VARYING MEANING COMPLEXITY TO EXPLAIN AND MEASURE COMPOSITIONALITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Compositionality is assumed to be a key property of language, but it is hard to observe in language emergence simulations. We posit that a common characteristic can underlie the emergence of intersective adjectives and argument structure: the complexity of meaning of the datapoints that agents discuss must vary. We show that this characteristic of the task facilitates the study of the emergent languages. Our first experimental results are promising.

Natural language is often thought to be *compositional*. For such languages, "the meaning of a complex expression is determined by its structure and the meaning of its constituents" (Szabó, 2020). However, we do not fully understand the nature of compositionality, how it emerges and how it can be integrated as inductive biases in artificial agents. For Zipf (1949), language maximizes information transmission while minimizing the effort of the speakers and listeners. This optimisation process is constrained by factors internal to the agents (memory constraints, preferences for iconicity, etc.) as well as by external factors such as the distribution of the meanings, pragmatics or discourse factors (Kirby, 1999). We can tease apart these factors using computer simulations of language evolution and acquisition (Steels, 1997; Lazaridou & Baroni, 2020).

However, such simulations are often not very successful, unless they use unrealistic assumptions in which case they lose part of their explanatory value. For example, Kottur et al. (2017) restrict the vocabulary size of the speaker to be exactly the number of different values to denote, and also erases its memory after each utterance. Korbak et al. (2019) proposes that compositionality evolves from simpler communication protocols where the primitives ("blue", "circle") are learned in a separate training process, but the syntax of the agents is hardcoded through very specific training procedure involving several agents. The difficulty of quantifying compositionality is yet another hurdle to progress. When agents exchange variable-length utterances without an explicit model of grammar, it is not clear whether there are any meaningful subconstituents or not, and if there are any, how to segment them in order to measure the quantities of interest. Thus it is often implicitly assumed that each symbol bears meaning. However, this can lead to strange conclusions: if we assume that each letter in "cat" is meaningful, then English is not compositional.

To tackle these problems, we posit a general functional factor, the *meaning complexity variation*. Due to Zipf (1949)'s least-effort principle, we think that grammatical structures grow in complexity to accommodate for the variation in complexity or the varying structures of the meanings to convey. Therefore, for non-trivial grammars to emerge, the inputs of agents need to vary in complexity. We argue that such variations are present in everyday communicative situations. We propose two very different but realistic setups to study the emergence of intersective adjectives and argument structure. Crucially, in both tasks, the complexity of the inputs vary, and our preliminary results show promising forms of systematicity. In addition, we show that meaning complexity variation sidesteps the need to segment artificial messages into smaller constituents. We conclude with an overview of the opportunities afforded, outlining some concrete next steps.

1 TWO REALISTIC SETUPS EXHIBITING MEANING COMPLEXITY VARIATION

Compositionality is typically studied mathematically as follows (Montague, 1970; Szabó, 2020). Consider a meaning function m that maps expressions taken from a language $\mathcal{L}$ (the space of grammatically correct expressions) to a space of meanings $\mathcal{X}$ (all the things that the language can denote). A language is compositional iff m is an homomorphism, that is, if there are two binary operators, $\cdot$

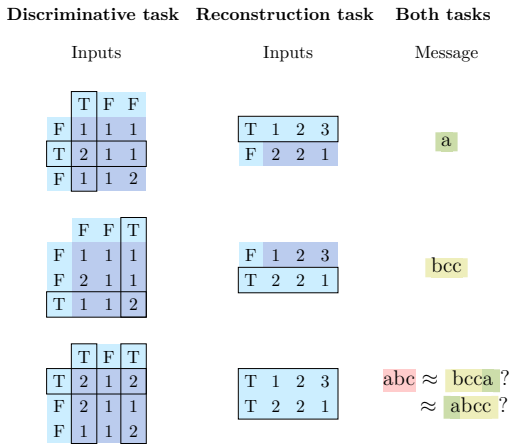

Figure 1: Simplified illustration of the two tasks. *Feature matrices* contain numbers from $\{1, 2, 3\}$. Rows: objects; columns: features. *Masks* contain booleans $\{\mathrm{T}, \mathrm{F}\}$, indicate to speaker what they should communicate about. Both speaker and listener observe information colored in dark blue, but only speaker observe cyan. Our metrics (cf. Section 2) compare utterances about simple meanings (two first rows) to utterance about complex objects (last row).

