a task in a language game, where language is necessary to succeed. However, subjects in language games are not allowed to use existing languages and are sometimes limited in the symbols that they can use. Emergent communication among humans and artificial agents do not share the same properties yet and also seem to show critical mismatches in the literature [GRR22]. Experimental data [LB20, RMLA19a, GRR22] shows that in contrast to natural language, which is characterized by complex grammatical structures and syntactic rules, artificial neural networks used in agent models may produce a language that does not always adhere to the same principles governing human languages. Specifically, emergent languages from artificial agents do not show the same structure or patterns as in human experiments.

When expressing complex concepts there is need for compositionality. Natural languages supply this requirement, as they typically provide the means to craft messages that refer to complex concepts. This is achieved through the systematic combination of expressions. Such a structure does not only supports the communication of multifaceted ideas but also underscores the rich interplay of semantics and syntax within human language [LHTC18, LB19]. Some researchers in language emergence seems to agree that compositionality is a required property for a language to handle unseen inputs [CLL+19, RMLA19a, Mar03, FL02]. However, emergent language researchers are also interested in how information is entangled in messages [CKB+20]. For instance, in the sentence "She paints the wall", "paints" is a verb or function, and "She" and "the wall" are arguments. The compositionality here involves more than just putting words together; it's about the relationship and interaction between the words. As such, in addition to compositionality it could be valuable in language emergence research to capture which properties certain symbols have in relation to attributes of objects.

Current research in language emergence focuses on investigating language and population properties, particularly their influence on generalization and language acquisition [GRR23]. Such properties are population size and input variability, which have been investigated by Raviv et al. [RMLA19b, RMLA19a] in experiments with human participants. The presented study aims to investigate the role of growing input variability in the acquisition of language in artificial agents.

Research by Raviv et al. [RMLA19b] has shown that when populations are larger they tend to have more systematic languages. This is due to the fact that languages in larger communities are under pressure for compression, such that they are more grammatical and easier to memorize in a large population. Raviv et al. also reasons that larger communities tend to handle larger input variability better. In other words, larger population size and a larger meaning space seem to play a role in the systematicity of a language. In this work, we empirically show, by means of computational experiments, that a growing meaning size indeed contributes to the systematicity of language. We show this by analysing the development of compositional structure in the evolving communication systems and reporting the results and the communication success rate of the language games.

2 Related Work

2.1 Compressibility Pressures

Galke et al. [GRR22] present three important phenomena that relate to the emergence of language, which have been confirmed experimentally in human studies:

1. Ease of learning compositional languages, how easy it is for a newly initialized agent to learn the emerged language.

2. Compositionality aids generalization and convergence of emergent languages

3. Larger population generally develops more structured languages

The presented paper will focus on one of these aspects, how compositionality aids generalization of emergent languages.

Socio-demographic factors have long been assumed to play an important factor in language evolution [LD10]. One of these socio-demographic factors is population size. Global cross-linguistic studies have found that larger populations tend to have languages with more systematic structures [LD10]. This effect has also been confirmed experimentally, where groups with more interacting participants created more systematic languages [RMLA19b]. The observed results can be explained due to the presence of compression pressures during the interaction. As group size grows, recalling partner-specific variations poses greater difficulty due to memory constraints. Consequently, bigger groups tend towards more easily interpretable languages [GRR22].

In earlier work, Raviv et al. [RMLA19b] conducted research to test the effect of group size and input variability on systematicity in a language. Two groups, one group consisting of four people and one group consisting of eight people, participated in a signaling game. Each group was shown several scenes, which they had to describe to each other using non-existent symbols (words). Following the descriptions, participants needed to select the correct scenes according to some descriptions provided by their group member. In the study Raviv et al. showed that growing group size and input variability facilitates systemacity in a language. In addition, they showed that a close relationship exists between linguistic structure and comprehensibility of a language.

Similarly, Raviv et al. [RMLA19a] conducted a meta-analysis, where they studied the effect of a growing meaning space and growing group size on compositionality. The results confirmed the findings in the previous study [RMLA19b], where it was shown that a growing group size facilitates compositionality. However, a significant effect from a growing meaning space on compositionality was not shown. Therefore, it seems growing group size is a stronger pressure for compositionality compared to a growing meaning space. However, both factors were studied simultaneously in one experimental set-up, and whether growing meaning spaces can boost the emergence of compositional structure in absence of a growing group size remains unknown. Finally, while other researchers [SKO04, AMPS08] have found that new learners are necessary for the formation of linguistic structure in a language, the meta-analysis [RMLA19a] showed this is not necessary. Therefore, it is justified to perform experiments using agent pairs, instead of larger groups.

