Frequency & Compositionality in Emergent Communication

Jean-Baptiste Sevestre*
ENS-PSL, EHESS, CNRS

Emmanuel Dupoux
ENS-PSL, EHESS, CNRS

Abstract

In natural languages, frequency and compositionality exhibit an inverse relationship: the most frequent words often resist regular patterns, developing idiosyncratic forms. This phenomenon, exemplified by irregular verbs where the most frequent verbs resist regular patterns, raises a compelling question: do artificial communication systems follow similar principles? Through systematic experiments with neural network agents in a referential game setting, and by manipulating input frequency through Zipfian distributions, we investigate if these systems mirror the irregular verbs phenomenon, where messages referring to frequent objects develop less compositional structure than messages referring to rare ones. We establish that compositionality is not an inherent property of the frequency itself and provide compelling evidence that limited data exposure, which frequency distributions naturally create, serves as a fundamental driver for the emergence of compositional structure in communication systems, offering insights into the cognitive and computational pressures that shape linguistic systems.

1 Introduction

Neural networks provide a modern framework for exploring how fundamental linguistic properties emerge during the evolution of language (Kottur et al., 2017; Lazaridou and Baroni, 2020; Rita et al., 2024). Among these properties, compositionality—the ability to combine discrete meaningful units to create complex expressions—is considered a defining feature of human language (Bickerton, 2007; Hockett, 1960). Our research explores the relationship between compositionality and frequency distributions in emergent languages.

Natural languages across cultures exhibit a universal tendency: highly frequent elements often resist regular patterns and develop idiosyncratic structures Pinker (1999); Bybee (2007). Irregular verbs provide a well-known example of this pattern, as the most frequently used verbs (e.g., "to be", "to go") tend to resist regular morphological conventions (Francis and Kučera, 1982) (Figure 1).

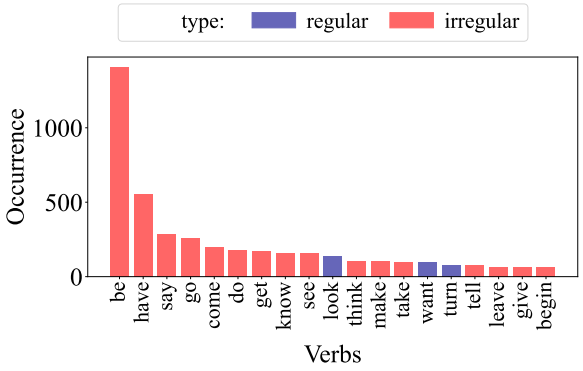

Figure 1: Most common verbs in The Great Gatsby[1]

Researchers in language evolution have questioned the prerequisites for this phenomenon. Kirby (2001) investigated linguistic structure evolution through an Iterated Learning Model (Smith et al., 2003) with symbolic algorithms, demonstrating that limited data exposure creates a learning bottleneck pushing toward structured language development. His work showed that with Zipfian distributions (Zipf, 2013), frequent elements develop irregular structure while being expressed in shorter forms.

With the recent advent of deep neural networks (LeCun et al., 2015), there has been a resurgence of computational simulations of language evolution (Lazaridou and Baroni, 2020; Rita et al., 2024), enabling scaled experiments with learning agents unbound by pre-established rules (Chaabouni et al., 2022). With such simulations, Chaabouni et al. (2019) demonstrated that standard neural network agents tend to develop in-

*Corresponding Author: jean-baptiste.sevestre@ens.psl.eu

[1]Data source: https://www.gutenberg.org/cache/epub/64317/pg64317.txt

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 27251–27263

efficient communication codes that violate Zipf's Law of Abbreviation (ZLA) (Zipf, 1949): the principle that more frequent words tend to be shorter. This inefficiency raises a compelling question: if those emergent languages struggle to conform to basic efficiency principles like ZLA, how might frequency affect their compositional properties?

