

# Discovering Properties of Inflectional Morphology in Neural Emergent Communication

Miles Gilberti

Shane Storks

Huteng Dai

University of Michigan

{milgil, sstorks, huteng}@umich.edu

## Abstract

Emergent communication (EmCom) with deep neural network-based agents promises to yield insights into the nature of human language, but remains focused primarily on a few subfield-specific goals and metrics that prioritize communication schemes which represent attributes with unique characters one-to-one and compose them syntactically. We thus reinterpret a common EmCom setting, the attribute-value reconstruction game, by imposing a small-vocabulary constraint to simulate double articulation, and formulating a novel setting analogous to naturalistic inflectional morphology (enabling meaningful comparison to natural language communication schemes). We develop new metrics and explore variations of this game motivated by real properties of inflectional morphology: concatenativity and fusion. Through our experiments, we discover that simulated phonological constraints encourage concatenative morphology, and emergent languages replicate the tendency of natural languages to fuse grammatical attributes.

## 1 Introduction

Emergent communication (EmCom) studies communication protocols developed between two or more deep neural agents engaged in a task requiring successful communication. Analyses of learned protocols in EmCom typically search for characteristics deemed crucial in natural language, prominently compositionality (Chaabouni et al., 2020). Such experiments promise to shed light on the pressures underlying human-like language evolution.

However, as interpreting emergent communication schemes and measuring compositionality are challenging, inhibitive assumptions are often made. First, most EmCom works afford agents an over-complete vocabulary of characters, making it possible for agents to describe each possible attribute value with a unique character one-to-one,

thus leading to treatment of generated characters exclusively as words composed only syntactically. In turn, common evaluation metrics developed for this large-vocabulary setting are overly simplified, favoring protocols that represent attribute values with consistent characters concatenated in a consistent ordering (Chaabouni et al., 2020; Resnick et al., 2020; Ueda et al., 2023). This disregards the possibility of many morphological phenomena attested in natural languages, such as nonconcatenative morphology (e.g., root-pattern morphology common in Semitic languages), and fusion of grammatical attributes in morphemes (e.g., the Spanish verb suffix *-amos* simultaneously communicates first person, plurality, and present tense).

In this work, we reinterpret EmCom as *emergent morphology*, addressing these shortcomings with two key design choices:

1. We target humanlike *double articulation* (Hockett, 1960) by enforcing a small character inventory, necessitating that individually meaningless characters form combinatorial morpheme-like symbols to encode attribute values and combinations thereof.
2. Shown in Figure 1, we introduce a novel configuration of attribute-value reconstruction inspired by *inflectional morphology*, where a high-cardinality attribute (analogous to roots) is paired with comparatively lower-cardinality attributes (analogous to grammatical information, e.g., tense and person), encouraging richer morphological phenomena.

Because agents learn to solve a problem directly analogous to information communicated by natural languages, we can thus make deeper and more direct comparisons between emerged language and natural language than in prior work. Toward this goal, we review EmCom evaluation metrics for various properties of communication schemes, proposing novel metrics for the underexplored properties of concatenativity and degree of fusion. To validate

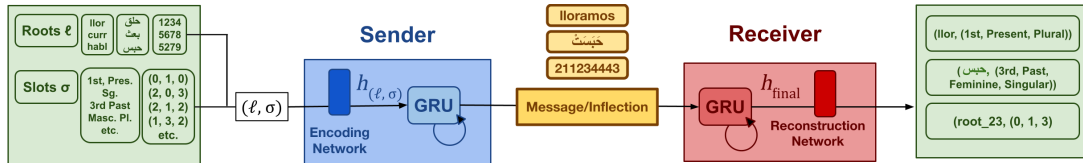


Figure 1: We reframe the typical attribute-value (Attr-Val) reconstruction game with an analogy to inflectional morphology, where roots  $\ell$  and slots  $\sigma$  (e.g., tense and person) comprise attributes which must be communicated through messages. Emergent communication schemes are then comparable to natural language inflections.

and establish expected ranges for these metrics, we analyze a representative group of artificial double articulatory communication schemes and inflection-based natural language communication schemes.

With these new paradigms and metrics established, we conduct some preliminary investigations, the results of which show how this approach can rediscover morphological phenomena in human language, as well as test new hypotheses about its evolution. Specifically, we:

- Implement a phonology-inspired constraint in communication agents, using concatenativity metrics to reveal ease of articulation as a pressure toward concatenative languages.
- Compare the topographic similarity of emergent and natural languages, finding significant room for improvement in emergent languages to reflect natural degrees of compositionality.
- Analyze fusion in emergent and natural languages for inflection, showing that emergent languages mostly fuse grammatical attributes rather than roots, mirroring natural language.

This work opens exciting new directions for future research into the evolution of morphology.

## 2 Related Work

Some other EmCom works address double articulation explicitly via selection of small inventory sizes (Ueda and Washio, 2021; Ueda et al., 2023; Ueda and Taniguchi, 2024), but the analysis focuses on words rather than morphemes, and only considers perfectly concatenative methods of combining them. Work utilizing bitstring messages in EmCom achieves this (Resnick et al., 2020; Gupta et al., 2020), but not in a linguistically realistic way.

Unlike our work, many works use large-vocabulary EmCom to investigate various phenomena relevant to natural language morphology. For example, Chaabouni et al. (2019) use a length pressure to rederive Zipf’s Law of Abbreviation. Lian et al. (2023, 2024) explore the case-marking versus word-order tradeoff in pair and group communi-

cation. Conklin and Smith (2023) discover naturalistic variability in emergent languages through proposed metrics for synonymy, homonymy, word-order freedom, and entanglement, complicating typical measures of compositionality. Galke et al. (2022); Galke and Raviv (2024) explore the importance of learnability, generalization, production effort, and group size pressures, among other psychosociolinguistic factors. Gualdoni et al. (2024) and Tucker et al. (2022) explore information theoretic constraints on the forms of emergent languages. Chaabouni et al. (2022) further explore the relationships between these aspects in a naturalistic visual task setting, while Kouwenhoven et al. (2024) investigate the effect of agents’ representational alignment on their performance.

Lastly, there is vast literature studying morphology in human languages. Notably, we adopt the term “emergent morphology” introduced by Archangeli and Pulleyblank (2016) within their *Emergent Grammar* framework, which hypothesizes how morphophonological structures arise from constraint interactions in natural languages. Especially related prior works attempt to model and analyze the emergence of various aspects of morphological systems (Nowak and Krakauer, 1999; Zuidema and de Boer, 2009; Elsner et al., 2020; Dekker, 2024). Unlike these works, we emphasize computational modeling of *artificial communication systems*, focusing on how minimal meaningless discrete units combine systematically into meaningful morphological structures like inflection. Automated morphological segmentation, especially for nonconcatenative morphology (Fullwood, 2018), is also relevant. While we only explored concatenative approaches, we expect such techniques to play a crucial role in future work toward understanding morphology in EmCom.

## 3 Toward Emergent Morphology

Prior EmCom work has largely focused on the Lewis Signaling Game (Kottur et al., 2017;

Chaabouni et al., 2020; Ueda et al., 2023; Lewis, 1979), in which a sender agent observes some state and generates a message to a receiver agent, which must take an action based on it. Success in the game relies on accurately communicating the relevant features of the environment. Morphology has been relatively understudied in EmCom due to a focus on individual characters representing atomic units of meaning (Boldt and Mortensen, 2024b).

### 3.1 Attribute-Value Reconstruction Game

The Attribute-Value Reconstruction Game (henceforth **Attr-Val**) has been extensively studied by EmCom researchers (Kottur et al., 2017; Chaabouni et al., 2020). This class of games models the state as an  $n$ -tuple of values where each value belongs to some attribute  $A_i \in \mathcal{A}$ , defined as a set of possible values for the attribute. That is, the set of states is  $S = \{s | s = (v_1 \in A_1, v_2 \in A_2, \dots, v_n \in A_n)\}$ .

Formally, Attr-Val can be defined as such: we have a state  $s \in S$ , a sender  $f_\theta : S \rightarrow M$ , and a receiver  $g_\theta : M \rightarrow S$ .  $M$  is the space of messages of up to length  $m$  characters, all of which must belong to a fixed character vocabulary  $C$ . This instance is considered successful if  $g_\theta(f_\theta(s)) = s$ .

### 3.2 Attr-Val as Inflectional Morphology

The attribute value sets  $A \in \mathcal{A}$  tend to have the same cardinality in prior work (Kottur et al., 2017; Chaabouni et al., 2020; Ueda et al., 2023), often motivating analogies to communicating properties of objects with few possible values, e.g., a red square or blue triangle. We claim that Attr-Val games also represent a rich analogy for inflectional morphology. The meaning of an inflected root in natural language consists of the root as well as additional attributes like tense and person for verbs, or number and gender for nouns. Like Attr-Val, these attributes take on a finite set of values. While the number of possible roots is unbounded in natural language, Attr-Val can simulate this through one attribute set (representing roots) being much larger than others. This paradigm provides a key advantage: natural languages solve an analogous and similarly challenging problem to this, enabling direct comparison with learned communication schemes.