Entities	α	**A**	**B**
$-, 8, 1$	$0, 1, 0$	$24, 79, 25$	$105, 47$
	$0, 0, 1$	$105, 16, 105$	$34, 34$
	$0, 1, 1$	105 , $79, 24$	34 , 47
$-, 8, 5$	$0, 1, 0$	$24, 79, 25$	$105, 47$
	$0, 0, 1$	$19, 24$	$18, 18$
	$0, 1, 1$	47, $79, 24, 25$	18 , 24
$-, 8, 190$	$0, 1, 0$	$24, 79, 25$	$105, 47$
	$0, 0, 1$	$16, 19$	19
	$0, 1, 1$	16 , 79 , 39, 79	105 , 24
$-, 8, 39$	$0, 1, 0$	$24, 79, 25$	$105, 47$
	$0, 0, 1$	$16, 44, 16, 72, 2$	$16, 19$
	$0, 1, 1$	$44, 16$, 59, 72	105 , 16

Table 1: Messages from the second task. Entities: list of entities ("-": no entity; number indicate rank of entity in the dataset; position in the list indicate role: AGENT, PATIENT, MISC). α: mask. **A**, **B**: messages produced by speakers from two different runs. Symbols are manually colored to identify phrases (first 2 rows in every block of 3 rows) within artificial sentences (third row in every block). We see that some systematicity has emerged.

on $\mathcal{L}$, and $\times$ on $\mathcal{X}$, such that for any expression made of two constituents e_1 and e_2 in $\mathcal{L}$, we have

$$m(e_1 \cdot e_2) = m(e_1) \times m(e_2). \tag{1}$$

In language emergence simulations, $\mathcal{L}$ is the set of variable-length sequences of symbols produced by an agent (the speaker). In our case, symbols are sampled from a vocabulary $\{1, \ldots, n_V\}$ and the length of the messages are bounded. The operator $\cdot$ is the concatenation operator, so we write $e_1 \cdot e_2 = e_1 e_2$.

The meaning space $\mathcal{X}$ contains sets of filler-role pairs, with the constraint that each role appears at most in one pair. As we will see, the fillers and roles differ in the two tasks. Since we use sets, we can take $\times$ to be the union on sets.[1] We define *simple meanings* as sets from $\mathcal{X}$ containing a single pair, while *complex meanings* contain at least two pairs.

A task with **meaning complexity variation** contains both simple and complex meanings as inputs. In both tasks, the objective function is to maximize information transmission (implemented as a cross-entropy loss) while minimizing least-effort (implemented as a linear function of the number of symbols in messages, as done by Chaabouni et al. (2019)). The speakers and listeners are implemented using architectures based on Transformers (Vaswani et al., 2017).

1.1 THE EMERGENCE OF COMPOSITIONALITY IN INTERSECTIVE ADJECTIVES

The first task is a discrimination task. A speaker observes a set of objects from $\mathcal{X}$ and one of these is marked as a target, while the others are called distractors. The speaker produces a message that characterizes the target. The listener observes the same set of objects and must pick out the target using the message. Each object $x \in \mathcal{X}$ is a set of attribute-value pairs. For example, a blue circle can be described as $x_1 = \{(\text{COLOR}, \text{BLUE}), (\text{SHAPE}, \text{CIRCULAR})\}$.

Crucially, there are simple meanings, for which a single attribute-value pair is necessary to pick out the target, and there are complex meanings, for which two pairs are necessary (cf. first column in Figure 1). In other words, the meaning complexity varies.

[1] This is an oversimplification: we need to ensure that each role appears at most in one pair.

This task simulates the emergence of intersective adjectives. These adjectives are concatenated sequentially, as in "the *blue circular* thing".[2] There are several important parallel between frequent, real-life communicative situations and this task, but we need to focus on the realism of our innovation: the variable meaning complexity. The primary function of adjectives is probably to distinguish between *similar enough* objects. One does not use adjectives to discriminate between different kind of concrete things such as a car and a tree, as these things are more naturally associated to basic level terms, which probably have a lesser cognitive cost. Thus, in our task, 75% of the meanings are such that a single attribute-value pair is necessary to solve the task. By contrast, works from Lazaridou et al. (2017); Havrylov & Titov (2017) seem to sample objects independently. We also assume that the speaker knows the minimal set of attributes that distinguish the target from the distractors. This set of attributes is directly given to the speaker via a mask over features, as illustrated in Figure 1.[3] This allows the speaker, under a least-effort pressure, to communicate exclusively about those attributes.