While compositionality (Kottur et al., 2017; Chaabouni et al., 2020; Li and Bowling, 2019; Ren et al., 2020; Rita et al., 2022) and frequency effects (Chaabouni et al., 2019; Rita et al., 2020) have been extensively studied in neural emergent communication, their relationship remains unexplored. To address this gap, we leverage neural network capabilities to scale Kirby (2001)'s experimental setup —originally limited to 25 objects and qualitative methods— to approximately one million distinct objects and quantitatively study the impact of frequency on compositionality.

In this paper, we first describe our experimental framework based on a referential game between neural agents. We then analyze how compositionality is affected by object frequency in various scenarios. Finally, we examine how the frequency and compositionality evolve during training.

Our findings reveal several key insights. We establish that emergent communication systems mirror the irregular verbs phenomenon observed in natural languages, where messages referring to frequent objects develop less compositional structure than messages referring to rare ones (Pinker, 1999; Francis and Kučera, 1982). We also demonstrate that compositionality is not directly caused by frequency itself, but rather emerges as a response to the limited data exposure that frequency distributions naturally create.

Aligning with Kirby (2001)'s work, these results provide evidence that limited data exposure, also facilitated through frequency distributions, serves as the fundamental driver for compositional structure emergence in these systems, offering insights into the cognitive and computational pressures that shape linguistic systems.

2 Experimental Framework

We study emergent communication in the context of a Lewis signaling game (Lewis, 1969; Skyrms, 2010), where neural network agents need to communicate to complete a cooperative task.

2.1 The Lewis Reconstruction Game

Following a standard approach in emergent communication (Lazaridou et al., 2017; Havrylov and Titov, 2017), we implement a referential game between two agents: Speaker and Listener.

The game proceeds as follows: The Speaker receives an input object i and produces a message m. The Listener receives the message m and attempts to reconstruct the original input object, producing output $\hat{i}$. The agents succeed if $\hat{i} = i$, meaning the Listener correctly reconstructs the Speaker's input.

Each input object i is composed of A attributes, with each attribute taking one of N possible values. We represent each attribute as a one-hot vector of size N, and the full input i is the concatenation of these attribute vectors. For a given configuration with A attributes and N values per attribute, the total size of the input space is $|\mathcal{I}| = N^A$. While this experimental framework builds on established approaches in emergent communication (Rita et al., 2022; Michel et al., 2022), our work extends it to an unprecedented scale. In our experiments, we use two different input space configurations, both yielding approximately 1 million distinct objects: (1) objects with 6 attributes, each taking 10 possible values, and (2) objects with 3 attributes, each taking 100 possible values. This allows us to study how attribute structure affects emergent communication while keeping the total input space constant. Unlike Rita et al. (2022) who only utilized a small fraction (<1%) of their similarly sized dataset, our approach actively engages with the entire input space.

2.2 Agent Architecture

Both the Speaker and the Listener are implemented as recurrent neural networks (Elman, 1990) with layer normalization (Ba et al., 2016).

Speaker is a single-layer LSTM (Hochreiter and Schmidhuber, 1997) with a hidden size of 320 units. The input i is first passed through an embedding layer of dimension 20, and this embedding is used to initialize the hidden state of the Speaker's LSTM. The message m is generated symbol by symbol using the LSTM, with each new symbol sampled from a Categorical distribution over a vocabulary V of size 40. Messages can contain up to 30 symbols, including the EOS (end-of-sentence) token. The symbols are recursively sampled until the EOS token is produced or the Speaker reaches the maximum length. During testing, instead of sampling, each symbols are deterministically selected by tak-

ing the argmax of the distribution.

Listener is also a single-layer LSTM with a hidden size of 800 units, drawing inspiration from Chaabouni et al. (2022) who used asymmetric LSTM sizes with a larger Listener than Speaker. It receives the message m, and passes it through an embedding layer of dimension 200. These token embeddings are then fed one by one into the Listener's LSTM. After consuming the entire message, the final hidden state is passed through a linear layer followed by a softmax activation to produce a distribution for each attribute over possible values. During testing, the argmax of these distributions is used to produce the reconstruction candidate $\hat{i}$.

