VARIABLE COMPOSITIONALITY RELIABLY EMERGES IN NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Human languages are pervasively compositional: because the meaning of an expression is composed from the meaning of its parts, we can productively leverage our prior experience in order to communicate about novel meanings. Work looking at the languages that emerge when neural networks solve communicative tasks has shown that networks regularly develop languages that allow them to communicate and generalize to novel examples; surprisingly, that work has struggled to show that compositional systems reliably develop, leading to claims that a language's degree of compositionality has little bearing on how well it can generalise. We argue that the languages that emerge between networks are in fact straightforwardly compositional, just with variation. We introduce a variation-based framework for interpreting the mappings produced by neural networks in emergent communication games and find that they reliably exhibit straight-forward compositional structure, with a degree of natural language-like variation that obscures their compositionality under measures used in previous work. We show that early in training measures of variation correlate with generalization performance, but that this effect goes away later in training as the languages become regular enough to compositionally generalize. In an effort to decrease the variability of the emergent languages we show how reducing a model's capacity results in greater regularity, in line with claims about factors shaping the emergence of regularity in human language.

1 INTRODUCTION

Compositionality is a defining feature of natural language; the meaning of a phrase is composed from the meaning of its parts and the way they're combined (Cann, 1993). This underpins the powerful generalization abilities of the average speaker: compositionality allows us to readily interpret novel sentences and express novel concepts.

Robust generalization like this is a core goal of machine-learning: foundational to how we evaluate our models is seeing how well they generalize to examples that were withheld during training (Bishop, 2006). Deep neural networks show remarkable aptitude for generalization in-distribution (e.g. Dong & Lapata, 2016; Vaswani et al.), but a growing body of work questions whether or not these networks are generalizing compositionally (e.g. Kim & Linzen, 2020; Lake & Baroni, 2018), highlighting contexts where models consistently fail to generalize (e.g. in cases of distributional shift; Keysers et al. (2020)). These questions follow from earlier work that challenges if these architectures are even capable of behaving compositionally (Fodor & Pylyshyn, 1988).

Recent work in language emergence has explored whether compositional communicative systems emerge in neural networks when placed in conditions analogous to those that gave rise to human language (e.g. Choi et al., 2018; Kottur et al., 2017). In these simulations, multiple separate networks need to learn to communicate with one another about concepts, environmental information, instructions or goals via discrete signals - like sequences of letters - but are given no prior information about how to do so. A common setup is a 'reconstruction game' modelled after a Lewisian signalling game (Lewis, 1970), where a sender network describes a meaning using a signal, and a receiver network needs to reconstruct that meaning given the signal alone. The resulting set of mappings from meanings to signals can be thought of as a language.

Previous work has shown that in this setup models reliably develop a language that succeeds not only in describing the examples seen during training but also successfully generalizes to a held-out test set,

allowing accurate communication about novel meanings. Despite this evidence of the capacity to generalize, which is a product of compositionality in natural languages, existing analyses of those emergent languages provide little evidence of reliable compositional structure (see Lazaridou & Baroni, 2020, for a review), leading some to suggest that compositionality is not required in order to generalise (Chaabouni et al., 2020; Kharitonov & Baroni, 2020). This aligns with some work in core NLP suggesting that the compositionality of a model's representations doesn't necessarily indicate how robustly it will generalize (Andreas, 2019).

However, this interpretation leaves us with a major puzzle: if the languages that develop in these emergent models are non-compositional, how do they allow successful communication about thousands of unseen examples (e.g. Lazaridou et al., 2018; Havrylov & Titov, 2017)? If the meaning of a form is completely arbitrary rather than being in some way composed from its parts there should be no reliable way to use such a mapping to generalize to novel examples (Brighton, 2002). Here we provide an answer to this question, showing that emergent languages are characterised by *variation*, which masks their degree of compositionality from many of the measures used in the existing literature to assess compositionality. Existing measures take regularity as the defining feature of a compositional system, assuming that in order to be compositional separate semantic atoms need to be represented separately in the signal (Resnick et al., 2020; Chaabouni et al., 2020), or that symbols in the signal must have the same meaning regardless of the context they occur in (Kottur et al., 2017). Expecting that each part of meaning will be encoded in only one way, or that the resulting languages will have a strict canonical word order (Brighton & Kirby (2006) used in Lazaridou et al. (2018)). However, natural languages exhibit rich patterns of variation (Weinreich et al., 1968), frequently violating these three properties: forms often encode multiple elements of meaning (e.g. fusional inflection of person and number or gender and case), language is rife with homonymy (where the meaning of a form depends on context) & synonymy (where there are many ways of encoding a meaning in form), and many natural languages exhibit relatively free word order.

This observation allows us to resolve the conundrum currently posed by the emergent communication literature: the question of an emergent language's compositionality is related to, but distinct from, questions about its regularity. Accordingly we look at compositionality and regularity separately: we take a language's generalization performance as an indication of whether or not it is compositional, and introduce 4 measures of natural language-like variation including those mentioned above. We present results from a standard emergent communication model, and show that every run results in a highly-generalizing, and therefore compositional language. We also show that the languages which emerge tend to exhibit a high degree of variability, which explains why previous metrics of compositionality would classify them as non-compositional. We compare scores across our variation measures with a perfectly regular compositional language and a maximally non-compositional random mapping, showing that the languages that emerge in our model are consistently more regular than a random mapping. Comparing each emergent language's regularity to its ability to generalize to an out of distribution test set, we show that degree of regularity correlates strongly with generalization performance early in training, but that this effect goes away as the model converges to a language that is sufficiently regular to compositionally generalize.

In a final set of experiments we look at how to increase the regularity of an emergent language. Limitations on human's memory and cognitive capacity are thought to be a driving force in the emergence of compositional structure and regularity in natural language (Kirby, 2001; Hudson Kam & Chang, 2009; Smith & Wonnacott, 2010). We reduce the capacity of our models by reducing the size of the hidden layers, and show that lower capacity models develop more regular languages, as predicted by accounts linking learner capacity and regularity in natural language. We summarise the core contributions of this work as follows:

- We introduce a variation based framework for analysing regularity in a mapping, and apply it to emergent languages in neural networks.
- We find that every run of our model results in clear compositional structure, with a degree of variation.
- We show that an emergent language's regularity does correlate strongly with generalization early in training, but that this effect goes away as the model converges to a sufficiently regular language to generalize compositionally

- Reducing model capacity reliably results in the emergence of languages with greater regularity, in line with work on regularity in natural language.