2.2 Language Properties

As there is a need to capture properties of emergent languages, several metrics have been proposed in order to capture such metrics. Topographic similarity [BK06] is often used in language emergence studies as a way to quantify compostionality [CKB+20, LHTC18]. The metric can detect the tendency for messages with similar meanings to be in similar form. However Chaabouni et al. [CKB+20] have pointed out there is no correlation between compositionality and the generalizability of new inputs for in language. Nonetheless, generalization of new inputs is seen as one of the core attributes of compositionality in natural languages [PW10]. Even though the results by Chaabouni et al. [CKB+20] show that compositionality is not necessary for a language to be generalizable, the general consensus is that compositional languages do support generalization [PW10, KCS08, GRR22]. This suggests, that the need to generalize is not sufficient for a language to be compositional. In conclusion, it can be said that compositionality makes it very likely for a language to be generalizable and thus is highly desired for language.

Several other metrics have been proposed in order to analyse what establishes compositionality in emergent languages [CKB+20, XNR22]. Taking inspiration from the representation learning literature [XNR22], Chaabouni et al. [CKB+20] propose that looking at disentanglement in learned representations could be of use. Messages in natural languages entangle information about concepts. Thus, studying how concepts are encoded in a message by disentangling the messages could give more insight in how different properties could arise. Consequently, using this inspiration Chaabouni et al. [CKB+20] introduced two new measures to capture disentanglement in a language. Positional disentanglement (posdis) and bag of Symbols disentanglement (bosdis). While, they were unable to show that disentanglement is strongly correlated with topographic similarity, they did observe that similar behaviour between the two. This is not surprising as topographic similarity is an agnostic measure and does not fully capture compositionality [CKB+20]. Given this understanding, it is sensible to use posdis and bosdis for analysis in emergent languagues.

2.3 Computational Modelling

Computational modeling has long been used to study language evolution [Ste12, KO21, KCS08]. Recent research has focused on the application of advanced machine learning methods such as deep neural networks and reinforcement learning algorithms to study emerging communication

patterns in multi-agent systems [LB20]. This approach enables analysis of the dynamics at play when multiple agents interact through learned signals, providing new insights into how populations might coordinate their behavior under different conditions [GRR22].

In most computational experiments either symbolic or visual inputs have been used to study language emergence [LHTC18]. Most studies employ either some type of Recurrent Neural Networks (RNN) [RHW85], such as LSTMs or GRUs [AS17], or Autoencoder models as architectures for multi-agent systems. Training is usually carried out utilizing REINFORCE or Gumbel-Softmax optimization for efficient agent coordination [GRR22].

In recent studies, the majority of investigations of computational language emergence have relied on Lewis' reference game as a means of evaluating language emergence [GRR22]. In the context of a Lewis game, the first participant (the speaker) perceives a randomly selected object within its environment. The aim of the speaker is to describe this object to the second participant (the listener), which needs to reconstruct the description in order to find the object. The success of this game is quantified by how well the pair is able to reconstruct the original object [Ste12]. However, variants of the Lewis game have also been presented. One such is the variant proposed by [KO21]. Compared to referential games wherein full information about the observations is available to all participants, the modified version features partial and disjoint knowledge possessed by the agents. As such, the focus of communication shifts to conveying the limited data obtained through respective incomplete perceptions.

Agent architecture, game design, input, and training protocol are the building pillars of a computational language emergence experiment [GRR22, GRR23]. However, the experiments also need to be analyzed. For this reason, researchers have used measures to analyze language emergence experiments to capture linguistic properties or generalization [CKB+20].

3 Method

3.1 Game Design

For the game design an existing framework, EGG, from Meta AI [KCBB19] has been used. The framework contains several language games which can be used to simulate language emergence among neural agents. Several other research papers have used this framework as well to study language emergence. [CKB+20, KCBB20] making it relevant to work in this framework.

The language game of choice for this research is the Object's game [LHTC18], which is a type of referential game similar to the Lewis' game [Ste12]. The Object's game consists of a speaker and listener which are able to *perceive* an observation. The speaker selects a random object from a set of objects from the observation and sends out a *message* to the listener, which in turn tried to find the selected object.

In each trial in the game, two agents perceive several discrete objects: A target o_T, a set of distractors $\mathcal{O}_\mathcal{D}$, and a message $i \in \mathcal{I}$. The task of the game is to choose the correct target object according to

some message M.

Agent Each agent pair is initialized newly at the start of a new game. An agent pair is denoted by a tuple $\mathcal{G} = (\mathcal{G}^s, \mathcal{G}^l)$. Where $\mathcal{G}^s$ and $\mathcal{G}^l$ represent the listener and speaker agent respectfully.