In this setup, the message space consists of sequences of length up to 30 with symbols from a vocabulary of size 40, including the EOS token. The message space size (approximately 10^{47}) therefore largely exceeds the input space size ($|\mathcal{I}| = 10^6$) as is standard in the literature (Rita et al., 2024).

2.3 Optimization

Training the agents presents a challenge because the communication channel is discrete, making the system non-differentiable end-to-end. We use a hybrid optimization approach:

For the Listener, we use standard cross-entropy loss, optimized with Adam (Kingma and Ba, 2014) and a learning rate of 0.001 with an entropy regularization coefficient of 0.2. This entropy regularization encourages exploration of different possible interpretations of messages during each retraining phase, which indirectly pressures the Speaker to develop clearer communication strategies.

For Speaker, which must produce discrete messages, we use the REINFORCE algorithm (Williams, 1992) with a batch normalization baseline to reduce variance (Sutton et al., 1999). We use log-based rewards and an entropy regularization coefficient of 1.0 to encourage exploration of the message space during early training. The Speaker is also optimized with Adam and a learning rate of 0.001.

To improve generalization and training stability, we implement the Early Stopping technique proposed by Rita et al. (2022) across all our experiments. While sharing characteristics with Iterated Learning Model (Ren et al., 2020; Li and Bowling, 2019; Cogswell et al., 2020), this technique implements a decoupled training paradigm wherein the speaker optimization proceeds independently, while the listener is reinitialized and trained from scratch until an early stopping criterion is met on a validation set.

We provide the decoupled training algorithm pseudocode and an ablation study of this technique in Appendix C. This study confirms enhanced convergence properties while demonstrating that our key findings about frequency and compositionality remain consistent regardless of whether listener resetting is employed.

The agents were trained for 250 training iterations, with 64 gradient steps and a batch size of 1024 for both agents.

We filtered for successful runs that converged to greater than 97% test accuracy.

Each training was performed using one NVIDIA V100 GPU for 12 hours.

For reproducibility, the code is available at `https://github.com/jbsevestre/frequency-and-compositionality`.

3 Analytical Method

3.1 Evaluation Metrics

We employ standard Emergent Communication metrics to evaluate our approach. All scores are averaged over 10 runs with different random seeds and evaluated on test splits.

3.1.1 Communication Success

We use accuracy as our primary measure of communication success between agents. Specifically, accuracy is calculated as the average number of correctly predicted attribute values, to give us a softer measurement of communication success.

3.1.2 Compositionality

To assess compositionality, we use Topographic Similarity (topsim) (Brighton and Kirby, 2006), which measures how well the structure of the meaning space is preserved in the message space. Topsim calculates the Spearman correlation between pairwise distances of objects (using Hamming distance) and their corresponding messages (using edit distance). Higher values indicate that similar objects receive similar messages, suggesting compositional structure. A score of 1.0 would indicate perfect preservation of meaning structure, while 0.0 indicates no systematic relationship.

3.2 Reference Distributions

To model frequency differences between attributes and objects, we sample attribute values from independent and identically distributed Zipfian distri-

butions, mimicking the frequency distribution of words in natural languages (Zipf, 2013).

For a given object's attribute, the marginal probability of sampling the k^{th} most frequent value is:

$$p(k) = \frac{k^{-\alpha}}{\sum_{l=1}^{N} l^{-\alpha}} \quad (1)$$

where k is the rank, α is the distribution parameter controlling skewness, and N is the number of values. As $\alpha \to 0$, $p(k)$ approaches a uniform distribution, and as $\alpha \to \infty$, probability concentrates entirely on the most frequent value ($k = 1$).