2 QUANTIFYING COMPOSITIONALITY

A number of measures already exist for assessing the degree of compositionality in a mapping from meaning to form. Many of these assess particular kind(s) of variation as a proxy for compositionality, and as a result treat certain natural-language-like variation as evidence of non-compositionality. We briefly discuss some of the most relevant measures then introduce four kinds of variation - drawn from sources of variation attested in natural language - and ways of quantifying each of them individually.

2.1 EXISTING MEASURES

Topographic Similarity (Topsim) (Brighton & Kirby, 2006) has been used as a measure of compositionality in a wide array of contexts from early computational models of language emergence (Smith et al., 2003), to human experiments (Kirby et al., 2008), and more recently in neural network based studies (Lazaridou et al., 2018). It assumes that in a compositional system where a whole signal is composed from reusable parts, similar meanings will map to similar signals. This can be assessed by measuring the correlation between pairwise meaning-distances and edit distances between their associated signals: a perfectly regular compositional language without variation achieves a correlation score close to 1, while a non-compositional (random) mapping between meanings and signals achieves a correlation close to 0. Often the emergent languages that develop in neural networks have topsim scores somewhere between these values (e.g. in our own results reported in Table 1, between 0.2 and 0.3), which is taken to indicate that the meaning signal mapping is not highly compositional.

Residual Entropy & Posdis Residual entropy (Resnick et al., 2020) and Posdis (Chaabouni et al., 2020) provide entropy-based measures capturing the extent to which each part of the signal univocally refers to a component of the meaning (e.g. subject, object, verb). These measures therefore assume disentanglement between meaning and form, with each part of a signal encoding one and only one part of the meaning. Similar to topsim, a perfectly regular, perfectly disentangled language will score near 1, while a completely entangled non-compositional language with every part of the meaning encoded in every part of the signal will score near 0.

While languages that score highly on topsim, residual entropy and posdis are necessarily compositional, not all compositional languages will score highly; these measures provide a way of evaluating the presence of sufficient but not necessary features of a compositional system. In particular, they assume that compositional languages will be highly regular, and do not allow for the kinds of variability that are seen in (compositional) natural languages.

2.2 KINDS OF VARIATION

Unanimity (Synonymy & Homonymy): Synonymy is minimised when each atom in a meaning maps to a single atom in a form. Homonymy is minimised when each atom of form maps to a single atom of meaning. A perfectly regular compositional language minimises both using one-to-one mappings between meaning and form, but natural language is rife with both synonymy and homonymy, often exhibiting one-to-many and many-to-one mappings (e.g. 'loves', 'adores', 'treasures', 'cherishes' all convey approximately the same concept, exemplifying a one-to-many meaning-word mapping; the homonymous 'bank' refers to a financial institution, the act of turning a plane, and the land at the side of a river).

A one-to-many mapping (synonymy) isn't a problem for compositionality, as each different realisation of a meaning can be composed with others in a sentence. A many-to-one mapping (homonymy) can similarly be used compositionally, with meaning disambiguated by context. Natural language shows us that a compositional system can exhibit high levels of synonymy and homonymy while retaining the productive generalisation afforded by compositionality.

Word Order Freedom: Word order freedom is minimized when each role in the meaning (e.g. subject, verb, object) is encoded in the same place in the signal each time, resulting in a single

canonical word order. Looking at a language like Basque we can see that a compositional language can support a number of different grammatical word orders. With at least two equivalently valid translations of 'Ollie saw Ernest:' *Ollie Ernest ikusi zuen, Ollie ikusi zuen Ernest.* Even in English which has relatively strict word order we see processes like topicalization that result in alternate orders that are equally acceptable *Let's go down to the lake for some fun; For some fun, let's go down to the lake*, or even more commonly dative alternations like *Ollie gave Orson a book; Ollie gave a book to Orson*.

The freedom of order by which meanings are expressed in form has little bearing on whether or not the resulting form can be composed or interpreted compositionally; even when there is maximal word order freedom the resulting language can still be clearly compositional, with order conveying little information.

Entanglement: Entanglement is minimised when each part of a meaning is encoded in a different part of form. While a dis-entangled language is likely compositional consider the English past tense form of 'go.' 'Went' is irregular, encoding meaning and tense together, in contrast to the hypothetical regular form 'goed' where the verb and tense are encoded in separate parts of form. Despite this we can go on to use the entangled form 'went' compositionally in a sentence: *Ollie went down to the shore.* While maximal entanglement where every part of a meaning is encoded in every part of form would be non-compositional, the existence of even a high degree of entanglement does not preclude compositionality, given the entangled forms can be straight-forwardly recomposed with others.

2.3 Variation-Based Measures

We define measures for the four kinds of variation outlined above: Synonymy, Entanglement, Word-Order Freedom, and Homonymy. Each of these measures is bounded between 0 and 1, where 0 indicates a perfectly regular language with no variation, and 1 represents a maximally variable language. We generate a perfectly regular compositional language, and maximally irregular non-compositional language (where each meaning maps to a unique randomly-generated signal). The regular language scores near 0 across our measures of variation, while the random mapping scores near 1, as shown in table 1.

All four measures start with the same underlying process, generating a tensor that describes the mapping between meanings and signals probabilistically, in terms of a probability distribution over characters in each signal position given a semantic atom in a semantic role. Encoding, for example, how likely it is that character 'A' is in signal position 1 because 'Ollie' is in the subject role of the signal's meaning. We can quantify this as a straightforward conditional probability using maximum likelihood estimation, shown in equation 1. We estimate this repeatedly for every atom ($\forall atom_{r,i} \in A_r$) in every role in the meaning ($\forall r \in R$), looking at every character ($\forall char_{p,j} \in C$) in every position of every signal ($\forall p \in P$).

$$\mathbb{P}(char_{p,j}|atom_{r,i}) = \frac{count(char_{p,j}, atom_{r,i})}{count(atom_{r,i})} \tag{1}$$

The resulting tensor describes how often each letter occurs in a position, given a certain atom (like Ollie) in a role in the meaning[1]. This tensor has dimensions semantic roles $\times$ semantic atoms $\times$ max signal length $\times$ characters[2], where the last axis is a probability distribution over all possible characters in a given position - here denoted by $\mathbb{P}(char_p|atom_{r,i})$.

Synonymy is evaluated by measuring, for each semantic atom in each role, how many different characters appear in each position in the signal. For example when $atom_i = Ollie$ and $r = Subject$ how many characters have non-zero probability in each signal position? A perfectly regular language where 'Ollie' is always encoded by 'A' in position 1 would have a probability of 1.0 on 'A' in position 1. A maximally variable language would have a uniform distribution over all characters. We can take the entropy over characters in a position $\mathcal{H}(\mathbb{P}(char_p|atom_{r,i}))$ as a measure of synonymy in that

[1]Here we use semantic roles given the meanings are sentences, this can be generalised to any attributes a dataset exhibits.