Actions Consider an environment where the action space is divided into a *message action space* $\mathcal{A}^m$ and a *decision action space* $\mathcal{A}^d$, such that $\mathcal{A}^d \cup \mathcal{A}^m = \mathcal{A}$ and $\mathcal{A}^d \cap \mathcal{A}^m = \emptyset$.
A speaker and listener pair $\mathcal{G} = (\mathcal{G}^s, \mathcal{G}^l)$ is initialized at the beginning of a round. The speaker $\mathcal{G}^s$ takes observations as inputs and sends a discrete message $m \in \mathcal{A}^m$ to the the listener $\mathcal{G}^l$ as output. The listener $\mathcal{G}^l$ observes the state of the environment and receives a message m from the speaker. According to m, the listener acts and decides what the correct object is with the action $d \in \mathcal{A}^d$.

The communication is successful between the pair $\mathcal{G}$ if d maps to the target T. Finally, both the speaker and the listener in a pair have access to the full state information of the objects in the environment.

Objects We consider an object $o \in \mathcal{O}$ as a discrete vector of length N and where the entries are $n_i \in N$, where $i \in M$. Here, N denotes the attributes of an object and M denotes the values the attributes can take. Consequently, $M^N = |\mathcal{O}|$ objects can be constructed.

State and observation Consider a state-tuple s, such that $s = (o_T, \mathcal{O}_D)$, where $\{o_T, \mathcal{O}_D\} \in \mathcal{O}$. o_T denotes the target object and $\mathcal{O}_D$ denotes the set of distractor objects. Given the state s the speaker generates an appropriate message $m \in \mathcal{A}^m$ to forward to the listener. Both agents have access to the full state information.

communication success rate To evaluate whether a language $\mathcal{L}$ has converged and **communication success rate** is used. Communication success rate is defined as the success rate of the listener selecting the the target object o_T.

3.2 Metrics

As there is a need to measure language properties of a language $\mathcal{L}$, inspiration is taken from earlier research that proposes the use of three metrics [CKB+20]. The metrics include *posdis*, *topsim* and *bosdis*

Topographic similarity (Topsim) Compositionality is considered crucial for generalization in a language [Mar03, FL02]. Therefore, Topsim [BK06] is used as a measure to capture compositionality. Topsim tests whether objects that are close to each other are also closely mapped in the message space. This is achieved by computing the Spearman correlation between the pairwise distances between object space and message space. We formally define Topsim by 1:

$$Topsim = \rho = 1 - \frac{6 \sum_{i=1}^{n} (r_{oi} - r_{mi})^2}{n(n^2 - 1)} \tag{1}$$

We denote, $n = |\mathcal{O}|$ is the total number of objects being compared. *oi* and *mi* are the set of distances between distances and objects respectfully. In addition, r_{oi} denotes the rank of the pairwise distance of the i-th object pair in the object space. Similarly, r_{mi} denots the rank of the pairwise distance of the i-th message pair in the message space. A higher value of ρ indicates a higher topographic similarity, as it means the ranks of the distances in the two spaces are more closely correlated.

Positional disentanglement (Posdis) Posdis [CKB+20] measures whether symbols in specific positions tend to univocally refer to the values of a specific attribute. This attribute is often encountered in natural language structures. For instance, take *yellow square*. The first symbol *yellow* refers to square in this example and is used as an adjective. Posdis tries to find structures like this in messages. Let's denote $s_j : a_1^j$ as the j-th symbol of a message and as the attribute with the highest mutual information. Likewise, a_2^j is the attribute with the second highest mutual information. Finally, $\mathcal{H}(s_j)$ is the entropy of the j-th symbol in a message m, which is used as normalizing term. Posdis, is in turn defined as 2.

$$Posdis = 1/c_{\text{len}} \sum_{j=1}^{c_{\text{len}}} \frac{\mathcal{I}\left(s_j; a_1^j\right) - \mathcal{I}\left(s_j; a_2^j\right)}{\mathcal{H}\left(s_j\right)} \tag{2}$$

Bag of symbols disentanglement (Bosdis) *Posdis* assumes that every symbol in a specific message refers to something else. However, we can imagine languages where that is not the case. For example, in English order of symbols does not always change the semantic meaning of a sentence, i.e. *bees and birds* and *birds and bees* are semantically the same. *Bosdis*[CKB+20] captures the tendency of symbols to refer to distinct meanings. Let n_j be the count of the j-th symbol in a message. Finally, *bosdis* is defined as 3.

$$Bosdis = 1/c_{\text{voc}} \sum_{j=1}^{c_{\text{voc}}} \frac{\mathcal{I}\left(n_j; a_1^j\right) - \mathcal{I}\left(n_j; a_2^j\right)}{\mathcal{H}\left(n_j\right)} \tag{3}$$

Generalization Generalization captures how well agents are able to generalize unseen objects. This is computed by taking the average test accuracy of various communication channels between an agent-pair (c, s)

3.3 Training and architecture

Speaker The training consists of $n\_episodes \in \mathbb{N}$. During each episode, one agent is initialized as speaker. Next, the speaker processes several discretely valued object-vectors $o \in \mathcal{O}$ into an embedding using either an vanilla-RNN, LSTM or GRU, which in turn outputs a message $m \in \mathcal{A}$.