The objects' joint distribution is a multivariate Zipfian distribution, and is derived from the attributes' marginal distributions product since the marginals are i.i.d.. During training, batches of objects are sampled with respect to this joint distribution. This creates a scenario where some objects appear much more frequently than others during training, allowing us to examine how frequency affects the language properties. For comparative analysis, we also give results with uniform marginal distributions as a control group.

For simplicity, we label attribute values according to their frequency rank: 'value 1' is the most frequent value, 'value 2' is the second most frequent, and so on until 'value N,' which is the rarest.

3.3 The Hypercube Approach

To analyze compositionality between frequency groups, we propose a systematic approach for partitioning the object space.

As illustrated in Figure 2, we can group objects based on their attribute values: the "frequent" group contains cats with only common attributes (brown or white fur with green or blue eyes), while the "rare" group contains cats with only uncommon attributes (hairless or blue fur with orange or heterochromatic eyes). In this example, messages referring to cats in the "frequent" group tend to be simple and non-compositional ("the cat"), while messages for the "rare" group would typically combine multiple descriptive elements ("the cat with the blue fur and orange eyes"), suggesting higher compositionality in communications about rare objects. This grouping facilitates the comparison of the compositionality between frequency groups.

We generalize this approach to object spaces with more attributes and values by partitioning it into N non-overlapping hypercubes:

$$\mathrm{H}_i = \mathrm{H}\left[(i-1) \times \frac{N}{C} + 1, i \times \frac{N}{C}\right] \quad (2)$$

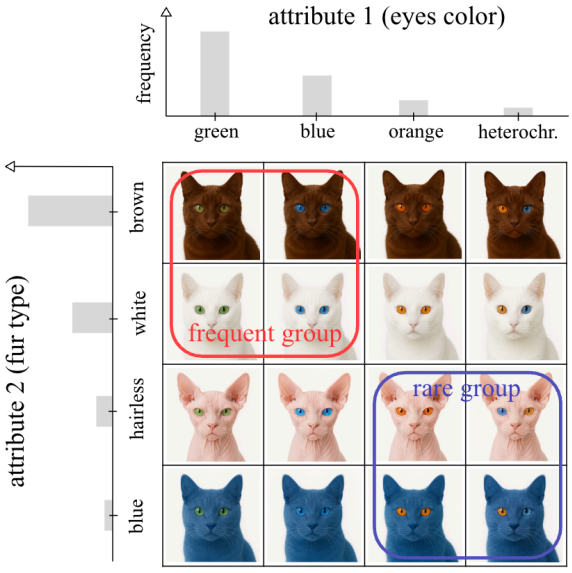

Figure 2: Hypercube approach example in a 2-attributes and 4-values setting. In the frequent group, animals are more likely to be qualified as just "the cat", when in the rare group, animals are more likely to be described with a concatenation of compositional adjectives: "the cat with the blue fur and the orange eyes".

and compute all metrics for objects within each hypercube separately, where i is the hypercube number, N is the number of values, and C is the number of hypercubes. Note that we deliberately use non-overlapping hypercubes to ensure statistical independence between frequency groups.

For our experiments, in the 6-attributes & 10-values setting, we split the object space into 2 hypercubes, each containing $15,625$ elements: the "frequent" group $\mathrm{H}_1 = \mathrm{H}[1,5]$ contains objects whose attributes only have the most frequent values (1 to 5), and the "rare" group $\mathrm{H}_2 = \mathrm{H}[6,10]$ contains objects whose attributes have the less frequent values (6 to 10). In the 3-attributes & 100-values setting, we split the object space into 10 non-overlapping hypercubes, each containing $1,000$ elements: from $\mathrm{H}_1 = \mathrm{H}[1,10]$: the "hyper-frequent" group containing objects with attribute's values between 1 and 10 to $\mathrm{H}_{10} = \mathrm{H}[91,100]$: the "hyper-rare" group containing objects with the most rare attribute's values between 91 and 100, offering a gradient of frequency between this two groups.