[2]For all experiments reported here these values are $3 \times 25 \times 6 \times 26$

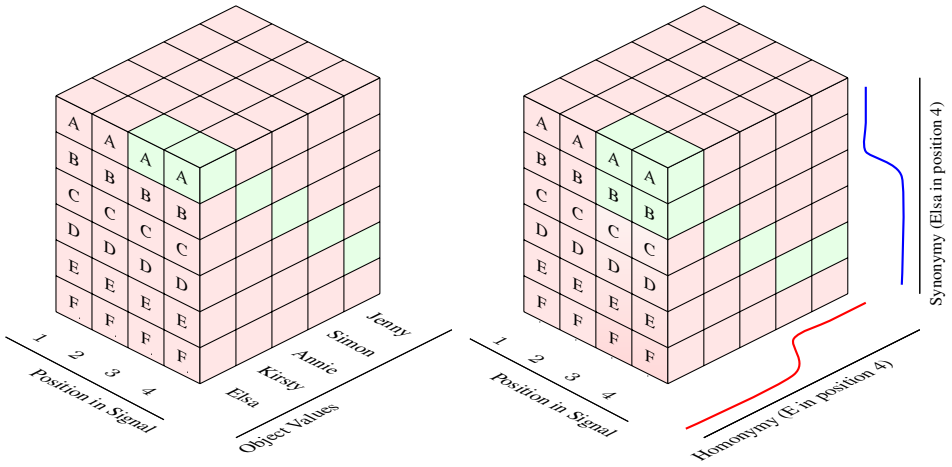

Figure 1: A depiction of the probability tensor built with equation 1. Values are artificial with green indicating high probability and red low. Only a portion of the tensor is shown where $r = Object$. **(Left:)** An ideal regular language, Elsa is always encoded by 'AA' in the final two positions, Kirsty by 'BB', Annie by 'CC' etc. **(Right:)** The same cube is shown (object labels removed) for a language with basic synonymy (Elsa can be encoded by 'A' or 'B') and homonymy (Jenny and Simon are both encoded by 'E'). We quantify the degree of synonymy by taking the entropy of each column (equation 2a) and the degree of homonymy by taking the entropy of each row (equation 5b)

position. Equation 2a gives us a set of $|P|$ values corresponding to entropy over characters in each position $\forall p \in P$. An illustration of this measure is shown in figure 1.

$$\mathbb{S}(atom_{r,i}) = \Big\{ \mathcal{H}\Big(\mathbb{P}(char_p|atom_{r,i})\Big) \colon \forall p \in P \Big\} \tag{2a}$$

We take the position with the lowest entropy as a lower-bound estimate of synonymy for that $atom_{r,i}$. [3] In order to bound the value we divide it by the entropy of a same-size uniform distribution $\mathcal{H}_u\Big(\mathbb{P}(char_p|atom_{r,i})\Big)$. The synonymy of an entire language ($\mathcal{L}$) is obtained by averaging across all semantic atoms in a role, and then across all roles.

$$Synonymy(\mathcal{L}) = \frac{1}{|R|} \sum_{r=1}^{|R|} \frac{1}{|A_r|} \sum_{i=1}^{|A_r|} \frac{min\Big(\mathbb{S}(atom_{r,i})\Big)}{\mathcal{H}_u\Big(\mathbb{P}(char_p|atom_{r,i})\Big)} \tag{2b}$$

The result is bounded between 0 and 1: a language with no synonymy where each word is encoded by a single character in a position achieving close to 0, and maximal synonymy where any character can refer to each word achieving close to one. These bounds are shown empirically in table 1.

Word Order Freedom We also want to know if each semantic role is consistently encoded in a particular part of the signal - reflecting a regular word order. A language with free word order is equally likely to encode any $Role \in R$ in any position in the signal, while a maximally regular language always encodes atoms from the same role in the same position(s). If a given $atom_{r,i}$ is not encoded in a position we would expect the distribution over characters in that position to be roughly uniform. So we can take the entropy for each position $\mathbb{S}(atom_{r,i})$ obtained in equation 2a, and average across all atoms in that role $\forall i \in A_r$. If all the atoms in a role are encoded in the same

[3] In practice we only take the minimum (as opposed a mean across positions) as positions where that atom is not encoded will have a roughly uniform distribution - any character is equally likely as it's not conditioned by $atom_{r,i}$, so averaging across positions artificially inflates the measure.

position the distribution resulting from the mean will be non-uniform, with some positions having lower mean entropy than others (an illustration is provided in appendix figure 4).

$$\mathbb{F}(role) = \frac{1}{|A_r|} \sum_{a=1}^{|A_r|} \mathbb{S}(atom_{r,i}) \tag{3a}$$

To get a lower-bound estimate of the language-level word-order freedom we take the minimum from the mean distribution $\mathbb{F}(role_r)$ and divide it by the entropy of a uniform distribution to bound between 0 and 1, then average across all roles:

$$Freedom(\mathcal{L}) = \frac{1}{|R|} \sum_{r=1}^{|R|} \frac{min\Big(\mathbb{F}(role_r)\Big)}{\mathcal{H}_u\Big(\mathbb{P}(char_p|atom_{r,i})\Big)} \tag{3b}$$

Entanglement In addition to looking at whether a role is consistently encoded in a position, we can see if two roles are consistently encoded in the same (or different) positions. We compare the means $\mathbb{F}(role)$ from equation 3a for each possible pair of roles $r_i, r_j \in {}^{R}C_2$ by taking the magnitude of their difference, if two roles are encoded in the same position the result will be close to zero. If the roles are maximally disentangled then the result will be close to the $max\Big(\mathbb{F}(role_i), \mathbb{F}(role_j)\Big)$ for that position. To get a lower bound estimate of two roles' entanglement we take the maximum of the difference and divide by the pre-difference max. An illustration of this measure is shown in figure 5.

$$Entanglement(\mathcal{L}) = 1 - \frac{1}{|{}^{R}C_2|} \sum_{r_i,r_j}^{{}^{R}C_2} \frac{max\Big(|\mathbb{F}(role_i) - \mathbb{F}(role_j)|\Big)}{max\Big(\mathbb{F}(role_i), \mathbb{F}(role_j)\Big)} \tag{4}$$

When the resulting value approaches 0 all distinct roles are expressed in different parts of the signal, as it approaches 1 all roles are encoded in the same positions in the signal (e.g. in a holistic language, and a random mapping, all roles are effectively encoded in all positions).