1.2 The emergence of compositional argument structure

In the second task, meanings in $\mathcal{X}$ are *events* which contain entity-role pairs. A complete description is out of scope, but we can think of them as representations of clauses. For instance, one such event would be $e = \{(\text{ABOVE}, x_1), (\text{BELOW}, x_2), (\text{REL}, \text{ABOVENESS})\}$, where x_1 and x_2 can be entities the geometric shapes of the previous task. To simplify the experiments and the analysis of the results, the pair with the REL attribute is observed by both agents and needs not be conveyed.

Crucially, each relation takes a variable number of roles. *Simple* meanings require a single entity-role pair to be conveyed while *complex* complex meanings require several. Furthermore, the listener observes a partial view of the input of the speaker. This is illustrated in the second column of Figure 1: a meaning made up of two entities can be partially observed (first 2 rows) or completely unobserved by the listener (last row). When all but one entity is observed by the listener, the meaning is simple, since only the entity hidden to the listener must be communicated. Otherwise, the meaning is complex. Thus meaning complexity varies across inputs in this task as well.

This task simulates the emergence of argument structure. Indeed, in all natural languages, the way arguments are composed seems to be highly systematic. In simple sentences, arguments are simply concatenated to the verb. This seems to be true, regardless if roles are encoded via word order (analytic languages), case marking (synthetic languages) or verb affixes (polysynthetic languages).

The most important characteristic of this task is probably partial observability, and it corresponds to a fundamental aspect of the human experience. Most events that have happened in the past or will take place in the future are not directly witnessed and are only partially represented in our minds. Yet we often want to represent them more completely, which prompts one to ask questions, or spontaneously share information that is ignored by the interlocutor. For instance, upon seeing a broken window, one can ask "who broke the window?". A knowledgeable interlocutor could answer "John" or "John did". In our experiments, the speaker is this knowledgeable person, answering such questions about unobserved entities. Even if the agents do not engage in dialogues but in one-time interactions, the mask models an inference made by the speaker about the listener's knowledge, and abstracts away over all the possible reasons to share this information. Going back to our example, "John did" is an acceptable answer which does not contain the phrase "the window". This parsimony is encouraged by the least-effort penalty.

In Mordatch & Abbeel (2018)'s and Bogin et al. (2018)'s works, agents also need to express such events. However, the complexity of the meanings do not vary.

2 Side-stepping the need for segmentation

To verify if compositionality holds as formalized in equation 1, we need to access the constituents e_1 and e_2 that make up the complex expressions of the form $e_1 e_2$. In some works, this structure is known (Andreas, 2018), but it is more typically assumed that every symbol is meaningful (for example, in Bogin et al. (2018) and Chaabouni et al. (2020)'s works). However, neither options seem

[2] Our setup does not explain if and how they are ordered.

[3] Indeed, Transformers Vaswani et al. (2017) seem to be incapable of computing this information.

appropriate to analyze the messages exchanged by neural agents. This problem is also identified by Baroni (2020), but to our knowledge, we are the first to address it, or rather to side-step it.

Using tasks with meaning complexity variation, we can encode simple meanings $\{x_1\}$ and $\{x_2\}$ in isolation to obtain meaningful constituents e_1 and e_2. There is no particular problem, since these meanings are in the training data. Let $e_{1,2}$ be the message produced by the speaker to discuss $\{x_1, x_2\}$. To quantify compositionality, and in particular, systematicity, we would like to measure the similarity between $m(e_1e_2)$ or $m(e_2e_1)$ and $m(e_{1,2})$ (cf. third column in Figure 1). To do so, we compare the loss incurred by the listener when it interprets e_1e_2 versus $e_{1,2}$. Formally, we define the information transmission loss (reconstruction or discrimination) incurred when the complex meaning $\{x_1, x_2\}$ is conveyed via e as $l(e) = -\log p(\{x_1, x_2\}|e)$, where p is the conditional distribution of the listener. Then, the *listener's concatenability* is defined as