Listener Similarly to the speaker in 3.3 the listener processes the observations in the same manner. However, instead of outputting a message m the listener processes the message m forwards it to the hidden layers and outputs an action $d \in \mathcal{A}^d$ with the highest probability.

Architecture Both agents are implemented with two GRUs. The sender processes the target object through a linear transformation and feeds it forward into the GRU. Using Gumbel-Softmax, it then generates a message m that is passed to the receiver. The receiver processes message m through a GRU, which in turn creates an latent embedding of the target message m. Additionally, the listener processes the target vector o_T and a set of distractor vectors $\mathcal{O}_\mathcal{D}$. For each object vector in the environment and the message embedding a pair is constructed for which a similarity score is calculated using the dot product. Following, the similarity score is used by the listener to select an object.

Training During training, messages are generated by the speaker and propagated to the listener through a communication channel. Using a message m and a decision a^d is selected. The loss is calculated with cross-entropy loss over the target objects and the selected object as this this helps with measuring the distribution of objects in the environment. Gradients are calculated using Gumbel-softmax relaxation [JGP16], which are used for reparametrization of the architectures of the agents. In addition, for loss optimization backpropagation is used and minimized through an Adam optimizer.

3.4 Experiments

In this work we aim to understand whether there is an effect between the increase of meaning space on compositionality of a language. To test this the following experimental setup has been performed.

There is a need to increase the meaning space in the referential game. To achieve this we simply adjust a single parameter in the experiment. Objects in the game are constructed using a discrete vector of length N, such that the entries yield values $n_i \in N$, where $i \in M$ and $N \in \mathbb{N}$. An entry n_i can yield a value ranging from 0 to N. In the proposed experimental setup $N = 4$. This leads to a configuration of the game where M^4 objects can be constructed.

For every experimental run we change M 10 times during a trial. The agents are trained for 30 epochs, after which we freeze and save the weights of the agents. Accuracy and linguistic metrics are outputted to a csv file as well as the messages that are outputted and objects vectors in the environment. Following, new objects are introduced by scaling M and the game is continued with an increase in meaning space. The vocabulary size $l_{voc} = 500$, epoch length $l_{epochs} = 20$, message length $l_m = 6$ and the number of distractors $l_{dist} = 9$ are kept constant. In addition, the hyperparameters are kept constant as well, where the learning rate and batch size for the Adam optimizer is 0.001 and a 32 respectively. Finally, 30 trials are performed to ensure a sufficient amount data for which we an analysis has been performed. Hyperparameters search has not performed as the hyperparameters from the original paper have been used [LB20].

4 Results

Effect of meaning space Averaged accuracy, topsim, bosdis, posdis and mutual information are reported in table 1. For a confidence interval of 95% a linear regression analysis has been performed for which the fitted curves can be found in figure 1. In addition, the slope coefficients, p-values of

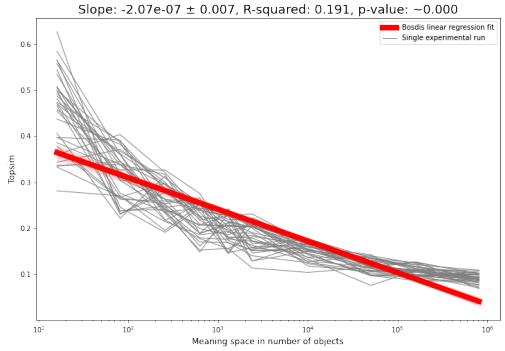
Table 1: Results of running the Object's game with agent pairs. Language metrics and test accuracy are reported when scaling meaning space averaged over 30 runs.