3.4 Random Block Split

In order to partition the full object space dataset into non-overlapping train, validation, and test sets, one key challenge is to preserve the frequency characteristics of our skewed joint distribution com-

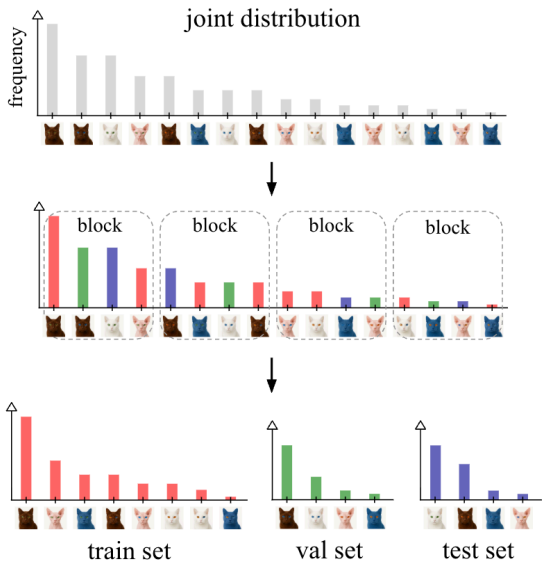

Figure 3: Random Block Split. Top: objects' joint distribution from attribute Zipfian marginals. Middle: division into blocks of consecutive values and random proportional allocation per block. Bottom: resulting train, val, and test sets: this approach preserves the original distribution characteristics.

puted from attribute Zipfian marginals.

Note that in this dataset, elements are unique (objects are distributional modalities), ordered (in our case, by frequency rank), and the distribution characteristics (frequencies) are known.

To tackle this, we propose Random Block Split (Algorithm 1), an approach specifically designed to split an ordered dataset while preserving its frequency characteristics.

The method operates on blocks of consecutive values, with block size determined by the desired split ratio. The Random Block Split algorithm works as follows: Given a dataset D of n ordered elements and a split ratio (r_{train}:r_{val}:r_{test}), it first determines the block size $b = r_{\text{train}} + r_{\text{val}} + r_{\text{test}}$. Then, for each successive block of b elements in D, it randomly assigns r_{train} elements to the training set, r_{val} elements to the validation set, and the remaining r_{test} elements to the test set. This process continues until the entire dataset is partitioned.

For our experiments, we use a 2/1/1 split ratio (train/validation/test), resulting in a 0.5M train set, 0.25M validation set, and 0.25M test set from the original 1M object dataset.

As exemplified in Figure 3, this approach successfully preserves the multivariate Zipfian distribution characteristics across the resulting datasets.

Algorithm 1 Random Block Split

From dataset D of n ordered elements
Define set ratio $r_{\text{train}} : r_{\text{val}} : r_{\text{test}}$
Set block size $b = r_{\text{train}} + r_{\text{val}} + r_{\text{test}}$
for each successive block of b elements of D **do**
  Randomly assign r_{train} elements to T_{train}
  Randomly assign r_{val} elements to T_{val}
  Assign remaining r_{test} elements to T_{test}
end for
Return T_{train}, T_{val}, T_{test}

4 Results

We present a series of experiments that investigate the relationship between frequency and compositionality in emergent communication systems.

4.1 Irregular verbs phenomenon

Our first experiment examines communication performance and compositionality across frequent and rare object spaces under both uniform and Zipfian distributions, as shown in Figure 4.

Under uniform distributions, accuracy and topographic similarity remain relatively constant across both frequent and rare object spaces (approximately $0.28 - 0.29$ for topsim), as expected from a control group where no frequency differences exist. On the contrary, for Zipfian distributions, we observe that frequent objects maintain high performance while rare objects show slightly decreased accuracy with greater variability. This accuracy drop is expected, as agents naturally struggle to learn rare attribute values that appear less frequently in the training data. We also observe that compositional structure (measured via topographic similarity) is significantly lower for frequent objects (median approximately 0.28) compared to rare objects (median approximately 0.38). Indeed, the p-value of the unilateral Wilcoxon signed-ranked test (Wilcoxon, 1945), testing if the compositionality on the rare group is significantly higher than the one on the frequent group, is statistically significant ($p < 0.01$), while the same test confirms no significant difference between the two groups under the uniform distribution ($p > 0.05$).