$$C^L = l(e_{1,2}) - \min(l(e_1e_2), l(e_2e_1)). \tag{2}$$

The $\min$ operator picks the lowest reconstruction loss among the two possible concatenation orders. Following (Lowe et al., 2019), we propose two metrics that capture the understanding of the two agents separately. We define $u(e) = \log q(e|\{x_1, x_2\})$, the log probability (according to the speaker's conditional distribution q) of an expression e when the complex meaning is observed. The *speaker's concatenability* is defined similarly as

$$C^S = \max(u(e_1e_2), u(e_2e_1)) - u(e_{1,2}). \tag{3}$$

These metrics are positive iff the concatenation e_1e_2 (or e_2e_1) is preferred to the actually sent message $e_{1,2}$. For the speaker, this means that the concatenation e_1e_2 (or e_2e_1) is more likely; for the listener, that e_1e_2 (or e_2e_1) allows it to recover the information better.

3 Discussion and conclusion

We can see samples obtained on the second task in Table 1, with additional details on how they are obtained in Appendix A. We also report some preliminary results on concatenability. A complete analysis is out of scope for this short paper, and secondary compared to the main arguments that we now summarize.

Meaning complexity variation is a novel, simple and promising functional explanation for the emergence of compositionality. We have argued that in conjunction with the least-effort principle, it could be a common cause behind the emergence of intersective adjectives and simple argument structure. This explanation is orthogonal to other characteristics of the tasks (reconstruction vs discriminative tasks) and could operate at different linguistic levels (phrase-level, sentence-level).

Our tasks could be merged within a single simulation. The meanings of the argument structure task are made of several entity-role pairs, but the entities themselves are meanings of the adjective task. By training agents to solve these two tasks at once, one expects adjectives to start appearing within arguments. More generally, this opens the door to simulations where various parts of speech emerge.

In addition to the functional explanation, we obtain a whole class of methods to analyze compositionality (and in particular, systematicity), as well as the ability to apply existing metrics such as those proposed by Chaabouni et al. (2020). Our concatenability metrics only scratch the surface. It is not a coincidence that meaning complexity variation both explains compositionality and helps us to study it. The easier languages are to analyze for scientists and machines, the easier they should be acquired by children as well, a reasoning that resonates with Gopnik (1996)'s theory.

However, this double role makes it harder to verify that our hypothesis is different from two alternative explanations. The first alternative explanation is that what matters is that inputs have the "right" meaning complexity. For instance, does compositionality emerge when all examples are complex, or when all examples are simple? We have already made progress to show that it is not the case. The second alternative explanation is that what matters is the variation in the *quantity* of information, not the structure or complexity of meanings. In our setup, the two go hand in hand. However, we do not think that varying the quantity of information is a sufficient condition. Indeed, if one uses a non-uniform input distribution over complex meanings only, the quantity of information varies, but the agents are not trained to convey the parts of meaning in isolation. We need to verify such claims as rigorously as possible in the future.

REFERENCES

Jacob Andreas. Measuring Compositionality in Representation Learning. In *International Conference on Learning Representations*, 2018.

Marco Baroni. Rat big, cat eaten! Ideas for a useful deep-agent protolanguage, 2020.

Ben Bogin, Mor Geva, and Jonathan Berant. Emergence of communication in an interactive world with consistent speakers. *arXiv preprint arXiv:1809.00549*, 2018.

Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. Anti-efficient encoding in emergent communication. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 6293–6303. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/31ca0ca71184bbdb3de7b20a51e88e90-Paper.pdf.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. Compositionality and Generalization In Emergent Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4427–4442, 2020.

Alison Gopnik. The Scientist as Child. *Philosophy of Science*, 63(4):485–514, 1996. doi: 10.1086/289970. URL https://doi.org/10.1086/289970.

Serhii Havrylov and Ivan Titov. Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 2149–2159, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/70222949cc0db89ab32c9969754d4758-Abstract.html.

Simon Kirby. *Function, selection, and innateness: The emergence of language universals*. OUP Oxford, 1999.