Meaning Space	16	81	256	625	1.296	2.401	10.000	50.625	160.000	810.000
Accuracy	0.75	0.93	0.94	0.94	0.93	0.93	0.93	0.92	0.92	0.92
Topsim	0.46	0.32	0.27	0.22	0.20	0.18	0.15	0.12	0.11	0.09
Bosdis	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05
Posdis	0.19	0.09	0.05	0.04	0.04	0.05	0.03	0.04	0.04	0.04
Mutual information	6.39	6.86	6.94	6.81	6.81	6.87	7.22	6.72	6.94	7.11

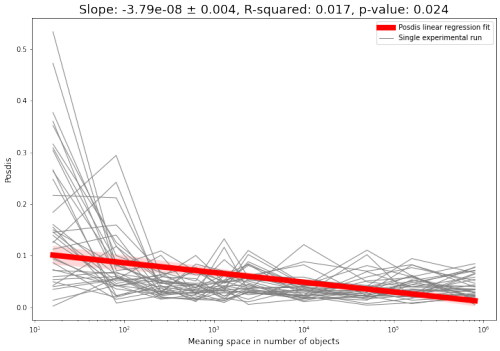
the variables and R^2 are reported. Meaning space fails to explain all the variation in the metrics, as can be shown by the low R^2. While the effect of meaning space on mutual information is minimal, there seems to be a causal relationship between the two.

In figure 2 the Spearman-correlation between the metrics are portrayed in a heatmap. All the reported correlations yield a p-value of $p < 0.05$, except for the correlation between meaning space and accuracy, and accuracy and bosdis, which have a p-value of 0.289 and 0.130 respectfully. Consistent with findings from previous research as highlighted by Chaabouni et al. [CKB+20], there appears to be little correlation between topsim and bosdis with generalization. However, a strong negative correlation can be found between generalization and posdis. In addition, A strong of positive correlation between Mutual information and meaning space can be found. Furthermore, a strong correlation between generalization and mutual information is observed. Finally, the results show that mutual information is strongly negatively correlated with topsim, posdis and bosdis.

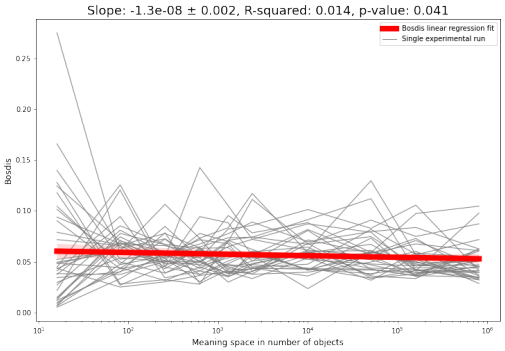
In addition, the results in figure 1 shows that generalization is strongly correlated with a scaling of the meaning space as can be shown by the increase in average accuracy on the test.



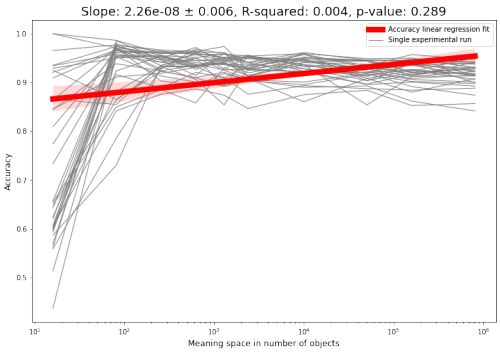
(a) Linear regression model for the predictor meaning space and outcome topograhic similarity (topsim)



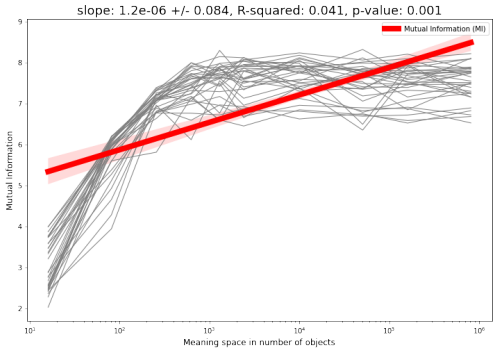
(b) Linear regression model for the predictor meaning space and outcome positional disentanglement (posdis)



(c) Linear regression model for the predictor meaning space and outcome bag of symbols disentanglement (bosdis)



(d) Linear regression model for the predictor meaning space and outcome accuracy



(e) Linear regression model for the predictor meaning space and outcome mutual information

Figure 1: Linear regression models for meaning spaces predicting one of the language language metrics and test accuracy. Reported metrics are averaged over 30 runs in addition a log(10) scale is used to capture the rapid changes in meaning space.