We consider 4 configurations for  $\mathcal{A}$ : a **default** setting resembling prior work with 3 attributes of 16 values each ( $16 \times 16 \times 16$ ), and 3 novel **inflection** settings with nonuniform attribute value sets ( $42 \times 2 \times 3$ ,  $42 \times 2 \times 2 \times 2$ , and  $42 \times 6$ ). In inflection settings, the largest attribute set symbolizes roots,

while others symbolize grammatical properties like tense and person. Settings have a comparable total number of attribute values (47-48).<sup>1</sup> To necessitate double articulation, i.e., composing meaningless characters into meaningful symbols to represent attributes, we enforce a consistent character vocabulary size  $|C| = 8$  and message length  $m = 9$ , smaller than the size of most attribute sets. We assume virtually all natural languages have significantly more roots than properties that can be inflected for due to the inherently limited nature of such grammatical information, so this design reflects the actual structure of natural languages. Importantly, one attribute in the inflection settings has many more possible values than  $|C|$  while others have fewer, possibly incentivizing more morphologically diverse communication schemes.

### 3.3 Agent Implementation

Following prior works, we use EGG (Kharitonov et al., 2019) to implement sender and receiver agents as single layer gated recurrent unit (GRU; Cho et al., 2014) models,<sup>2</sup> with hidden dimensions of 500. Senders were optimized using REINFORCE (Williams, 1992) and receivers with back-propagation following Chaabouni et al. (2019) using the accuracy (i.e., proportion of attributes correctly reconstructed) as a loss, with cross-entropy loss as a supplementary loss term.<sup>3</sup> Additional implementation details are provided in Appendix B.

## 4 Interpreting Communication Schemes

To investigate double articulation in EmCom, we next introduce methods to evaluate communication schemes by extracting meaningful symbols from messages, and measuring various properties of them. We then establish reference points for emergent communication schemes based on both idealized artificial languages and natural languages.

### 4.1 Evaluating Communication Schemes

The need for stronger evaluations in EmCom is well-established (Boldt and Mortensen, 2024b;

<sup>1</sup>We also ran experiments controlling for total number of combinations, but models did not converge (see Appendix A).

<sup>2</sup>We also attempted to implement agents with transformers (Vaswani et al., 2017). However, even the best hyperparameter settings failed to converge to the degree of accuracy found in the GRU agents, even when given significantly more time (24 hours vs. 8). Due to the very low accuracy, it is impossible to make meaningful conclusions, so these results are omitted.

<sup>3</sup>This supplementary loss allowed smoother and more reliable convergence with lower sequence lengths, desirable for obtaining interpretable and minimally complex languages.

Chaabouni et al., 2020), and especially prominent in our problem setting, which aims to consider a greater number of plausible communication schemes than prior work. As such, we next introduce and appraise several algorithms for segmenting meaningful symbols from sequences of characters, as well as evaluation metrics for sequences of symbols in communication schemes.

#### 4.1.1 Symbol Segmentation Algorithms

To extract meaningful, morpheme-like symbols from messages (essential for double articulation), we consider two approaches: one from prior work based on Harris’ articulation scheme (Ueda et al., 2023), and a previously unexplored approach based on byte-pair encoding (Gage, 1994). We define the resulting segmented symbol vocabulary  $V$  as the set union of symbols extracted from each message.

**Harris’ Articulation Scheme** Harris’ articulation scheme (HAS; Harris, 1955; Tanaka-Ishii, 2021) has been proposed as a desirable, naturalistic trait for emergent languages to exhibit (Ueda et al., 2023; Ueda and Taniguchi, 2024). HAS captures the tendency of word boundaries to be identifiable purely by character-level entropy. Specifically, the scheme claims that while entropy generally decreases across an utterance, it will increase at morpheme boundaries. This implies an algorithm for segmenting morphemes by drawing boundaries at indices  $i$  where  $\mathcal{H}_i - \mathcal{H}_{i+1} > \tau$ , where  $\tau$  is a threshold that determines how sensitive the algorithm is. In all of our analyses,  $\tau$  is set to 0 to maximize sensitivity. We use the implementation provided by Ueda et al. (2023).

**Byte-Pair Encoding** HAS is a powerful tool for identifying symbols in communication schemes, but it is difficult to control the degree of segmentation by it. As such, we propose an alternative approach based on the Byte-Pair Encoding (BPE) algorithm (Gage, 1994), which has been widely used in tokenizers for NLP systems (Sennrich et al., 2016). Importantly for our purposes, the properties of the subwords generated by BPE have been shown to enable characterization of languages by established typological categories, such as analytic and synthetic (Gutierrez-Vasques et al., 2023). BPE is configurable with a maximum vocabulary size  $|V|$ . Early compressions tend to be most informative (Gutierrez-Vasques et al., 2021, 2023), so we apply BPE with a merging cutoff of 96, approximately twice the expected number of symbols (i.e.,

the total number of unique attribute values), and additionally apply BPE with maximum compression.

#### 4.1.2 Evaluation Metrics

Given a segmented symbol vocabulary  $V$ , we next discuss common metrics for its compositionality from prior work. We then introduce methods to evaluate concatenativity and degree of fusion, more specific forms of compositionality that typical compositionality metrics do not directly measure, as well as learnability, which has previously been connected to compositionality (Kirby et al., 2014; Li and Bowling, 2019; Chaabouni et al., 2020).

**Compositionality** Compositionality is the core objective for many EmCom works, motivated by interpretability and its prominence in natural languages. Topographic similarity (**TopSim**) is a common measure for this (Brighton and Kirby, 2006; Lazaridou et al., 2018), formally defined as the correlation of the distance (typically Levenshtein distance; Levenshtein, 1965) between pairs of messages and the distance between their meanings. This provides a global measure of how similar the symbols in messages for similar configurations.

However, TopSim places minimal restrictions on how similarity of messages is calculated. As such, Chaabouni et al. (2020) propose bag-of-symbols disentanglement (**BoSDis**), an alternative metric accounting for permutation-invariant communication schemes<sup>4</sup> (e.g., “A and B” versus “B and A”):

$$\frac{1}{|V|} \sum_{v \in V} \frac{\mathcal{I}(n_v; a_v) - \mathcal{I}(n_v; b_v)}{\mathcal{H}(n_v)} \quad (1)$$

BoSDis calculates the mean entropy-normalized difference in mutual information  $\mathcal{I}$  between the count  $n_v$  of a vocabulary symbol  $v \in V$  and attributes  $a_v$  and  $b_v$ , which have the 2 highest  $\mathcal{I}$  values with  $v$ . This measures whether a character’s presence in a message uniquely communicates a specific attribute value. This should generally not happen in our small-vocabulary setting, which necessitates reuse of characters in multi-character symbols to induce double articulation,<sup>5</sup> thus BoSDis should be low. However, if an algorithm effectively segments symbols which uniquely refer to attributes, the BoSDis with respect to those symbols should be higher. We thus gauge whether HAS

<sup>4</sup>They also propose positional disentanglement, which we omit but provide more details about in Appendix D.

<sup>5</sup>One-character symbols are an exception, but an accurate communication scheme has very limited capacity for these.

and BPE extract meaningful symbols through a ratio  $\text{BoSDis}_C^V$  of BoSDis for a segmented symbol vocabulary  $V$  and unsegmented characters  $C$ . If  $\text{BoSDis}_C^V > 1$ , then  $V$  creates a more disentangled representation than individual characters  $C$ , thus  $V$  contains meaningful multi-character symbols.

**Concatenativity** TopSim and BoSDis fail to account for non-trivial and variable forms of compositionality, including some present in natural language (Korbak et al., 2020; Conklin and Smith, 2023). For example, TopSim based on Levenshtein distance of characters reflects an assumption that characters should not be modified or reordered under composition. Contrarily, we aim to consider any scheme where messages are formed via a systematic composition of symbols (i.e., as opposed to a memorized “hash function”) as compositional.

One particular form of compositionality we aim to tease apart from others is *concatenativity*, i.e., the degree to which composable symbols consist of uninterrupted streams of characters (analogous to phonemes) that are concatenated under composition.<sup>6</sup> Using English verb conjugation as an example, “played” is a concatenative message consisting of “play” and “ed.” By contrast, “sang” is a non-concatenative message since the characters “s\_ng” correspond to the root “sing,” but are interrupted by “a,” which communicates tense.

As HAS and BPE require symbols to be composed of consecutive characters, we estimate concatenativity as the average number of symbols in segmented messages of a communication scheme. We refer to these length metrics as **HASLen** and **BPELen**<sub>|V|</sub>.<sup>7</sup> A smaller number of symbols indicates greater concatenativity, as more consecutive characters belong in self-contained units of meaning. Of course, not all compositional communication schemes are concatenative. In Section 5.1, we will apply concatenativity metrics to contrived concatenative and nonconcatenative compositional schemes to analyze how these properties relate.

**Degree of Fusion** *Fusion* is the expression of multiple attributes of meaning within one morpheme or symbol. Prior work examines this by taking conditional probabilities of surface forms

<sup>6</sup>This is closely related to *integrity* in optimality theory, which prohibits output forms from containing multiple correspondents to an input form (McCarthy and Prince, 1995).

<sup>7</sup>For HASLen, we use the number of boundaries. Since HAS relies on entropy, we note that HASLen may be artificially low in communication schemes that include some random or otherwise non-meaningful characters.

after ablating feature pairs from data (Rathi et al., 2021, 2022), but this requires computationally expensive training of models for each feature pair. The fusion of two attributes  $A_1, A_2$  is equivalent to treating them as one attribute  $A_{12} = A_1 \times A_2$ , for which symbols may be chosen arbitrarily, and a language may be compositional with respect to a particular fused attribute set. As such, we define a novel metric of fused TopSim (**F-TopSim**), calculated by the maximum TopSim over the fusions of all attribute pairs. A language is fusional if  $\text{F-TopSim} \geq \text{TopSim}$ , so we summarize the degree of fusion with F-TopSim  $\delta = \text{F-TopSim} - \text{TopSim}$ .