Homonomy Finally we assess homonomy to see how uniquely a character in a position of a signal is used to refer to a single word in the meaning. Or if many different meanings map to that character. This measure is similar to the synonymy measure applied to a different axis of the probability tensor $\mathbb{P}$ as depicted in figure 1. For each character in each position we create a probability distribution over atoms in a role $\mathbb{H}(char_{p,j}, r)$

$$\mathbb{H}(char_{p,j}, r) = \frac{\Big\{\mathbb{P}(char_{p,j}|atom_{r,i}) : \forall i \in A_r\Big\}}{\sum_{i=1}^{|A_r|} \mathbb{P}(char_{p,j}|atom_{r,i})} \tag{5a}$$

To get a lower-bound estimate of language level homonymy we take the position with the lowest entropy over words and divide by the entropy of a uniform distribution to bound between 0 and 1, then average across all characters and roles.

$$Homonomy(\mathcal{L}) = \frac{1}{|R|} \sum_{r=1}^{|R|} \frac{1}{|C|} \sum_{j=1}^{|C|} \frac{min\Big\{\mathcal{H}\Big(\mathbb{H}(char_{p,j}, r) : \forall p \in P\Big)\Big\}}{\mathcal{H}_u\Big(\mathbb{H}(char_{p,j}, r)\Big)} \tag{5b}$$

When the resulting value is close to 1 every semantic atom maps to the same character(s) in each position. When it approaches 0 each character in each position univocally refers to a single semantic atom. [4]

[4] Note that the model consistently uses fewer characters than are available. For the homonomy measure this means the unused characters need to be dropped before calculating the measure as an additional step before equation 5a.

3 METHODS

Models We implement a reconstruction game with a sender network and receiver network. The sender network is comprised of an embedding layer, linear layer, and a GRU (Cho et al., 2014). A linear layer is used as the input is of fixed length, so can be presented at once as a one-hot encoding - while a GRU spells out the variable-length signal a character at a time. The maximum signal length used here is 6, with 26 characters available to the model in each position. The receiver architecture is the inverse of the sender, with a GRU mapping signal to hidden state, and a linear layer mapping that to the output space. The sender is optimized using REINFORCE (Williams, 1992) due to the discrete channel, while the receiver is optimised using ADAM (Kingma & Ba, 2014). Models are implemented using pytorch (Paszke et al., 2019), and make use of portions of code from the EGG repository (Kharitonov et al., 2019).

Data The sender is shown examples drawn from a meaning space of two place predicates (e.g. *Ollie loves Osgood*) generated using a context free grammar, with three roles: subject, verb, and object. These are mapped to a signal space comprised of a sequence of discrete characters. This setup, while simple, is in-line with previous work in this area (e.g. Guo et al., 2021; Chaabouni et al., 2020).

O.O.D. Evaluation Previous emergent communication work typically evaluates generalization on an in-distribution held out test-set. In order to better align our findings with the broader literature on compostional generalization in neural networks we implement a version of the maximum compound divergence (MCD) algorithm from Keysers et al. (2020), and report results for both in-distribution generalization, and out-of-distribution generalization to an MCD split.

The effect of capacity on variation Constraints on human memory and cognitive capacity are thought to be key drivers of regularity in natural language (Kirby, 2001; Hudson Kam & Chang, 2009). To test if similar factors influence the variability of emergent languages in neural networks, and to confirm that our measures of variation work as intended, we vary the capacity of our models to see if this interacts with the resulting degree of variation. We look to see if models with less capacity arrive at more regular languages than their larger counterparts. We vary model capacity by varying the size of the hidden layers used by the model reporting and comparing results of three different model sizes $small$, $medium$, and $large$ with hidden layer sizes 250, 500, and 800 respectively.

4 RESULTS AND DISCUSSION

Every run of the model in every condition scores well on o.o.d. generalization, which (as argued above and by e.g. (Brighton, 2002)) likely indicates a compositional language. We start by considering the relationship between measures of generalization, compositionality and variation in our $small$ model, before considering the effects of model capacity on those measures.

Figure 2 shows the relationship between o.o.d generalization accuracy and various measures of variation & regularity, for our $small$ model, over the timecourse of the language's emergence across 20 runs of the model. As can be seen in the figure, early on variation measures correlate negatively with generalization while topsim and posdis correlate positively - in other words, measures of variation and regularity are predictors of generalization as would be expected. These correlations trend towards 0 as the models converge on languages which are all sufficiently regular to enable compositional generalization but which each differ in the variability and regularity they converge to [5].

Our results for all model sizes are summarised in table 1.

Results from all four variation measures show that the languages from every condition of the model exhibit greater regularity than the random language. Additionally each condition reduces the degree of variation over the course of the run, with the variation at the best generalizing epoch lower than at an epoch selected from early in training [6]. This supports our assertion that correlations between

[5]We include plots of raw measure data from early and late epochs in appendix figure 6, along with more implementation details of the rolling mixed effects model used to analyse their relationship.

[6]At epoch 0 the language is a direct product of the random seed used to initialize the model. As we're interested in the languages that emerge from the models trying to communicate we select an epoch by sampling the least regular language that occurs between epochs 1 and 10 for each run to compare with later in training.