Tomasz Korbak, Julian Zubek, Lukasz Kucinski, Piotr Milos, and Joanna Raczaszek-Leonardi. Developmentally motivated emergence of compositional communication via template transfer. *CoRR*, abs/1910.06079, 2019. URL http://arxiv.org/abs/1910.06079.

Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2962–2967, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1321. URL https://aclanthology.org/D17-1321.

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

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-Agent Cooperation and the Emergence of (Natural) Language. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Hk8N3Sclg.

Ryan Lowe, Jakob N. Foerster, Y.-Lan Boureau, Joelle Pineau, and Yann N. Dauphin. On the Pitfalls of Measuring Emergent Communication. In Edith Elkind, Manuela Veloso, Noa Agmon, and Matthew E. Taylor (eds.), *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, pp. 693–701. International Foundation for Autonomous Agents and Multiagent Systems, 2019. URL http://dl.acm.org/citation.cfm?id=3331757.

Richard Montague. Universal grammar. *Theoria*, 36(3):373–398, 1970. doi: https://doi.org/10.1111/j.1755-2567.1970.tb00434.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1755-2567.1970.tb00434.x.

Igor Mordatch and Pieter Abbeel. Emergence of Grounded Compositional Language in Multi-Agent Populations. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 1495–1502. AAAI Press, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17007.

Luc Steels. The synthetic modeling of language origins. *Evolution of communication*, 1(1):1–34, 1997.

Zoltán Gendler Szabó. Compositionality. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2020 edition, 2020. URL https://plato.stanford.edu/archives/fall2020/entries/compositionality/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

George Kingsley Zipf. Human behavior and the principle of least effort. 1949.

A Results

The results presented here will be of limited interest, since we have neither described the architectures of the agents used, nor any optimisation detail, nor even details about the implementation of the least-effort penalty. We hope to publish these details in a longer format soon. However, it might be particularly useful to show some examples of messages to demonstrate the intuition behind the variation of meaning complexity.

A.1 Qualitative analysis

Meaning	**A**	**B**
$1, -, -$	64	7
$-, 3, -$	104	113
$1, 3, -$	64 , 104	6, 7
$1, -, -$	64	7
$-, -, 4$	112	27
$1, -, 4$	64 , 112	7 , 27
$-, 3, -$	104	113
$-, -, 1$	66	7
$-, 3, 1$	66 , 104	27, 113
$4, -, -$	58	
$-, -, 4$	112	27, 113
$4, -, 4$	97, 112	113

Table 2: Messages obtained using the first task. Meaning: sequence, where position denotes the attribute and number: value. **A**, **B**: messages produced by speakers from two different runs.

In order to obtain the messages shown in Table 1, we took several meanings where the same entity (the 8th) is in the same role (in our case, 4 different meanings). By varying the mask of the speaker, we have obtained different messages that denote entities in isolation or together. We have not yet done the exact same procedure with the first task, but we have done something comparable. After having selected two models, we have encoded all utterances from the training set and computed the

most frequent message for each meaning (complex and simple). We report 4 complex meanings as well as their most frequent messages, and similarly for the simple meanings that they are made of. Instead of presenting the feature matrices and the mask, we directly present the meaning to convey as the masked feature matrix. For example, the sequence $1, -, 2$ denotes $\{(1, 1), (3, 2)\}$.

We see that messages are obtained rather systematically, but not completely. However, we need to keep in mind that this analysis is very crude: there might be other likely messages that are more systematic, but that our greedy decoding process does not allow us to find. Quantitative methods such as concatenability take into account the probability distributions of the speakers and listeners and are thus much more fine-grained.

A.2 Concatenability

We will limit our analysis to the discriminative experiments. We have compute concatenability metrics C^S and C^L on the training set, averaging over examples. It is only computed on examples where e_1 and e_2 are not empty messages. In the discrimination task, 75% of the datapoints are simple whereas the rest is complex.

Perhaps surprisingly, C^L is on average equal very close to 0 (-0.033, over 30 runs), whereas the average reconstruction loss is 0.59. There are runs above 0, that is, for which $e_1 e_2$ yields a better loss than the actually sent message computed by the speaker using the complex meanings. On the other hand, from the speaker's point of view, the average C^S is -10.7, meaning that the concatenations of the simple expressions are much less likely than the messages actually sent by the speaker. Thus even if concatenations are on average rather well interpreted by the listener, they are not likely according to the speaker.