**Learnability** To explore its relationship with concatenativity and fusion, we measure learnability as the mean number of training **epochs** that a receiver takes to exceed 99% reconstruction accuracy.

## 5 Artificial and Natural Reference Points

Next, we curate reference compositional communication schemes. To evaluate how metrics capture the above properties and establish expected ranges, we first propose artificial languages to target them. To contextualize EmCom results with real-world communication schemes, we develop additional languages based on natural language inflections.

### 5.1 Artificial Languages

As shown in Figure 2, we generate 5 artificial languages with a range of morphological complexity using message length  $m = 8$  and vocabulary size  $|C| = 9$ . The first 4 types compose symbols of constant and mixed length with varied levels of concatenativity. These languages preserve symbols and the ordering of their characters under composition, favored by TopSim. The fusion language violates this by assigning symbols to combinations of attribute values rather than individual ones. Tables 1-2 report metrics for receivers trained on these languages. We list key observations below.<sup>8</sup>

First, the degree of concatenation has minimal effect on learnability, implying a lack of inductive bias toward concatenation in recurrent receiver networks.  $\text{BoSDis}_C^V$  metrics confirm that HAS and BPE extract more meaningful morphemes from more concatenative languages, despite all languages being compositional. HASLen and BPELen are lower in more concatenative languages, suggesting these metrics effectively reflect concatenativity. Mixed concatenative languages have

<sup>8</sup>More discussion and artificial languages in Appendix E.

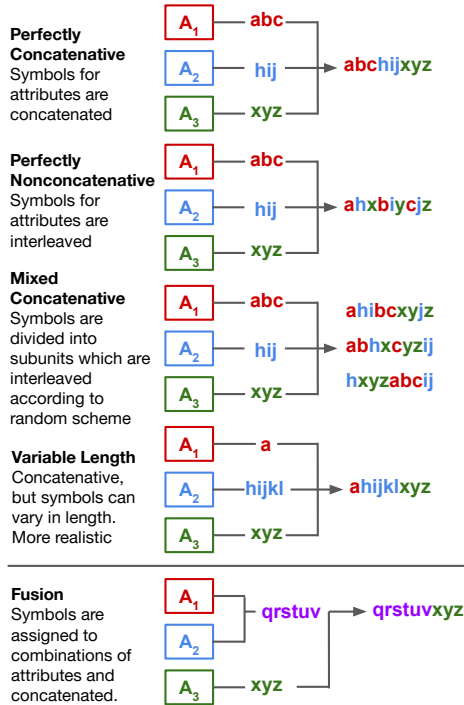


Figure 2: Artificial languages. Each language defines a unique symbol for each possible attribute value, but varies the composition operation for symbols.

higher variance in BPELen, falling along a spectrum between perfectly concatenative and nonconcatenative languages, and random languages have markedly higher asymptotic BPELen than others, further evidence of BPE’s effectiveness. As expected, HAS is less effective for languages with randomness, evidenced by low HAS  $\text{BoSDis}_C^V$  and HASLen in mixed concatenative and variable length languages, with the latter also low in random languages.<sup>9</sup>  $\text{BoSDis}_C^V$  metrics are large for random languages, suggesting an over-sensitivity to noise in the absence of other signal. Lastly, F-TopSim  $\delta$  effectively reflects fusion in the fusion languages.

## 5.2 Natural Languages

As a comparison to humanlike communication schemes, we also analyze subsets of the verb conjugations of two natural languages, Spanish and Arabic, as Attr-Val languages. We chose these due to the availability of high-quality datasets, the fact that both conjugate for tense and person, and the fact that they represent fairly divergent communication schemes along the concatenativity axis. Spanish data comes from Jehle (2011), and Arabic data from Nawaz et al. (2025). We select a set of verb attributes comparable to the inflection

<sup>9</sup>Supplementary results with HASLen in Appendix F.

Attr-Val setting: root, tense, and person. We then generate Attr-Val configurations based on combinations of these verb attributes sampled from the data, which serve as cognitively plausible (albeit simplified) representations of actual cognitive objects and events encoded and reconstructed in natural communication. For other attributes beyond tense and person, we choose a constant value (e.g., mood is always indicative). We generate 50 such inflectional sublanguages for Spanish and Arabic, each of which are mappings between lexeme-slot tuples  $(\ell, \sigma)$  and surface forms  $w$  (Wu et al., 2019).

Up to 76% of generated Spanish sublanguages can be meaningfully segmented by either HAS or BPE (i.e.,  $\text{BoSDis}_C^V > 1$ ), while only up to 8% of the Arabic sublanguages can be<sup>10</sup>, demonstrating how both tools for symbol segmentation are insufficient for nonconcatenative morphology. Other metrics for these sublanguages are discussed in Section 6.

## 6 EmCom Experiments

Our Attr-Val settings targeting double articulation and inflection, general and specific evaluation metrics for communication schemes, and illustrative artificial and natural language reference points enable us to ask previously unexplored research questions around morphology in EmCom. To scratch the surface of this space, we ask:

1. *Do more concatenative languages emerge under a human-inspired evolutionary pressure for ease of articulation?*
2. *How do topographic similarity in emergent and natural languages for inflection compare?*
3. *Do emergent languages exhibit fusion in inflection similarly to natural languages?*

We next introduce experimental details, including the above articulation pressure. Guided by these questions, we then present the results of this initial exploration as a proof of concept.

### 6.1 Experimental Design

Prior EmCom experiments have often explored pressures inspired by human evolution, such as ease of learning (Kirby et al., 2014; Li and Bowling, 2019) among others (Vithanage et al., 2023; Galke and Raviv, 2024), to encourage compositionality in emergent communication schemes. Meanwhile, prior work often assumes or favors concatenativity in evaluation (Resnick et al., 2020; Ueda et al., 2023) despite the possibility of compositional but

<sup>10</sup>Detailed analysis in Appendix G.

Language	TopSim	BoSDis	HAS BoSDis <sub>C</sub> <sup>V</sup>	BPE <sub>96</sub> BoSDis <sub>C</sub> <sup>V</sup>	BPE <sub>Max</sub> BoSDis <sub>C</sub> <sup>V</sup>	F-TopSim $\delta$	Epochs
Perf. Conc.	.628 $\pm$ .01	.076 $\pm$ .03	4.889 $\pm$ 4.47	7.836 $\pm$ 3.72	.751 $\pm$ .34	-.236 $\pm$ .01	4.875 $\pm$ .83
Mixed Conc.	.621 $\pm$ .00	.076 $\pm$ .03	2.146 $\pm$ .70	1.829 $\pm$ .84	.430 $\pm$ .23	-.236 $\pm$ .01	3.875 $\pm$ .64
Nonconcat.	.624 $\pm$ .01	.076 $\pm$ .03	.509 $\pm$ .18	.825 $\pm$ .43	.379 $\pm$ .15	-.237 $\pm$ .01	4.875 $\pm$ .99
Var. Length	.451 $\pm$ .03	.085 $\pm$ .03	1.965 $\pm$ 1.01	5.110 $\pm$ 1.77	.796 $\pm$ .28	-.131 $\pm$ .02	10.00 $\pm$ 3.85
Fusion	.238 $\pm$ .03	.126 $\pm$ .01	9.685 $\pm$ .08	4.474 $\pm$ .15	.402 $\pm$ .03	.175 $\pm$ .01	29.13 $\pm$ 3.09
Random	.000 $\pm$ .00	.001 $\pm$ .00	3.437 $\pm$ 1.14	8.263 $\pm$ 2.71	19.51 $\pm$ 7.32	.000 $\pm$ .00	100+

Table 1: Evaluation of compositionality, degree of fusion, and learnability of artificial languages, averaged over 8 random seeds.  $1\sigma$  confidence intervals provided to within 2 decimal points, or whole numbers for larger values.

nonconcatenative communication schemes (e.g., our artificial and natural reference languages).

In order to achieve concatenative emergent languages with double articulation, we explore an *ease of articulation* pressure. Human spoken languages are universally subject to phonological rules which determine the possibility of vocal sounds based on the environment and vocal tract physiology (Dediu et al., 2017). The interface between phonology and morphology is a key area of study in linguistics, and a computational evolutionary model of morphology (as we hope to lay the groundwork for) would be incomplete without considering that morphemes are made up of meaningless components subject to similar rules independent of their meanings. Specifically, there are disallowed clusters of sounds in spoken languages (e.g., Hawaiian disallows consonant clusters), which may represent a significant pressure towards concatenative languages. To test this, we added a term to the training loss for a toy simulated phonology for agents, where even-valued characters can not occur next to even-valued characters, and odd-valued characters can not occur next to odd-valued characters. Formally, for a message  $s$  comprised of a sequence of integer tokens, the loss term for ease of articulation pressure is computed as:

$$a(s_i, s_j) = \begin{cases} 1 & s_i \equiv s_j \pmod{2} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{L}_{\text{articulation}} = \epsilon \sum_{i=1}^{|s|-1} a(s_i, s_{i+1}) \quad (2)$$

$\epsilon$  is a hyperparameter which controls how strict the phonological rule is. We tried 1, 10, and 100 as values and found that 10 resulted in fairly strict adherence without incurring optimization problems, so  $\epsilon = 10$  in our reported experiment.