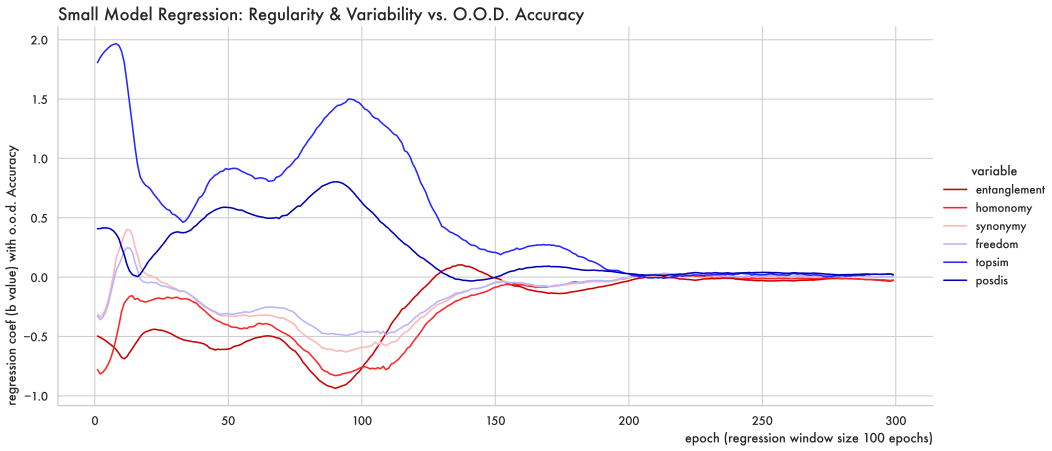

Figure 2: The model is fit to a window of data from 100 epochs at a time across 20 initializations. The window slides forward one epoch at a time (i.e. epochs 0-100, 1-101 ...) and fits a different model between o.o.d. accuracy and each measure of variation for each window. Shown are the regression coefficients (b values) of our four measures of variation, and two previous measures of regularity (topsim and posdis) with o.o.d. generalization accuracy for the *small* model for each window. This process shows how the relationship changes throughout training; later on as the model converges correlations trends towards 0 and non-significance. Suggesting that early in training more regular languages generalize better but eventually all languages become sufficiently regular to generalize robustly.

Model	i.i.d. acc		o.o.d. acc		synonymy		entanglement		freedom		homonomy		topsim		posdis	
ideal					0.00		0.00		0.00		0.12		0.62		1.00	
random					0.99		1.00		0.99		0.99		0.00		0.00	
small	**97.54**	±0.99	**72.86**	±14.14	**0.46**	±0.05	**0.54**	±0.07	**0.49**	±0.06	**0.53**	±0.05	**0.21**	±0.02	**0.24**	±0.05
Δ *best o.o.d.*					−0.20	±0.05	−0.42	±0.07	−0.19	±0.06	−0.18	±0.05	0.12	±0.03	0.22	±0.06
medium	**97.73**	±1.18	**82.13**	±7.22	**0.52**	±0.09	**0.60**	±0.14	**0.54**	±0.10	**0.58**	±0.09	**0.19**	±0.05	**0.19**	±0.09
Δ *best o.o.d.*					−0.13	±0.09	−0.35	±0.14	−0.12	±0.10	−0.13	±0.10	0.10	±0.05	0.17	±0.10
large	**97.53**	±1.04	**81.33**	±6.06	**0.63**	±0.05	**0.80**	±0.08	**0.66**	±0.04	**0.69**	±0.04	**0.14**	±0.02	**0.08**	±0.04
Δ *best o.o.d.*					−0.01	±0.07	−0.13	±0.11	−0.01	±0.06	−0.03	±0.04	0.03	±0.03	0.05	±0.04

Table 1: Main accuracy and variation results. We show the mean and 95% confidence interval across 20 runs, taken from the epoch with the best o.o.d. generalization performance, along with the change in measures $\Delta best$ between an epoch early in training and the best generalizing one. Also included are the variation measures applied to a rule-generated ideally regular compositional language, and a maximally variable random one. Two measures of regularity from previous work (topsim (Brighton & Kirby, 2006) and posdis (Chaabouni et al., 2020)). For measures in the grey cells a more regular language scores closer to 1 - the inverse of our variation measures.

variation and generalization trend to 0 as languages become sufficiently regular to succeed at the task. That said the models' languages rarely approach the degree of regularity exhibited by the ideally compositional language. Instead they exhibit a substantial degree of variation across all four measures. Figure 4 visualizes the word order for the run of our *small* model with the highest variation. It shows that while the language is still much more variable than a perfectly regular one (the illustrative language has $freedom = 0.57$, a compositional language with fixed word order has $freedom = 0$), it nonetheless exhibits a high degree of word order regularity, with verbs most likely encoded at the start of the signal, subjects in the middle, and objects at the end, but with each individual word sometimes being encoded slightly differently.

We see a strong effect of capacity on regularity, with an increase in the number of trainable parameters resulting in an increase in variability across all measures with *large* arriving at significantly more variable languages than *small* or *medium* ($p < 0.05$). Spearman correlations show model size does not correlate significantly with o.o.d. accuracy ($p = 0.24$) but correlates with synonymy ($r = 0.67$), word order freedom ($r = 0.69$), entanglement ($r = 0.68$), and homonymy ($r = 0.68$) indicating larger models develop more variable languages. Interestingly model size correlates negatively with

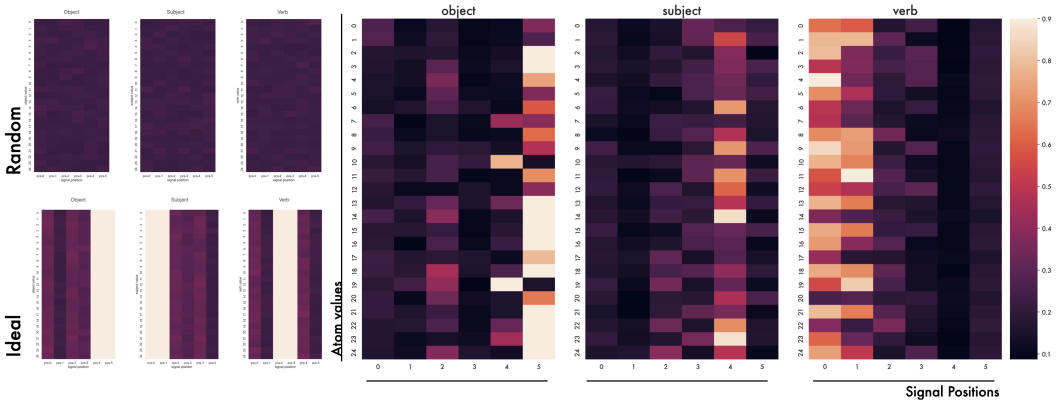

Figure 3: Plots showing the max from the distribution over characters for each semantic atom in each signal position, with a plot for each separate role (object, subject, verb). The signal positions 1 to 6 are on the x axis, and on the y axis are ids for each $atom_i \in A_r$. Shown to the left are these plots for the synthetic ideally regular compositional language (with SVO order), and the maximally variable random mapping. The large plot shows data from the run of the $small$ model with the *highest* variation, illustrating how these languages often develop a surprising degree of word-order regularity despite being quite variable. This run's variation scores: $freedom = 0.57, entanglement = 0.61, homonymy = 0.61, synonymy = 0.51, topsim = 0.28, posdis = 0.26$

topsim ($r = -0.63$) and posdis ($r = -0.71$) all correlations are well below the threshold for significance ($p < 0.00001$). This result is in line with work that points to constraints on human cognition as a key driver of regularization in natural language, suggesting that similar factors shape the regularity of emergent communication in neural networks.