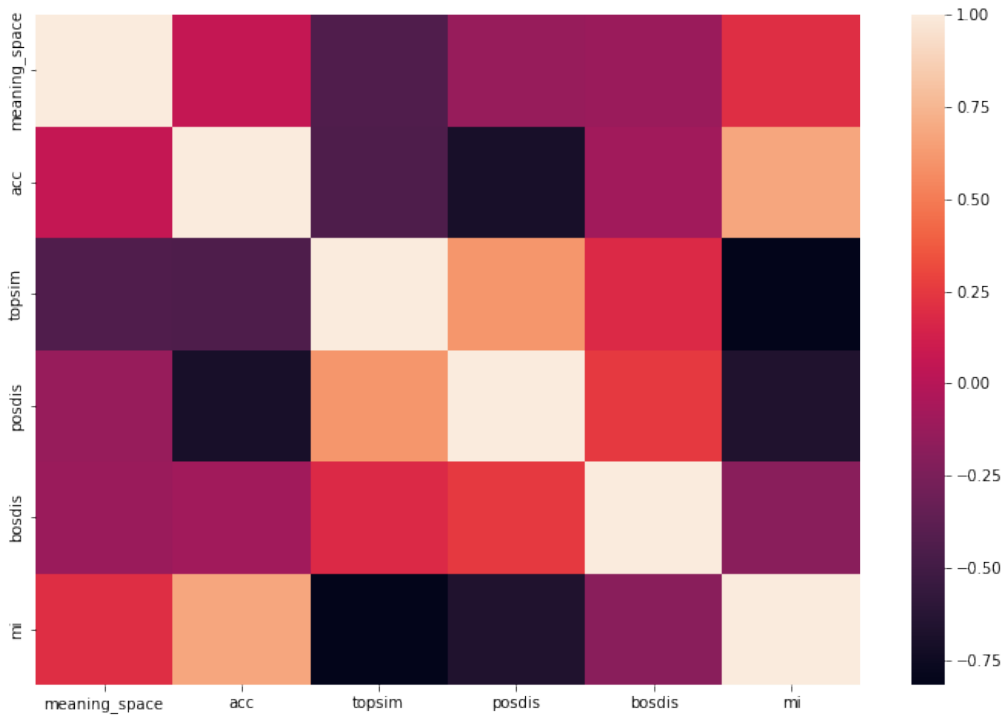


Figure 2: Spearman correlations between the variables *meaning spaces*, *accuracy*, *topographic similarity*, *positional disentanglement*, *bag of symbol disentanglement* and *mutual information*. Reported correlations are statistically significant ($p < 0.05$), with exception of meaning space and accuracy, and accuracy and bosdis.

Role of compositionality measures The analyses failed to show a positive causal effect between meaning space and topsim shown by the regression analysis in figure 1a and the Spearman-correlation analysis in figure 2. This raises the question how other compositionality measures behave in regard to topsim. The Spearman-correlation analysis has already shown that there appears to be a correlation between *posdis* and *topsim*, and between *bosdis* and *topsim*. Whether the portrayed relationships also account for causality remains unclear. For this reason, an additional multiple regression analysis has been performed to assess the importance of the variables. The results are presented in figure 3, for which we obtain the following statistics in table 2.

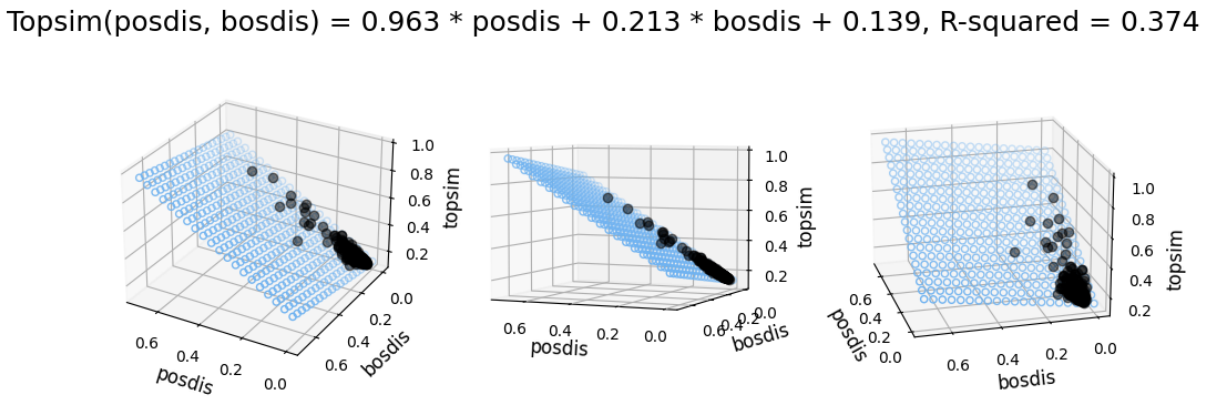


Figure 3: Multiple linear regression model with posdis and bosdis as predictors for topsim. Posdis shows a highly predictive relationship with topsim, whereas bosdis does not show a highly predictive relationship with topsim.

Table 2: Parameters for the multiple linear regression model using predictor variables posdis and bosdis, with topsim as the outcome variable.