In all experiments, agents are trained with 8 different random initializations. Emerged languages are considered successful if the agent achieves 90%

Language	HASLen	BPELen <sub>96</sub>	BPELen <sub>Max</sub>
Perf. Conc.	2.758 $\pm$ .22	3.185 $\pm$ .06	2.001 $\pm$ .00
Mixed Conc.	3.791 $\pm$ .57	4.489 $\pm$ .20	2.178 $\pm$ .15
Nonconcat.	4.558 $\pm$ .32	4.722 $\pm$ .07	2.294 $\pm$ .05
Var. Length	2.463 $\pm$ .30	3.078 $\pm$ .18	1.995 $\pm$ .00
Fusion	3.502 $\pm$ .21	4.341 $\pm$ .02	2.004 $\pm$ .01
Random	1.593 $\pm$ .19	4.997 $\pm$ .01	3.105 $\pm$ .01

Table 2: Artificial language concatenativity metrics averaged over 8 random seeds, and  $1\sigma$  confidence intervals.

accuracy on the task, which occurs in most cases. In both the default and inflection settings, all 8 languages were successful, with the exception of 1 language in the  $42 \times 2 \times 3$  inflection setting with the above articulation pressure.

## 6.2 Experimental Results

In Table 3, we summarize concatenativity and degree of fusion metrics in each Attr-Val setting, without and with articulation pressure. As shown there,  $\text{BPE}_{96} \text{BoSDis}_C^V > 1$  on average across all emerged languages, implying the existence of multi-character meaningful segments and confirming that double articulation is achieved.

### 6.2.1 Concatenativity and Articulation

We visualize the concatenativity (measured by BPELen) of emergent communication schemes in the default and inflection Attr-Val settings respectively in Figures 3 and 4. First, we observe no inductive bias toward concatenativity in our RNN-based agents as previous works implicitly assume; instead, emergent languages fall along a spectrum between perfectly concatenative and nonconcatenative. Table 4 demonstrates a particularly obvious nonconcatenative symbol from an emergent language, where repetitions of e seem to indicate 2 as the second attribute value. While manual inspection does not prove that such patterns play a causal role in the receiver’s selection, we find this example illustrative of the fact that EmCom agents can and

Setting	Art.	TopSim	BoSDis	HAS BoSDis <sub>C</sub> <sup>V</sup>	BPE <sub>96</sub>	BoSDis <sub>C</sub> <sup>V</sup>	BPELen <sub>96</sub>	F-TopSim $\delta$
16 × 16 × 16	✗	.339±.034	.213±.047	3.25±4.31	1.25±.359		3.82±.235	-.0418±.0210
16 × 16 × 16	✓	.330±.028	.196±.076	5.98±8.53	1.04±.359		3.534±.133	-.0691±.0350
42 × 2 × 3	✗	.247±.119	.504±.0629	1.79±.0943	1.23±.0604		3.257±.225	-.0078±.028
42 × 2 × 3	✓	.144±.0102	.509±.0861	1.42±1.03	1.56±1.79		2.34±.182	-.0083±.0178
42 × 2 × 2 × 2	✗	.0859 ±.039	.6585 ± 0.593	11.9±6.07	7.23± 2.82		2.69 ± 0.229	.0415±.0289
42 × 2 × 2 × 2	✓	0.146±.0860	.323±0.103	8.79± 10.7	11.2 ± 8.04		2.33± .226	.0219±.0185
42 × 6	✗	.0880±.0367	.582±.136	2.17± 6.07	1.47±2.82		2.69±0.212	N/A
42 × 6	✓	.239 ±.0645	.645±.452	2.16±.775	2.51±1.25		2.33±0.181	N/A

Table 3: Evaluation of compositionality, concatenativity, and degree of fusion of emergent languages in default and inflection Attr-Val settings, without and with articulation pressure (Art.).

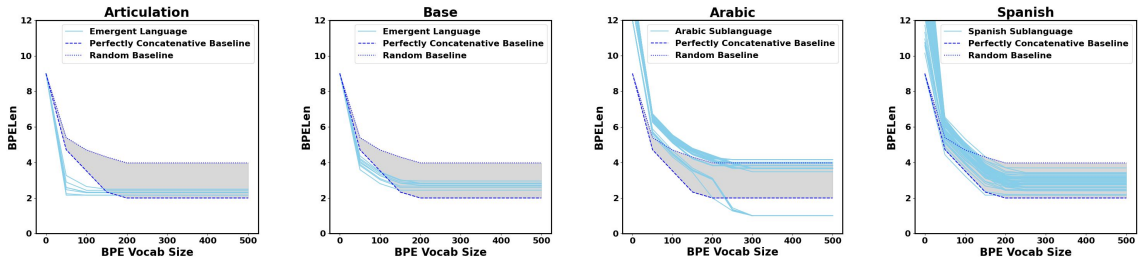


Figure 3: BPELen at varying  $|V|$  for emergent and natural languages in the  $42 \times 2 \times 3$  inflection Attr-Val setting. Comprehensive plots for the full range of inflection experiments with natural language comparisons in Appendix I.

do find interesting, nonconcatenative structures.

This may not be unrealistic: the natural languages also exhibit such a spectrum. Spanish appears relatively diverse, likely due to the presence of vowel patterns and auxiliary verbs. Arabic is more bimodal, with some very nonconcatenative sublanguages and some which have a high maximal compression due to being very fusional.

However, the introduction of articulation pressure sharply decreases BPELen in emergent languages. We perform additional  $t$ -tests on the mean BPELen<sub>96</sub> of emergent languages between conditions, observing statistical significance in this change for both the default and inflection Attr-Val setting ( $p \approx 0.009$  and  $p \approx 4.5 \cdot 10^{-7}$  respectively). This suggests ease of articulation is indeed a strong pressure towards concatenation, and the presence of “unpronounceable” clusters at the phonological (character) level is sufficient to induce concatenation at the morphological (symbol) level. More broadly, this may help explain the dominance of concatenative morphology in natural languages. Table 4 identifies a concatenative symbol in a language from this setting, where the string bebe communicates a first attribute value of 10.

It is also worth noting that in the default Attr-Val setting, all of the successful articulatory languages exhibited  $\text{BoSDis}_C^V > 1$ . In the inflection setting, only 3 of 7 successful languages exhibited

$\text{BoSDis}_C^V > 1$ , indicating that not all of the segmentations were meaningful. We hypothesize that the additional constraint caused difficulties in segmenting, or alternatively that the pressure may have also resulted in highly nonconcatenative languages that appear concatenative due to the additional uninformative pattern induced by the articulation pressure. The fact that the intersection of these two conditions results in a somewhat unintuitive result underscores the need for future work to understand the fine-grained details of EmCom.

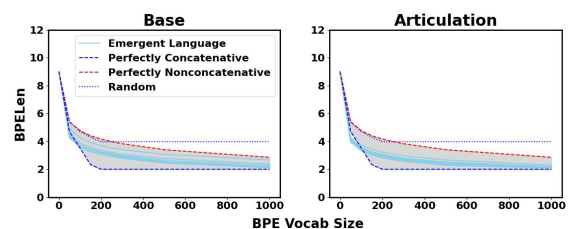


Figure 4: BPELen at varying  $|V|$  for emergent languages in the default Attr-Val setting.

## 6.2.2 TopSim in Inflectional Languages

To shed more light on the compositionality of naturalistic communication schemes, we compare the TopSim of emergent and natural languages in the inflection Attr-Val setting in Figure 5, including emergent languages with the articulation pressure. Compared to natural languages, TopSim for emer-

State	Message	State	Message
(0, 2)	eedhedhee	(10, 0, 0)	hehgbebeb
(1, 2)	eehedeb	(10, 0, 1)	bebgbebeb
(2, 2)	eedeedeeh	(10, 0, 2)	bebgbebeb
...	...	(10, 1, 0)	ebcbgbebe
(40, 2)	eehebebeb	(10, 1, 1)	ebcbgbebe
(41, 2)	eedeeeebe	(10, 1, 2)	ebcbgbebe

Table 4: Example nonconcatenative (left) and concatenative (right) symbols from successful emergent communication schemes. Message vocabulary displayed as Latin letters for readability. *Left*: The repeated character e represents states with a second attribute value of 2 across all combinations in the  $42 \times 6$  inflection-inspired setting (most omitted for space). *Right*: The string bebe represents states with a first attribute value of 10 across all combinations in the  $42 \times 2 \times 3$  inflection-inspired setting with ease of articulation pressure.

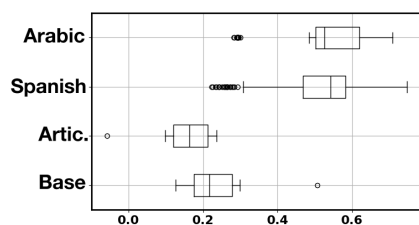


Figure 5: Topographic similarities of  $42 \times 2 \times 3$  emergent (without and with articulation pressure) and natural languages. See Appendix I for additional conditions.

gent languages was low, except for one instance with TopSim of .507 (in the range of natural languages). In the  $42 \times 2 \times 2 \times 2$  and  $42 \times 6$  settings, but not in the  $42 \times 2 \times 3$  setting, the articulation pressure resulted in a significant increase in TopSim. There seems to be a complex interplay between this constraint and the specific attribute-value setup, which we leave for future work to explore in depth. Together, this suggests that emergent languages still have room for improvement in achieving human-like topographic similarity, and there may remain more pressures toward this to identify and explore.