The negative correlation between existing measures of compositionality (topsim and posdis) and model size strongly suggests that rather than tracking compositionality these measures implicitly track the degree of regularity in a language, especially given that the magnitude of their correlation coefficient is similar to that of our measures of variation. Indicating these measures are affected in a similar way, and to a similar degree to explicit assessments of variability. This helps to explain the apparent conundrum that compositionality appears not to correlate with generalization; instead our results suggest that compositional but variable languages reliably emerge, with the degree of variability being influenced by the network's capacity to tolerate variability. Similar to how high variation in natural language doesn't impede its compositionality, the emergent languages can generalize robustly without needing to be perfectly regular.

5 Conclusion

We have presented a novel approach to interpreting the languages that develop in deep-learning models of emergent communication. We take generalization ability as an indication of whether or not a language is compositional and show how previous assessments of compositionality can be best framed as assessing the degree of regularity in a language. We introduced four novel measures for assessing the degree of variation in a mapping, using them to show that the languages that emerge, while reliably compositional, exhibit a high degree of variation. By accounting for that variation we can see striking similarities between the structure of the languages that emerge and structures in natural language. Early in training variation is a strong predictor of generalization, but this effect goes away as models converges to a language sufficiently regular to succeed at the task. We also show that the higher a model's capacity the less regular the language that emerges, showing that limitations on memory and cognitive capacity, hypothesised to influence regularity in natural language, also influence the regularity of emergent languages in neural networks.

REFERENCES

Jacob Andreas. Measuring Compositionality in Representation Learning. *arXiv:1902.07181 [cs, stat]*, April 2019. URL http://arxiv.org/abs/1902.07181. arXiv: 1902.07181.

Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, New York, 2006. ISBN 978-0-387-31073-2.

Henry Brighton. Compositional Syntax From Cultural Transmission. *Artificial Life*, 8(1):25–54, January 2002. ISSN 1064-5462, 1530-9185. doi: 10.1162/106454602753694756. URL https://direct.mit.edu/artl/article/8/1/25-54/2396.

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

Ronnie Cann. *Formal semantics an introduction*. Cambridge University Press, Cambridge [etc., 1993. ISBN 978-1-139-16631-7. URL http://0-www.ebooks.cambridge.org.cataleg.uoc.edu/ebook.jsf?bid=CBO9781139166317. OCLC: 1120437841.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. Compositionality and Generalization In Emergent Languages. pp. 16, 2020.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *SSST@EMNLP*, 2014.

Edward Choi, Angeliki Lazaridou, and Nando de Freitas. Compositional Obverter Communication Learning From Raw Visual Input. *arXiv:1804.02341 [cs]*, April 2018. URL http://arxiv.org/abs/1804.02341. arXiv: 1804.02341.

Li Dong and Mirella Lapata. Language to Logical Form with Neural Attention, June 2016. URL http://arxiv.org/abs/1601.01280. arXiv:1601.01280 [cs].

Jerry A. Fodor and Zenon W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, March 1988. ISSN 00100277. doi: 10.1016/0010-0277(88)90031-5. URL https://linkinghub.elsevier.com/retrieve/pii/0010027788900315.

Shangmin Guo, Yi Ren, Kory Mathewson, Simon Kirby, Stefano V. Albrecht, and Kenny Smith. Expressivity of Emergent Language is a Trade-off between Contextual Complexity and Unpredictability. *arXiv:2106.03982 [cs]*, June 2021. URL http://arxiv.org/abs/2106.03982. arXiv: 2106.03982.

Serhii Havrylov and Ivan Titov. Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols. *arXiv:1705.11192 [cs]*, November 2017. URL http://arxiv.org/abs/1705.11192. arXiv: 1705.11192.

Carla L. Hudson Kam and Ann Chang. Investigating the cause of language regularization in adults: Memory constraints or learning effects? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3):815–821, 2009. ISSN 1939-1285, 0278-7393. doi: 10.1037/a0015097. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/a0015097.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. MEASURING COMPOSITIONAL GENERALIZATION: A COMPREHENSIVE METHOD ON REALISTIC DATA. pp. 38, 2020.

Eugene Kharitonov and Marco Baroni. Emergent language generalization and acquisition speed are not tied to compositionality. *arXiv:2004.03420 [cs]*, April 2020. arXiv: 2004.03420.

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. EGG: a toolkit for research on Emergence of lanGuage in Games. *arXiv:1907.00852 [cs]*, October 2019. URL http://arxiv.org/abs/1907.00852. arXiv: 1907.00852.

Najoung Kim and Tal Linzen. COGS: A Compositional Generalization Challenge Based on Semantic Interpretation. *arXiv:2010.05465 [cs]*, October 2020. URL http://arxiv.org/abs/2010.05465. arXiv: 2010.05465.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

S. Kirby. Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110, April 2001. ISSN 1089778X. doi: 10.1109/4235.918430. URL http://ieeexplore.ieee.org/document/918430/.

S. Kirby, H. Cornish, and K. 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, August 2008. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0707835105. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0707835105.

Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog. *arXiv:1706.08502 [cs]*, August 2017. URL http://arxiv.org/abs/1706.08502. arXiv: 1706.08502.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *arXiv:1711.00350 [cs]*, June 2018. URL http://arxiv.org/abs/1711.00350. arXiv: 1711.00350.

Angeliki Lazaridou and Marco Baroni. Emergent Multi-Agent Communication in the Deep Learning Era. *arXiv:2006.02419 [cs]*, July 2020. URL http://arxiv.org/abs/2006.02419. arXiv: 2006.02419.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. Emergence of Linguistic Communication from Referential Games with Symbolic and Pixel Input. *arXiv:1804.03984 [cs]*, April 2018. URL http://arxiv.org/abs/1804.03984. arXiv: 1804.03984.

David Lewis. *Convention: A philosophical study*. John Wiley & Sons, 1970.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Cinjon Resnick, Abhinav Gupta, Jakob Foerster, Andrew M. Dai, and Kyunghyun Cho. Capacity, Bandwidth, and Compositionality in Emergent Language Learning. *arXiv:1910.11424 [cs, stat]*, April 2020. URL http://arxiv.org/abs/1910.11424. arXiv: 1910.11424.

Kenny Smith and Elizabeth Wonnacott. Eliminating unpredictable variation through iterated learning. *Cognition*, 116(3):444–449, September 2010. ISSN 00100277. doi: 10.1016/j.cognition.2010.06.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0010027710001320.