These findings reflect the irregular verbs phenomenon in emergent communication: frequent items develop less compositional structure than rare items, analogous to how frequently used verbs in natural languages often maintain irregular forms

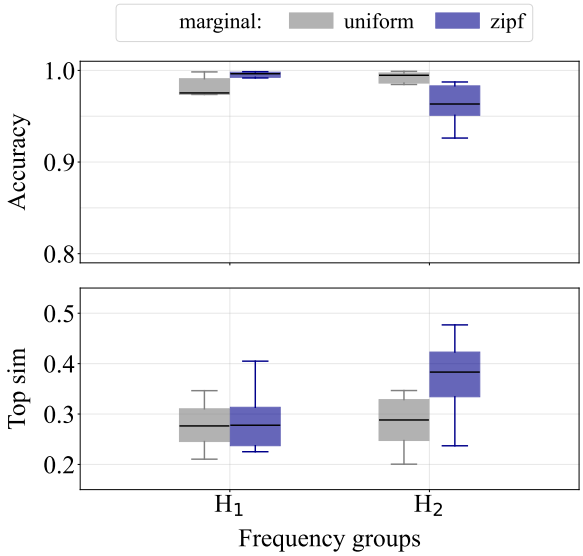

Figure 4: Topographic similarity evaluated from runs on the 6-attributes and 10-values setting. H_1 and H_2 correspond respectively to the frequent and the rare groups.

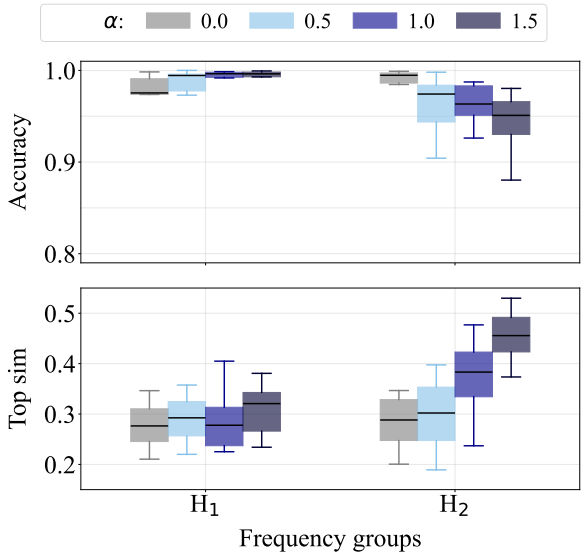

Figure 5: Topographic similarity evaluated from runs on the 6-attributes and 10-values setting for different values of the Zipf's alpha parameters. H_1 and H_2 correspond respectively to the frequent and the rare groups.

(Pinker, 1999; Wu et al., 2019).

To test how message length compression affects the compositional structure differences observed between frequent and rare objects, we conducted additional experiments using the LazImpa mechanism (Rita et al., 2020), which explicitly encourages shorter messages. As detailed in Appendix A, our analysis shows that the irregular verbs phenomenon is independent of the message length.

For comparison, we established a baseline using randomly generated messages. As detailed in Appendix B, this random baseline yields topographic similarity values of approximately 0.0 and chance-level accuracy (0.1).

4.2 Impact of Zipf Parameter α

Our second experiment investigates how varying the skewness of the Zipfian distribution through different α parameters affects communication performance and compositional structure (Figure 5).

The results reveal a systematic relationship between the strength of the Zipfian skew (α) and both communication accuracy and compositional structure. As α increases from 0.0 (uniform) to 1.5 (highly skewed), we observe distinct patterns for each object space.