### 6.2.3 Fusion Under Inflection

As we would predict, the  $16 \times 16 \times 16$  setting was least fusional, with a quite negative F-TopSim  $\delta$ . In the  $42 \times 2 \times 3$  setting, F-TopSim  $\delta$  was near 0, with 0 within a standard deviation of the mean. This implies there was some degree of fusion within these languages, since a completely non-fusional language would suffer a large drop in TopSim by fusing two attributes. The  $42 \times 2 \times 2 \times 2$  setting was quite fusional, with a mean F-TopSim  $\delta$  of 0.0415 (i.e., by fusing the right attributes, we get a TopSim

increase of over 0.04.)

Suppletion (fusion of a root with grammatical features, e.g., *went* as the past tense of *go*) is much rarer than fusion between grammatical features. Our EmCom results mirror this: the pair of lowest-cardinality attributes (representing tense and person) had the highest F-TopSim  $\delta$  in 6 of 8 of the  $42 \times 2 \times 3$  emergent languages, suggesting these attributes are most likely to fuse. Similar results were observed for the  $42 \times 2 \times 2 \times 2$  setting. This may be explained by symbolic complexity (Zhang, 2025) increasing the least in fusing these attribute pairs; while 5 symbols are required to communicate the values for these two attributes before fusion, only 6 are required after. Meanwhile, fusing the root attribute (with 42 values) with another attribute can drastically increase the symbolic complexity (i.e., 2-3 times). This trend mirrors the fusion of fixed grammatical features in natural language, which we also observed in most of our generated Arabic and Spanish sublanguages.

## 7 Conclusion

In this work, we outline an experimental paradigm for emergent morphology, motivated by the principle of double articulation and a rich analogy between the Attr-Val game and inflectional morphology. We review existing metrics for EmCom, proposing new ones for concatenativity and degree of fusion. We calibrate our metrics towards a wide array of idealized compositional schemes that are possible under this paradigm, as well as schemes based on natural language inflection. We then demonstrate through EmCom experiments the power of this reimagined setting to produce linguistically interesting analyses grounded in comparison to natural language. In particular, the double articulation aspect of our paradigm reveals simulated articulatory constraints as a promising pressure toward concatenativity, which prior work has often implicitly assumed in evaluations of compositionality and attempted to elicit through other pressures, and future work may use to test more specific morphophonological hypotheses. Further, our emergent languages rederive the bias towards fusion of grammatical features in natural language. Future work should adopt small-vocabulary and inflectional Attr-Val settings for focused, linguistically motivated studies of morphology allowing direct comparison to natural language.

## Limitations

**Naturalness of task inputs.** Like related works (Resnick et al., 2020; Chaabouni et al., 2020; Ueda et al., 2023), we chose to focus on Attr-Val reconstruction as our EmCom task setting. This setting removes the requirement of agents to perceive key features from stimuli (e.g., images), instead only requiring agents to learn to communicate ground truth features, significantly simplifying the learning problem. However, we intentionally chose this setting to ensure comparability with Ueda et al. (2023), and because our work is entirely focused on how EmCom agents learn to communicate features, thus possible perception errors would needlessly obstruct this effort. Further, considering that we aim to explore inflectional morphology in EmCom, the Attr-Val game is a natural analogy to existing models of morphology in linguistics, e.g., the lexeme-slot encoding of inflectional morphology (Wu et al., 2019). Nonetheless, while out of scope for this work, an interesting direction for future work would be to investigate relationships between aspects of visual stimuli and morphology in emergent languages with small vocabularies.

**Representativeness of natural languages.** In our natural language reference points, we only considered Arabic and Spanish, which are not a representative sample of the world’s natural languages, and were chosen in part due to the availability of data. Without a much larger and more diverse dataset, we cannot definitively say whether a given emergent language is naturalistic. Additionally, the Attr-Val scheme we focused on is a rather small subset of the complexity of natural morphologies. A full analysis of a large set of natural languages, to cover this diversity of morphologies, is outside the scope of this paper, but we believe our examples will provide a solid initial point of comparison for emergent languages. Boldt and Mortensen (2024a) explored larger-scale cross comparisons of emergent languages, and we anticipate that the addition of parallel natural language data to such analyses would prove useful for future work.

**Representativeness of articulation pressure.** The choice of an even-odd toy phonological constraint for the articulation pressure was arbitrary, and not representative of the range of such constraints in natural language. An alternative explanation for the increased concatenativity from this pressure is that the predictable even-odd alternation

pattern resulting from this pressure simply made the data more compressible due to the fewer number of possible variations in messages. While this is only a simple proof of concept of the sorts of linguistically motivated experiments we can do when embracing morphology and double articulation in EmCom, we argue that phonological constraints in natural languages may provide a similar advantage in human language learning, and encourage future investigation of how they may influence concatenativity and other properties in EmCom.

## Acknowledgments

We thank our anonymous reviewers for their thoughtful and constructive feedback. We thank Zheyuan Zhang for helpful discussion about this work, and Ryo Ueda for consulting with us on the implementation of metrics in Ueda et al. (2023). This research was supported in part through computational resources and services provided by Advanced Research Computing at the University of Michigan, Ann Arbor.

## References

- Diana Archangeli and Douglas Pulleyblank. 2016. Emergent morphology. In *Morphological metatheory*, pages 237–270. John Benjamins Publishing Company.
- Brendon Boldt and David Mortensen. 2024a. [Elcc: the emergent language corpus collection](#). *Preprint*, arXiv:2407.04158.
- Brendon Boldt and David R Mortensen. 2024b. [A review of the applications of deep learning-based emergent communication](#). *Transactions on Machine Learning Research*.
- Henry Brighton and Simon Kirby. 2006. [Understanding linguistic evolution by visualizing the emergence of topographic mappings](#). *Artificial Life*, 12:229–242.
- Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. 2020. [Compositionality and generalization in emergent languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4427–4442, Online. Association for Computational Linguistics.
- Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. 2019. [Anti-efficient encoding in emergent communication](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Rahma Chaabouni, Florian Strub, Florent Alché, Eugene Tarassov, Corentin Tallec, Elnaz Davoodi,

- Kory Wallace Mathewson, Olivier Tieleman, Angeliki Lazaridou, and Bilal Piot. 2022. [Emergent communication at scale](#). In *International Conference on Learning Representations*.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. [Learning phrase representations using rnn encoder-decoder for statistical machine translation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, page 1724–1734.
- Henry Conklin and Kenny Smith. 2023. [Compositionality with variation reliably emerges in neural networks](#). In *The Eleventh International Conference on Learning Representations*.
- Dan Dediu, Rick Janssen, and Scott R. Moisik. 2017. [Language is not isolated from its wider environment: Vocal tract influences on the evolution of speech and language](#). *Language & Communication*, 54:9–20. The multimodal origins of linguistic communication.
- Peter Dekker. 2024. *Identifying drivers of language change using agent-based models*. Ph.D. thesis, Vrije Universiteit Brussel.
- Micha Elsner, Martha B Johnson, Stephanie Antemaso, and Andrea D Sims. 2020. Stop the morphological cycle, i want to get off: Modeling the development of fusion. *Society for Computation in Linguistics*, 3(1).
- Sara Finley and Elissa Newport. 2021. [Segmentation of root and pattern morphology](#). *PsyArXiv*.
- Michelle Alison Fullwood. 2018. *Biases in segmenting non-concatenative morphology*. Ph.D. thesis, Massachusetts Institute of Technology.
- Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 2:23–38.
- Lukas Galke, Yoav Ram, and Limor Raviv. 2022. [Emergent communication for understanding human language evolution: What’s missing?](#) In *EmeCom at ICLR 2022*.
- Lukas Galke and Limor Raviv. 2024. [Learning and communication pressures in neural networks: Lessons from emergent communication](#). *Language Development Research*, 5.
- Eleonora Gualdoni, Mycal Tucker, Roger Levy, and Noga Zaslavsky. 2024. [Bridging semantics and pragmatics in information-theoretic emergent communication](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 21059–21078. Curran Associates, Inc.
- Abhinav Gupta, Cinjon Resnick, Jakob Foerster, Andrew Dai, and Kyunghyun Cho. 2020. [Compositionality and capacity in emergent languages](#). In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 34–38, Online. Association for Computational Linguistics.
- Ximena Gutierrez-Vasques, Christian Bentz, and Tanja Samardžić. 2023. [Languages through the looking glass of bpe compression](#). *Computational Linguistics*, 49(4):943–1001.
- Ximena Gutierrez-Vasques, Christian Bentz, Olga Sozinova, and Tanja Samardzic. 2021. [From characters to words: the turning point of BPE merges](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3454–3468, Online. Association for Computational Linguistics.
- Zellig S. Harris. 1955. [From phoneme to morpheme](#). *Language*, 31(2):190–222.
- Charles Hockett. 1960. The origin of speech. *Scientific American*, 203:88–96.
- Fred Jehle. 2011. [Fred Jehle’s conjugated Spanish verb database](#).
- Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 2019. [Egg: a toolkit for research on emergence of language in games](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, page 55–60.
- Simon Kirby, Tom Griffiths, and Kenny Smith. 2014. [Iterated learning and the evolution of language](#). *Current Opinion in Neurobiology*, 28:108–114. SI: Communication and language.
- Tomasz Korbak, Julian Zubek, and Joanna Rączaszek-Leonardi. 2020. [Measuring non-trivial compositionality in emergent communication](#). *Preprint*, arXiv:2010.15058.
- Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. 2017. [Natural language does not emerge ‘naturally’ in multi-agent dialog](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, page 2962–2967. Copenhagen, Denmark. Association for Computational Linguistics.
- Tom Kouwenhoven, Max Peeperkorn, Bram Van Dijk, and Tessa Verhoef. 2024. [The curious case of representational alignment: Unravelling visio-linguistic tasks in emergent communication](#). In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 57–71, Bangkok, Thailand. Association for Computational Linguistics.
- Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. 2018. [Emergence of linguistic communication from referential games with symbolic and pixel input](#). In *International Conference on Learning Representations*.
- Vladimir I. Levenshtein. 1965. [Binary codes capable of correcting deletions, insertions, and reversals](#). *Soviet physics. Doklady*, 10:707–710.