Kenny Smith, Henry Brighton, and Simon Kirby. Complex systems in language evolution: the cultural emergence of compositional structure. *Advances in complex systems*, 6(04):537–558, 2003.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. pp. 11.

Uriel Weinreich, William Labov, and Marvin Herzog. Empirical foundations for a theory of language change. *University of Texas Press*, pp. 100, 1968.

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

A APPENDIX

A.1 WORD ORDER FREEDOM GRAPHIC

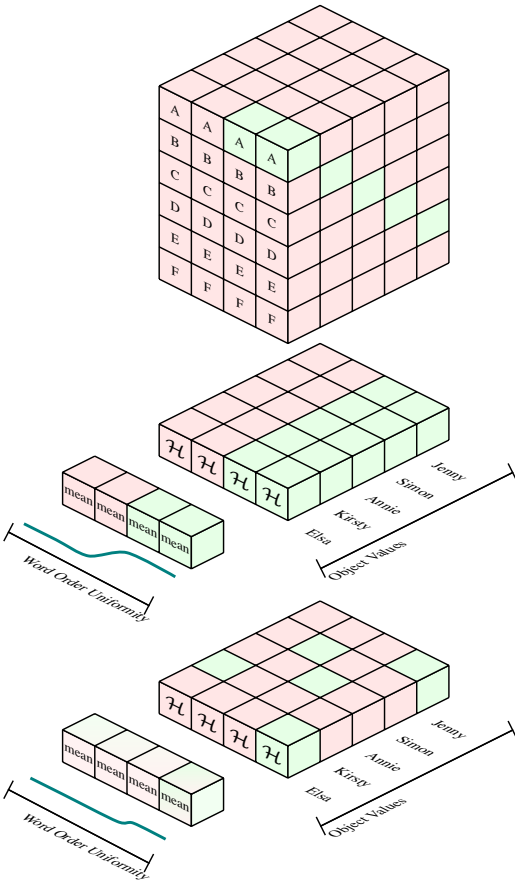

Figure 4: The figure shows an illustration of the Freedom measure applied to a regular and variable language. At the top is a reproduction of the perfectly regular language cube from figure 1 where $role = Object$ where every semantic atom in the object position is encoded by two letters at the end of the signal. Red indicates low probability, and green high. Directly below it is the entropy of each column, which reveals the word order freedom (conversley for the entropy plots green corresponds to low entropy, red to high) this is the operation completed in equation 3a. As the language is perfectly regular the entropy is low in the last two positions across all semantic atoms in the object role. As a result when we take the mean of the entropies we get a non-uniform distribution: entropy is consistently lower in the two final positions. To get a measure between 0 and 1 we divide the lowest value in the mean distribution by the entropy of a uniform distribution of the same size this is done in equation 3b. The bottom graphic shows the entropy illustration for an alternate, irregular language, where different atoms in the object position are most likely to be encoded in a variety of signal positions. As a result the mean distribution is nearly uniform, so when we take the minimum it will be closer to the entropy of a uniform distribution resulting in a score close to 1.

A.2 ENTANGLEMENT GRAPHIC

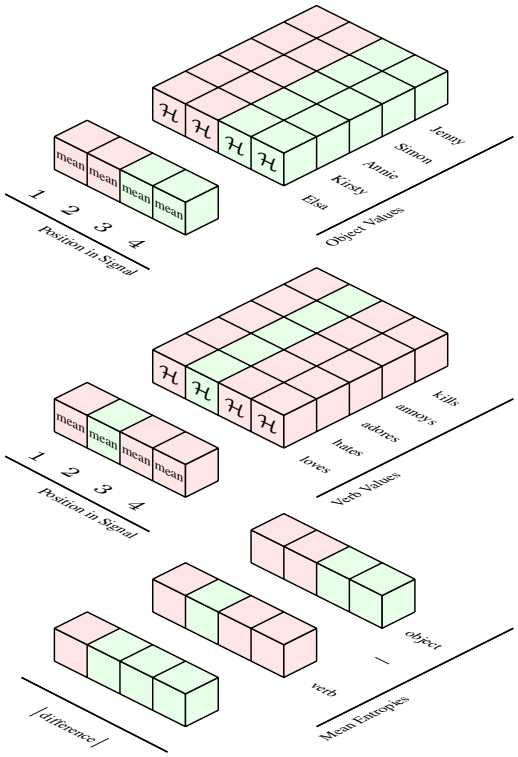

Figure 5: The figure shows the entanglement measure applied to a perfectly regular language. We start with the column entropies shown in the uniformity graphic, and which are calculated using equation 3a. Shown are the column entropies for two different roles, object and verb. The mean of both are taken and then compared in the lower part of the graphic. We take the absolute value of the difference of the two averaged entropies. If two roles are encoded in the same part of the signal then the result will be close to zero: they both have consistently similar in the same parts of the signal. In the illustration this would look like the final difference bar being entirely red, here it's green in 3 positions indicating disentanglement there. The maximum of the difference is then divided by the maximum of the mean entropy for the respective roles before they were subtracted - this bounds the measure between 0 and 1 and makes the measure a kind of percentage of the overlap between roles (this operation is shown in equation 4). If the roles are disentangled we end up subtracting a high entropy value from a low entropy value in each position, meaning the magnitude of the difference will be substantially greater than zero. We do this for each possible paring of roles, getting entanglement scores for each pairing, then mean them to get language level entanglement.

A.3 ROLLING MIXED EFFECTS MODEL IMPLEMENTATION

We implement a rolling mixed effects model using the python statsmodels package. We fit a separate model for each of our 6 independent variables: synonymy, entanglement, freedom, homonymy, topsim, posids. The dependent variable across all of them is o.o.d. generalization performance. For each model we include two random effects, random intercepts based on the seed used to initialize that run of the model, and random slopes for the epochs of training. This allows the model to account for variation between different models, given that some initializations outperform others.

The model is fit to a window of 100 epochs of training data at a time, at each step it fits a regression to 100 epochs of data for 20 initializations of the model. It then moves forward one epoch at a time (e.g. the first fit of the model is on epochs 0-100, the second on 1-101, the third 2-102, etc.). We plot the resulting regression coefficient (b value) obtained from each mixed effects model fit to each IV, at each window of data.

For reference the corresponding command to run this model with the LMER package in R (a standard method for fitting these kinds of models) is: lmer("ood_acc IV + (1 + Epoch | Seed)", data=results) where IV is one of the variation measures.