For frequent objects, accuracy remains consistently high (approximately 99%) across all α values, while topographic similarity shows only modest increases (from approximately 0.28 at $\alpha = 0.0$ to approximately 0.32 at $\alpha = 1.5$). For rare objects, accuracy progressively decreases with increasing α (from approximately 0.98 at $\alpha = 0.0$ to approximately 0.94 at $\alpha = 1.5$), while topographic similarity increases substantially (from approximately 0.29 at $\alpha = 0.0$ to approximately 0.45 at $\alpha = 1.5$).

The effect is particularly pronounced for rare objects, where the highest α (1.5) yields the lowest accuracy but the highest topographic similarity.

As the distribution becomes more skewed with higher α values, rare objects become even less frequent in the training data, creating greater challenges for accurate communication. In response, agents develop increasingly compositional language for these rare objects, which helps them generalize from limited exposure.

4.3 Hyper-frequent group

Our third experiment employs a more fine-grained approach, partitioning the object space into 10 hypercubes (H_1-H_{10}) ranging from extremely frequent (H_1) to extremely rare (H_{10}) (Figure 6).

The results reveal nuanced patterns across the frequency spectrum. For accuracy, under uniform distributions, performance remains consistently high across all hypercubes. Under Zipfian distributions, we observe a clear frequency gradient: accuracy decreases progressively from H1 (hyper-frequent) to H_{10} (hyper-rare), with increasing variability for rarer hypercubes. For topographic simi-

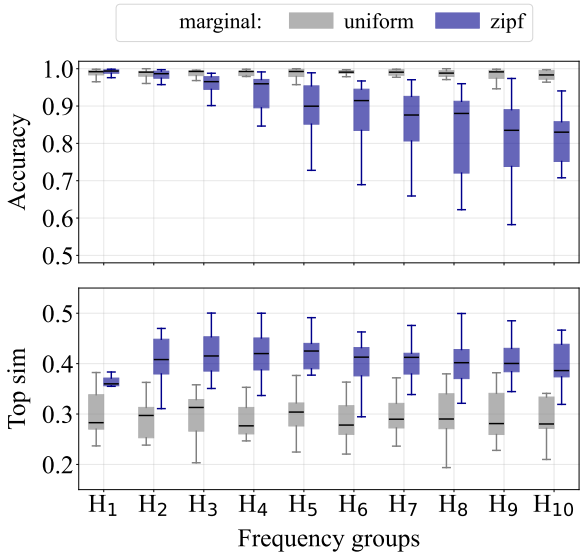

Figure 6: Topographic similarity evaluated from runs on the 3-attributes and 100-values settings. Frequency groups are arranged from most frequent (H_1, leftmost) to least frequent (H_{10}, rightmost).

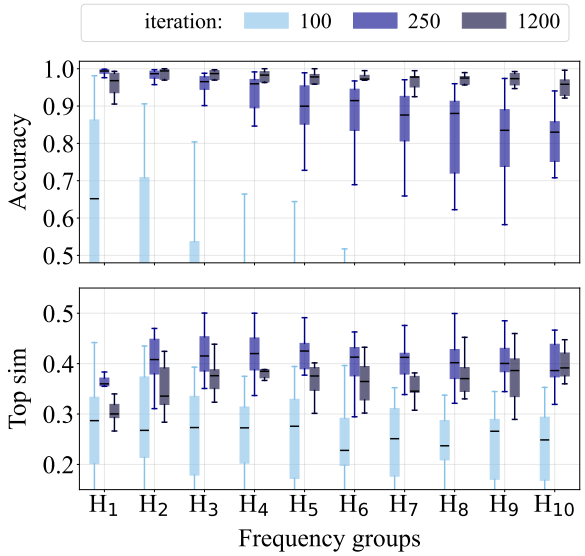

Figure 7: Topographic similarity evaluated from runs on the 3-attributes and 100-values settings for different numbers of iterations. Frequency groups are arranged from most frequent (H_1, leftmost) to least frequent (H_{10}, rightmost).