β posdis	β bosdis	p-value posdis	p-value bosdis	R^2	confidence level
0.963	0.213	~0.000	0.322	0.374	95%

The presented regression model shows a significant positive relationship between *posdis* and *topsim* ($\beta \approx 0.963$, $p \approx 0.000$). On the other hand, the regression model was unable to find a significant relationship between *bosdis* and *topsim* ($\beta \approx 0.213$, $p \approx 0.322$). Finally, the variables of the regression model do not account for all the variation in the data as can be shown by squared R value.

5 Conclusions and Further Research

The results from the experimental setup failed to show a strong causal relationship between meaning space and the language metrics, topsim, bosdis, posdis and mutual information. This suggests that the need to generalize to new inputs does not appear to be a sufficient pressure to to develop a compositional language. However, while the effect is not strong, meaning space does appear to

influence generalization and mutual information slightly. An increase in mutual information can be explained by two things. First, the rise in mutual information could be attributed to increased compositionality in the emergent language. Second, it might be the result of agents developing holistic languages at some point in the experimental setup. The latter explanation seems more plausible as the results show that topographic similarity and mutual information are not strongly positively correlated. In many languages, high mutual information might suggest that words are related in a compositional manner (e.g., "blue" might frequently occur with "sky"). However, in a holistic language, this is not the case. Instead, words with high mutual information might frequently appear in the same situations or contexts, even if they do not share semantic components.

The results in this work are parallel to the work done by Raviv et al. [RMLA19a, RMLA19b]. The researchers were not able to observe a causal effect between meaning space and compositionality, whilst there seemed to be an effect between group size and compositionality. This suggests that increasing group size is a sufficient compressibility pressure to increase compositionality. The effect could be explained due to an increase in input variability. When populations increase in size different styles or dialects can emerge [CT17]. In turn, language learners get exposed to this variability and are under pressure to develop a language that is robust. Moreover, increasing meaning space had not been individually studied before. Here we find no positive relationship between scaling meaning space and its effect on topsim.

In addition, bosdis did not seem to have a significant relationship with topsim, whereas, a significant relationship between posdis and topsim was observed. Even though all three measures capture compositionality in a different way bosdis seemed to capture an exclusive property. The difference is not surprising due to the definition of of the measures. Posdis captures the tendencies for symbols in specific positions to refer to another symbol, while topsim captures the tendency for semantically similar messages to be in similar form. The type of compositionality bosdis captures is different given by its definition 2. The nature of bosdis is rooted in the concept of permutation invariance; it is primarily concerned with the mere presence or absence of symbols, rather than their sequential arrangement. On the other hand, posdis and topsim are sensitive to the arrangement of symbols. This difference underlies the observed deviation in behavior among the three metrics.

There are a number of shortcomings in the study's design and execution. Firstly, the research was constrained by the limited size of the population tested. A larger population size could increase the internal validity of the study. Furthermore, the reporting methodology was not as detailed as wanted. Due to the usage of the EGG framework, presenting the measures for each epoch was not achieved. Only the test measures after 50 training epochs were presented, which might not capture the full dynamism of the system under study. Lastly, an opportunity was missed in not utilizing unsupervised learning methods. By clustering the latent representations or messages of the agents, we could have delved deeper into the changes and patterns that might emerge when scaling the meaning space. This kind of analysis has been done previously by [KO21]. These oversights, while highlighting areas for improvement, also underscore the potential avenues for future research.

In conclusion, an increase in meaning space is not sufficient for increased compositionality. Additionally, languages with high communication success rate are not necessarily highly compositional. The experimental setup showed that high communication success rate could also be caused by

holistic languages. Finally, we find parallels between the disentanglement metrics bosdis and posdis, and topsim. However, bosdis seems to behave differently from posdis and topsim. While this work was not able to replicate the results of Raviv et al. [RMLA19b], it provides an opportunity to redirect our attention to study different compressibility pressures in language emergence.

Finally, the disentanglement metrics are a promising new direction of measuring language properties such as compositionality. In future work, it would be valuable how we can ensure disentanglement occurs and how these metrics would behave in different experimental setups. Another promising direction is to use architectures that are able to capture disentanglement. Inspiration can be taken from the various ideas that have been proposed in recent years in the representation learning literature. For example, architectures have been introduced that are able to capture interpretable factorised latent representations[HMP+16, XNR22] in visual data.

6 Acknowledgments

I would like to express my profound gratitude to my supervisors, Tessa Verhoef and Tom Kouwenhoven. Their support, invaluable insights, and mentorship have been pivotal to this research. I am deeply appreciative of the time and effort they invested in guidance for this work.