- David Lewis. 1979. [Scorekeeping in a language game](#). *Journal of Philosophical Logic*, 8:339–359.
- Fushan Li and Michael Bowling. 2019. [Ease-of-teaching and language structure from emergent communication](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Yuchen Lian, Arianna Bisazza, and Tessa Verhoef. 2023. [Communication drives the emergence of language universals in neural agents: Evidence from the word-order/case-marking trade-off](#). *Transactions of the Association for Computational Linguistics*, 11:1033–1047.
- Yuchen Lian, Tessa Verhoef, and Arianna Bisazza. 2024. [NeLLCom-X: A comprehensive neural-agent framework to simulate language learning and group communication](#). In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 243–258, Miami, FL, USA. Association for Computational Linguistics.
- John McCarthy and Alan Prince. 1995. Faithfulness and reduplicative identity. *Papers in Optimality Theory. University of Massachusetts Occasional Papers*, 18.
- Haq Nawaz, Manal Elobaid, Ali Al-Laith, and Saif Ullah. 2025. [Automated generation of Arabic verb conjugations with multilingual Urdu translation: An NLP approach](#). In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 136–143, Abu Dhabi, UAE. Association for Computational Linguistics.
- Martin Nowak and David Krakauer. 1999. [The evolution of language](#). *Proc. Natl. Acad. Sci. USA*, 96:8028–8033.
- Neil Rathi, Michael Hahn, and Richard Futrell. 2021. [An information-theoretic characterization of morphological fusion](#). *An Information-Theoretic Characterization of Morphological Fusion*.
- Neil Rathi, Michael Hahn, and Richard Futrell. 2022. [Explaining patterns of fusion in morphological paradigms using the memory–surprisal tradeoff](#). In *Proceedings of the annual meeting of the cognitive science society*, volume 44.
- Cinjon Resnick, Abhinav Gupta, Jakob Foerster, Andrew Dai, and Kyunghyun Cho. 2020. [Capacity, bandwidth, and compositionality in emergent language learning](#). In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-agent Systems*.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Kumiko Tanaka-Ishii. 2021. [Articulation of Elements](#), pages 115–124. Springer International Publishing, Cham.
- Mycal Tucker, Roger Levy, Julie A Shah, and Noga Zaslavsky. 2022. [Trading off utility, informativeness, and complexity in emergent communication](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22214–22228. Curran Associates, Inc.
- Ryo Ueda, Taiga Ishii, and Yusuke Miyao. 2023. [On the word boundaries of emergent languages based on harris’s articulation scheme](#). In *The Eleventh International Conference on Learning Representations*.
- Ryo Ueda and Tadahiro Taniguchi. 2024. [Lewis’s signaling game as beta-vae for natural word lengths and segments](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Ryo Ueda and Koki Washio. 2021. [On the relationship between Zipf’s law of abbreviation and interfering noise in emergent languages](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 60–70, Online. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Kasun Vithanage, Rukshan Wijesinghe, Alex Xavier, Dumindu Tissera, Sanath Jayasena, and Subha Fernando. 2023. [Accelerating language emergence by functional pressures](#). *PLOS ONE*, 18(12):1–28.
- Ronald J. Williams. 1992. [Simple statistical gradient-following algorithms for connectionist reinforcement learning](#). *Mach. Learn.*, 8(3–4):229–256.
- Shijie Wu, Ryan Cotterell, and Timothy O’Donnell. 2019. [Morphological irregularity correlates with frequency](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5117–5126, Florence, Italy. Association for Computational Linguistics.
- Zheyuan Zhang. 2025. [A combinatorial approach to neural emergent communication](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1660–1666, Abu Dhabi, UAE. Association for Computational Linguistics.
- Willem Zuidema and Bart de Boer. 2009. [The evolution of combinatorial phonology](#). *Journal of Phonetics*, 37:125–144.

## A Additional Notes On Inflection Setting

In the inflection setting of Attr-Val, one possible limitation is the lack of control on the number of combinations of attribute values (rather than unique number of attribute values, which we control for instead). We conducted a preliminary experiment where we controlled for number of combinations rather than number of values, however these models universally failed to converge, instead learning to distinguish either grammatical features or roots, but never both.

We believe that this failure to converge is related to the model capacity, as explored by Resnick et al. (2020). Specifically, we hypothesize that for languages with more unique values (i.e., symbolic complexity as defined in Zhang, 2025), the minimum model capacity for a successful language increases. While we could simply increase the size of this model (e.g., embedding and/or hidden size) to better encourage its convergence, it would no longer be comparable to other results. Meanwhile, if we increased the sizes of all models, this may influence compositionality in less complex configurations. As such, we decided that a comparison controlling for the number of combinations would not be possible within the scope of this paper.

Nevertheless, we chose our specific Attr-Val configuration in the interest of comparability to natural language. To make a good comparison, we needed to be able to select features from natural language with equal or greater cardinalities to our artificial features. As such, there are only so many grammatical features we can introduce and find high-quality data for, and considering features with larger cardinalities would preclude a lot of natural language features like gender, number, person, etc. and severely limit our dataset. The specific values were chosen to allow this comparison and control for symbolic complexity.

## B Supplementary Agent Implementation Details

Observed states for sender agents were implemented by concatenating one-hot vectors for each attribute. Each run occurred on a single A40 GPU with 48GB VRAM and took about 8 hours, although many models converged before then. Models generally converged after about 400 training epochs.

In the inflection setting, we additionally weighted the root accuracy so it represented 90% of

the accuracy optimized by agents, both to stabilize training and reflect that far more of the semantic content of a word is stored in the lexeme rather than the slot (thus the lexeme is more important to communicate). We also experimented with larger configurations with more “roots,” but we found that higher input sizes led to unstable training where agents would learn either to inflect or distinguish roots, but not both. We attribute this to increased symbolic complexity of the setting (Zhang, 2025).

## C Harris’ Articulation Scheme

We replicated the results in Ueda et al. (2023) that C-TopSim is greater than W-TopSim. This result led the authors to suggest that Criterion 3 for Harris’ Articulation Scheme (i.e., W-TopSim should be higher than C-TopSim) may be unsolvable. As discussed in Section 4.1.2, we use ratios of BoSDis to judge whether segmented symbols are meaningful. We assert that this is a reasonable replacement for the intended purpose of the criterion, since it asks the question "do segments act as symbols which tend to refer to some meanings more than others."

Ueda et al. (2023) report negative results on the basis of low W-TopSim, which does not hold for our replacement metric of the BosDis ratio. Thus, we believe that emergent languages do exhibit HAS, and that HAS does not represent a significant gap between natural and emerged languages, as is claimed. The HAS algorithm, therefore represents a useful tool for interpreting emergent communication.

## D Positional Disentanglement

Motivated by the possibility that symbols in different positions could indicate different attributes (common in natural languages), Chaabouni et al. (2020) propose positional disentanglement (PosDis):

$$\frac{1}{m} \sum_{v \in V} \frac{\mathcal{I}(s_j; a_v) - \mathcal{I}(s_j; b_v)}{\mathcal{H}(s_j)} \quad (3)$$

For a communication scheme with messages of length  $m$ , PosDis calculates the average difference

C-TopSim	W-TopSim
0.611	0.362

Table 5: W-TopSim and C-TopSim for a perfectly concatenative language.

in mutual information  $\mathcal{I}$  between the value of a symbol  $s_j$  in position  $j \in [1, m]$  and attributes  $a_v$  and  $b_v$ , which respectively have the most and second-most mutual information with  $v$ .<sup>11</sup> This is then normalized by the entropy  $\mathcal{H}$  of  $s_j$ . PosDis will be high if attributes are uniquely distinguishable by positions of symbols in messages. However, as we did not observe insightful trends with respect to this metric, we omitted PosDis from results presented in the paper to conserve space.

## E Artificial Languages Extended Analysis

In this appendix, we provide extended discussions about concatenation and fusion in artificial languages. We then propose and evaluate additional nontrivial compositional artificial languages based on mutation and more general functional compositionality.