A.4 O.O.D. Accuracy Vs. Variation Slices

In addition to the regression analysis presented in the results section we show relational plots for two different epochs in training. One from mid way through and one from late in training near convergence. In line with the regression analysis a more linear relationship between o.o.d. performance and variation is visible earlier in training before the model converges. Entanglement in particular shows a steep negative relationship in the 100 epoch plot but is totally scattered by epoch 500. Were we to only assess the relationship between generalization and variation at the end of training we would could easily conclude in line with previous work that they were not meaningfully related.

It's worth noting that the pattern here may not appear as salient as it appears in the rolling mixed effects model presented earlier, there are two major reasons for this: first the rolling model considers 100 epochs at a time, rather than a single slice with only 20 data points, providing it with 100 times the data visualized here by which to assess the relationship between variation and generalization. Secondly the rolling model has a random intercept based on the seed used in each run of the model. This is important because in line with other work on o.o.d. generalization we see a substantial effect of initialization on generalization performance, by including it as a random effect the rolling model can look at each seed separately to see if each seed's generalization performance over the run is related to its language's variation. So while in the visualizations below we may see some seeds which appear like outliers, the rolling model accounts for this, fitting a separate intercept for each run.

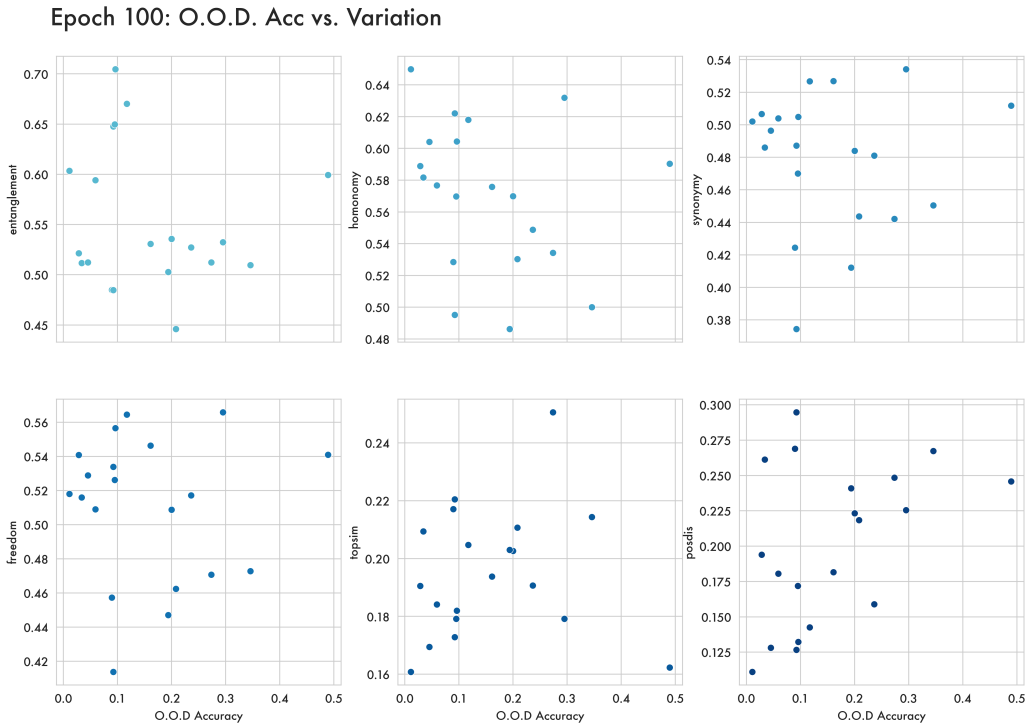

Figure 6: Plots show each of the variation measures plotted against o.o.d. accuracy at the 100th epoch of training.

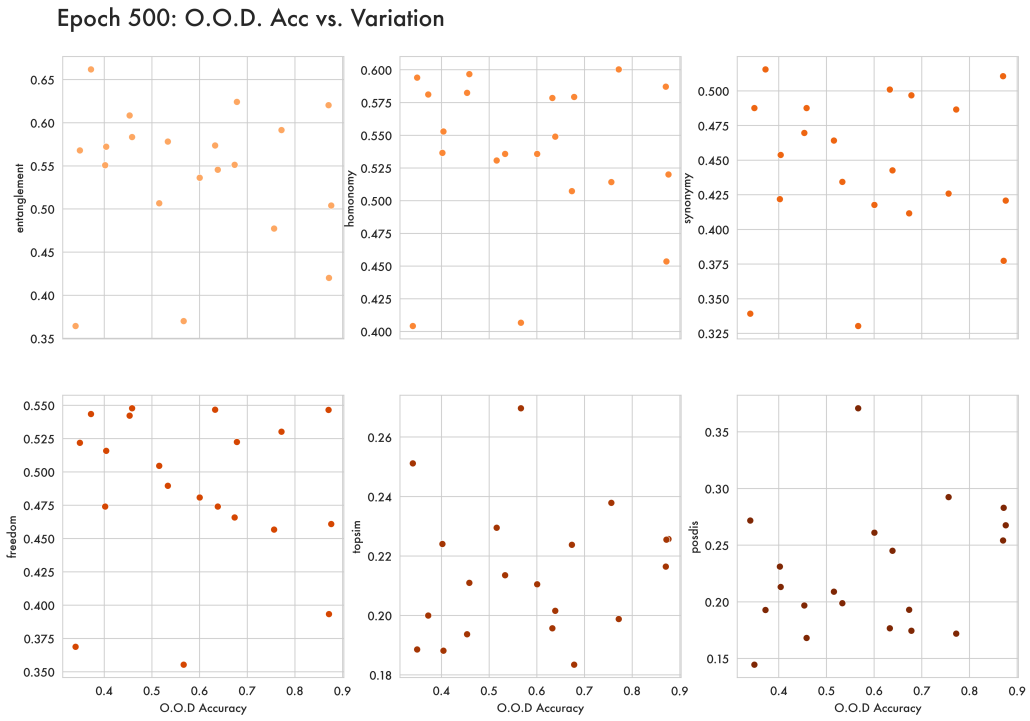

Figure 7: Plots show each of the variation measures plotted against o.o.d. accuracy at the 500th epoch of training.

A.5 Significance Testing for Variation Differences

params	synonymy	entanglement	freedom	homonomy	topsim	posdis
250 vs 500	0.0275	0.1049	0.0574	0.0629	0.2565	0.0993
250 vs 800	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
500 vs 800	0.0001	0.0000	0.0002	0.0001	0.0005	0.0001

Table 2: P Values obtained from a t-test comparing variation measures from different sized initialization. The difference between $large$ and $small$, and $large$ and $medium$ are significant. Of differences between $small$ and $medium$ only synonymy and posdis are significant