References

[AAS20] Arohan Ajit, Koustav Acharya, and Abhishek Samanta. A review of convolutional neural networks. In *2020 international conference on emerging trends in information technology and engineering (ic-ETITE)*, pages 1–5. IEEE, 2020.

[AMPS08] Mark Aronoff, Irit Meir, Carol A Padden, and Wendy Sandler. The roots of linguistic organization in a new language. *Interaction studies*, 9(1):133–153, 2008.

[AS17] Ben Athiwaratkun and Jack W Stokes. Malware classification with lstm and gru language models and a character-level cnn. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2482–2486. IEEE, 2017.

[Bic07] Derek Bickerton. Language evolution: A brief guide for linguists. *Lingua*, 117(3):510–526, 2007.

[BK06] Henry Brighton and Simon Kirby. Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial life*, 12(2):229–242, 2006.

[CKB+20] Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. Compositionality and generalization in emergent languages. *arXiv preprint arXiv:2004.09124*, 2020.

[CLL+19] Michael Cogswell, Jiasen Lu, Stefan Lee, Devi Parikh, and Dhruv Batra. Emergence of compositional language with deep generational transmission. *arXiv preprint arXiv:1904.09067*, 2019.

[Cro08] William Croft. Evolutionary linguistics. *Annual review of anthropology*, 37:219–234, 2008.

[CT17] Jack K Chambers and Peter Trudgill. Dialect grammar: data and theory. In *Dialects of English*, pages 291–296. Routledge, 2017.

[FL02] Jerry A Fodor and Ernest Lepore. *The compositionality papers*. Oxford University Press, 2002.

[GRR22] Lukas Galke, Yoav Ram, and Limor Raviv. Emergent communication for understanding human language evolution: What's missing? *arXiv preprint arXiv:2204.10590*, 2022.

[GRR23] Lukas Galke, Yoav Ram, and Limor Raviv. What makes a language easy to deep-learn? *arXiv preprint arXiv:2302.12239*, 2023.

[HMP+16] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016.

[JGP16] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[KCBB19] Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. Egg: a toolkit for research on emergence of language in games. *arXiv preprint arXiv:1907.00852*, 2019.

[KCBB20] Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. Entropy minimization in emergent languages. In *International Conference on Machine Learning*, pages 5220–5230. PMLR, 2020.

[KCS08] Simon Kirby, Hannah Cornish, and Kenny Smith. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686, 2008.

[KO21] Jooyeon Kim and Alice Oh. Emergent communication under varying sizes and connectivities. *Advances in Neural Information Processing Systems*, 34:17579–17591, 2021.

[LB19] Fushan Li and Michael Bowling. Ease-of-teaching and language structure from emergent communication. *Advances in neural information processing systems*, 32, 2019.

[LB20] Angeliki Lazaridou and Marco Baroni. Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*, 2020.

[LD10] Gary Lupyan and Rick Dale. Language structure is partly determined by social structure. *PloS one*, 5(1):e8559, 2010.

[LHTC18] Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. Emergence of linguistic communication from referential games with symbolic and pixel input. *arXiv preprint arXiv:1804.03984*, 2018.

[Mar03] Gary F Marcus. *The algebraic mind: Integrating connectionism and cognitive science*. MIT press, 2003.

[MCH+20] Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, 2020.

[MM12] April McMahon and Robert McMahon. *Evolutionary linguistics*, volume 223. Cambridge University Press, 2012.

[PW10] Peter Pagin and Dag Westerståhl. Compositionality ii: Arguments and problems. *Philosophy Compass*, 5(3):265–282, 2010.

[RHW85] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[RMLA19a] Limor Raviv, Antje Meyer, and Shiri Lev-Ari. Compositional structure can emerge without generational transmission. *Cognition*, 182:151–164, 2019.

[RMLA19b] Limor Raviv, Antje Meyer, and Shiri Lev-Ari. Larger communities create more systematic languages. *Proceedings of the Royal Society B*, 286(1907):20191262, 2019.

[SKO04] Ann Senghas, Sotaro Kita, and Asli Ozyurek. Children creating core properties of language: Evidence from an emerging sign language in nicaragua. *Science*, 305(5691):1779–1782, 2004.

[Ste12] Luc Steels. Grounding language through evolutionary language games. *Language grounding in robots*, pages 1–22, 2012.

[XNR22] Zhenlin Xu, Marc Niethammer, and Colin A Raffel. Compositional generalization in unsupervised compositional representation learning: A study on disentanglement and emergent language. *Advances in Neural Information Processing Systems*, 35:25074–25087, 2022.

7 Code repository

The code and data can be found in the following repository:
https://github.com/imnublet/growing_meaning_spaces