**On Concatenation** First, the level of concatenation in the first three artificial languages has minimal effect on TopSim and learnability, implying a lack of inductive bias toward concatenation in recurrent receiver neural networks. Notably, this is not consistent with human results in linguistics (Finley and Newport, 2021), and implies additional pressures may be required to induce concatenative morphology. Meanwhile, as expected,  $\text{BoSDis}_C^V$  metrics indicate that both HAS and BPE successfully extract more meaningful morphemes from more concatenative languages, despite all artificial languages being compositional. In the mixed concatenative and variable length languages, HAS yields lower metrics, suggesting it may indeed be less effective in languages with randomness. BPE can seemingly compensate for this as long as a sufficiently limited vocabulary size  $|V|$  is chosen. The right vocabulary size is not obvious, however, as the  $\text{BoSDis}_C^V$  for  $\text{BPE}_{\text{Max}}$  is relatively low for artificial languages, suggesting an over-segmented language that erases the disentanglement of symbols. While segmentation algorithms behave as expected for systematic languages, these metrics are greater than 1 for random languages. This may result from over-sensitivity to noise in the absence of other signal, e.g., the presence of a few rare, highly predictive symbols may skew the metric upwards. This may also be attributed to a vanishingly small BoSDis for character-level segmentation. Fu-

<sup>11</sup>Given the reliance on a consistent message length, we only apply PosDis based on individual characters rather than symbols extracted from HAS and BPE.

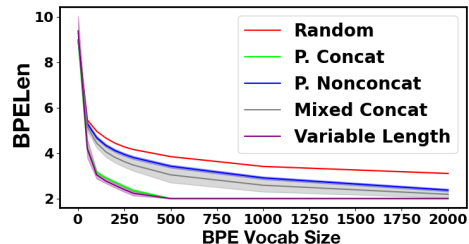


Figure 6: BPELen at varying  $|V|$  for artificial languages, averaged for random seeds with  $1\sigma$  confidence interval.

ture research should consider this edge case when comparing BoSDis before and after symbol segmentation.

Focusing in on concatenativity as measured by HASLen and BPELen metrics, we unsurprisingly observe that the perfectly concatenative and variable length languages (which are both entirely concatenative) exhibit the highest concatenativity, except for  $\text{BPELen}_{\text{Max}}$ , which shows no sensitivity to the level of concatenativity in artificial languages. This further supports the possibility of over-segmenting languages into symbols. As such, in Figure 6, we visualize the BPELen for a range of symbol vocabulary sizes. While minimal compression and maximal compression are similar across all artificial languages except the random language, there is a wide range in which BPELen behaves as expected. The mixed concatenative languages appear to fall on a spectrum between perfect concatenation and nonconcatenation, thus BPELen appears to be an effective measure of concatenativity. Meanwhile, due to the marked difference in BPELen between random and systematic languages, maximal compression may be an effective way to judge the amount of randomness in messages.

**On Fusion** F-TopSim  $\delta$  values indicate that the fusion language is indeed the most fusional. We observe that F-TopSim  $\delta$  for random languages is zero, which does not suggest fusion, rather that TopSim does not change with respect to fusion of attributes (expected for a random language, which should not assign any meaningful symbols to any attribute values or pairs thereof). We observe that degree of fusion does not vary significantly with respect to concatenativity, but increases in the variable length concatenative language.

**On Mutation and Functional Compositionality** Traditional metrics for compositionality such as TopSim do not account for some types of nontrivial compositionality, such as phonemic mutation (e.g., of vowels in inflection of *sing*, *sang*,

and *sung*), or more general function-based compositionality, where symbols from some attributes serve as functions applied to other attributes. As visualized in Figure 7, we defined additional languages for these phenomena. Specifically, following Korbak et al. (2020), the *mutation* language allows symbols to overlap, and combines overlapping characters’ values by addition modulo  $|C|$ . We specifically consider two forms of the mutation language:  $\text{Mutation}_3$ , which consists of length 5 symbols each overlapping each other at 3 character positions, and  $\text{Mutation}_0$ , which consists of length 9 symbols each entirely overlapping. Lastly, the *symbolic function* language defines one symbol as a function, specifically a *reordering* function applied to the characters of another symbol. Beyond these artificial languages, we also consider a baseline *random language*.

**On  $\text{BoSDis}_C^V$  in Random Languages** One may note that the  $\text{BoSDis}_C^V$  for random languages is quite high. We expect this for 2 reasons:

1. They have character-level  $\text{BoSDis}$  close to 0.
2. By sheer random chance, some segments of characters will be better predictors of attributes than others. This results in a slight increase in symbol-level  $\text{BoSDis}$ . This results in a very high ratio, since even a small absolute increase may be 10x as great as the original  $\text{BoSDis}$ . For this reason we don’t advocate putting much stock in the magnitude of the ratio.

Whether this is a problem for the metric is a matter of interpretation. We believe that since random languages are very unique when all metrics, especially  $\text{TopSim}$ , are taken into account, the risk of misinterpreting a random language as a meaningfully segmented one is low. Furthermore, this enhances the confidence we can place in low  $\text{BoSDis}$  ratio being an indicator of non-concatenativity, as these languages are qualitatively different from both random and concatenative languages (see  $\text{BoSDis}_C^V < 1$  for perfectly nonconcatenative languages in Table 1).

The evaluation metrics for these artificial languages are listed in Tables 6 and 7. Despite being compositional, these languages drastically impact evaluation metrics. The reordering and  $\text{Mutation}_3$  language are substantially less compositional (according to  $\text{TopSim}$  and  $\text{BoSDis}$ ), segmentable, and

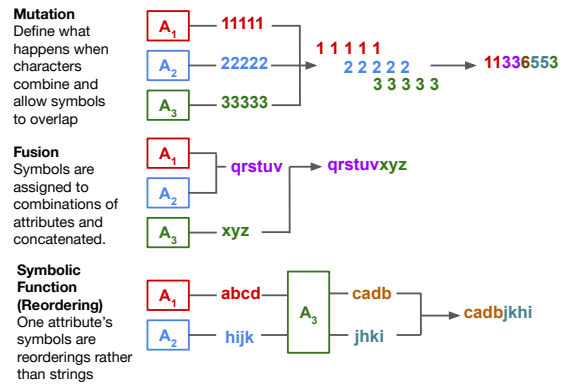


Figure 7: Artificial languages targeting mutation and symbolic functions (reordering). These languages define a unique symbol for each possible attribute value, and apply nontrivial composition operation for symbols.

learnable than fully concatenative languages. In  $\text{Mutation}_0$ , where all characters are mutated twice, the languages were essentially indistinguishable from random languages by almost all metrics, including learnability. Even though this language consists entirely of messages formed by composing symbols according to simple, defined rules, in practice this compositionality may not matter or be recognized through existing metrics. The only metrics where there was a noticeable difference are the  $\text{BoSDis}_C^V$  ratios for HAS and BPE, but the interpretations of these are unclear, since segments cannot be disentangled symbols for this type of language.

## F HASLen for Natural and Emergent Languages

In Figure 8, we visualize the HASLen for emergent languages (on average) along with Spanish and Arabic natural language reference points. Following observations in Section 5.1, we observe that Spanish and Arabic are scored with a higher HASLen than random languages. This further demonstrates HAS’ sensitivity to randomness, making it difficult to use for reference points of concatenativity (despite its effectiveness in identifying symbol boundaries).

## G Segmentation Analysis of Natural Languages

As shown in Table 8, when calculating the  $\text{BoSDis}_C^V$  ratio for these languages, we find that

Language	TopSim	BoSDis	HAS BoSDis <sub>C</sub> <sup>V</sup>	BPE <sub>96</sub> BoSDis <sub>C</sub> <sup>V</sup>	BPE <sub>Max</sub> BoSDis <sub>C</sub> <sup>V</sup>	F-TopSim	Epochs
Reordering	.227±.01	.042±.03	23.24±22.30	88.64±76.56	13.44 ±11.84	-0.006±0.01	20.25±5.47
Mutation <sub>3</sub>	.359±.01	.060±.00	40.37±0	17.31±.00	.578±.25	-.023±.01	24.88±9.85
Mutation <sub>0</sub>	.000±.00	.001±.00	3146±1736	7461±3091	17.204±5.46	.001±.00	100+

Table 6: Evaluation metrics for compositionality, degree of fusion, and learnability of additional artificial languages averaged over 8 random seeds. Learnability is measured by the average number of epochs that a receiver agent takes to learn the language, i.e., exceed 99% reconstruction accuracy.  $1\sigma$  confidence intervals are provided to within 2 decimal points, or whole numbers for large BoSDis<sub>C</sub><sup>V</sup> values.

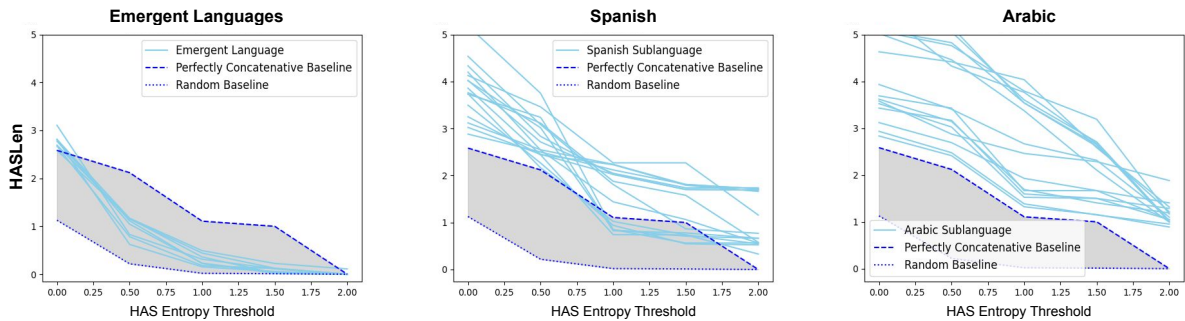


Figure 8: HASLen of emergent and natural languages with respect to HAS entropy threshold.

Language	HASLen	BPELen <sub>96</sub>	BPELen <sub>Max</sub>
Reordering	2.268±.17	4.127±.04	2.086±.01
Mutation <sub>3</sub>	4.843±.17	4.461±.04	2.273±.05
Mutation <sub>0</sub>	1.560±.38	4.998±.03	3.107±.02

Table 7: Evaluation metrics for concatenativity of artificial languages, averaged over 8 random seeds.  $1\sigma$  confidence intervals are provided to within 2 decimal points, or whole numbers for large BoSDis<sub>C</sub><sup>V</sup> values.

	HAS	BPE <sub>96</sub>
Spanish	0.76	0.74
Arabic	0.08	0.00

Table 8: Proportion of sublanguages meaningfully segmented (BoSDis<sub>C</sub><sup>V</sup>>1) by BPE<sub>96</sub> and HAS.

up to 76% of generated Spanish sublanguages<sup>12</sup> can be meaningfully segmented by either HAS or BPE (i.e., ratio greater than 1), while virtually none of the Arabic sublanguages can be. On the latter, this demonstrates how existing tools for symbol segmentation are insufficient to capture highly non-concatenative morphologies.

## H Ease-of-Learning Pressure Experiment

Inspired by prior work in iterated learning (Kirby et al., 2014), we implement an experiment incen-

<sup>12</sup>Some failures here could be attributed to nonconcatenativity in the dataset caused by prepended auxiliary verbs.

tivizing ease of learning in emergent languages (Li and Bowling, 2019). Specifically, every 100 epochs, we reset the receiver’s network weights. In theory, this should encourage emergent languages that are easier for the receiver to learn quickly, which Li and Bowling show improves compositionality of the resulting languages. To understand whether this ease of learning is attributed to concatenativity, which prior work has typically not distinguished from compositionality, Figure 9 visualizes the BPELen of emergent languages with this pressure in the default Attr-Val setting (i.e., 3 attributes each with 16 values). Interestingly, we find that the languages resulting from this pressure are not significantly more concatenative than those without it shown in Figure 4. Such a result is unexpected from an intuitive standpoint, but aligns with the results in 6, which show that concatenativity, by itself, does not have a relationship with learnability. This motivates future research around how various pressures may influence fine-grained properties of emergent languages, and how much of human language structure can be explained by the generational (iterated) learning paradigm.

## I Additional Natural Language Comparisons

Figure 10 shows BPE compression computed for Spanish, Arabic, emergent languages without articulation pressure (Em. Base) and emergent lan-

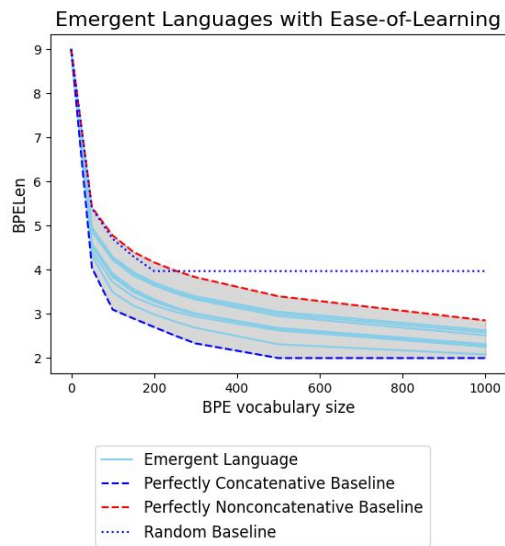


Figure 9: BPELen (concatenativity) of emerged languages with ease-of-learning pressure at varying symbol vocabulary sizes  $|V|$ .

guages with articulation pressure (Em. Artic.). Figure 11 shows TopSim ranges for the same experiments. Due to lack of a sufficiently large grammatical feature in our Arabic dataset, we were unable to compute an Arabic baseline for the  $42 \times 6$  condition.

## J License Information

We provide license information and links for the following artifacts used in this work:

- EGG Framework: MIT License.  
<https://github.com/facebookresearch/EGG>  
License: <https://github.com/facebookresearch/EGG/blob/main/LICENSE>
- Fred Jehle’s Spanish Verb Dataset: Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.  
<https://github.com/ghidinelli/fred-jehle-spanish-verbs>  
License: <https://creativecommons.org/licenses/by-nc-sa/3.0/>
- Arabic-Urdu Conjugation Dataset

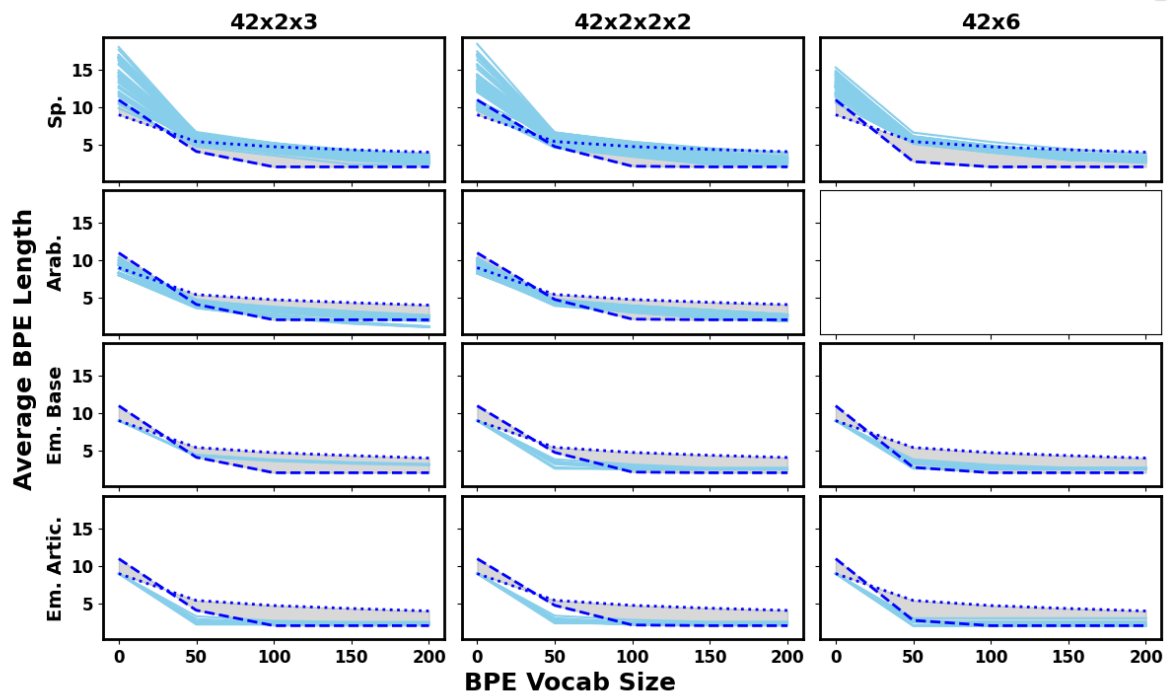


Figure 10: BPE compression for natural languages (Spanish and Arabic) and emergent languages across all inflection conditions. Dotted blue lines: random baselines; dashed blue lines: perfectly concatenative baselines; solid light blue lines: sublanguages or emergent languages, depending on condition.

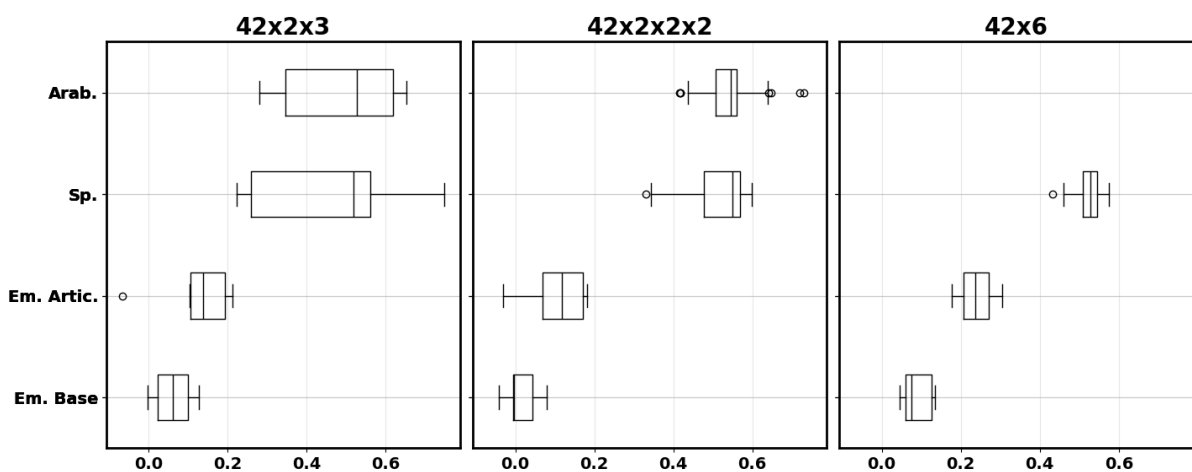


Figure 11: TopSim for natural languages (Spanish and Arabic) and emergent languages across all inflection conditions.