larity, under uniform distributions, compositional structure remains remarkably stable (approximately $0.28 - 0.30$) across all hypercubes. Under Zipfian distributions, we observe two distinct patterns: (1) the hyper-frequent hypercube H_1 shows the lowest topographic similarity (approximately 0.35), and (2) hypercubes H_2-H_{10} maintain consistently higher topographic similarity levels (approximately $0.39 - 0.42$) with no frequency-based gradient. Statistical analysis provided in Appendix D confirms the distinctly lower topographic similarity of H_1 compared to all other groups under Zipfian distributions (unilateral Wilcoxon signed-ranked test, $p < 0.01$) as shown in the p-value matrix (Fig 11).

These findings strongly confirm the irregular verbs phenomenon: while the hyper-frequent objects (H_1) indeed show reduced compositional structure, the remaining object space exhibits remarkably consistent levels of compositionality despite decreasing frequency and accuracy.

While these consistent compositional levels could potentially indicate attribute independence and a common compositional framework across frequency groups, confirming whether agents employ shared encoding strategies would require qualitative message analysis that extends beyond our current topographic similarity measures.

4.4 Compositionality throught training

Our final experiment examines how communication systems evolve over different training durations (100, 250, and 1200 iterations), illustrated in Figure 7. Note that the mid-training stage (250 iterations) represents the same training duration used in our previous experiments with Zipfian distributions.

The results reveal a complex relationship between training duration, communication accuracy, and compositional structure: during early training (100 iterations), accuracy varies dramatically across hypercubes: high for frequent items (approximately $0.50 - 0.65$ for H_1-H_2) but poor for rare items (dropping to approximately 0.10 for H_{10}), while topographic similarity remains relatively low across all hypercubes (approximately $0.28 - 0.30$). As training progresses to mid-stage (250 iterations), accuracy improves substantially across all hypercubes (approximately $0.83 - 0.99$), though still showing a frequency gradient, and topographic similarity reaches its peak values (approximately $0.35 - 0.45$), with the highest values observed in hypercubes H_3-H_5. With extended training (1200 iterations), accuracy approaches ceiling performance across all hypercubes (approximately $0.95 - 1.0$), with minimal frequency effects; interestingly, topographic similarity decreases from the 250-iteration levels across all hypercubes (approxi-

mately $0.30 - 0.40$), though remaining above the 100-iteration values.

These dynamics reveal a two-phase development process in emergent communication systems.

First, a *Systematization Phase* (up to approximately 250 iterations): This phase is characterized by rapidly improving accuracy and increasing topographic similarity. Agents initially develop increasingly compositional language to improve communication accuracy, particularly for rare objects.

Later, a *Conventionalization Phase* (beyond 250 iterations): During this stage, while accuracy keeps increasing, topographic similarity starts decreasing across most of the frequency groups. With extended training, moderately frequent objects gain sufficient exposure for agents to develop conventional forms that may depart from compositional structure.

The observed two-phase development process resembles the historical evolution of language forms in natural languages, where initially descriptive and compositional terms often give way to conventional alternatives once they become established through repeated use (consider how 'electronic mail' evolved into 'email').

As training progresses to this later stage (1200 iterations), Zipf-based runs achieve comparable accuracy to uniform-based ones across all frequency groups while matching the same compositionality on the frequent groups. Notably, compositional differences between hyperfrequent and other groups become blurred, and the irregular verbs phenomenon is no longer as pronounced, suggesting that with sufficient training, the system may eventually converge toward more uniform compositional structures regardless of initial frequency distributions. This observation reveals that compositionality is not inherently tied to frequency itself, but rather to the exposure to the data during learning.

While not mandatory, frequency significantly facilitated this discovery by creating a natural imbalance in the learning environment, providing important exposure to frequent objects from which agents could infer rules and structure, alongside limited exposure to rare objects that incentivized generalization. Indeed, as detailed in Appendix E, frequency creates a situation where a significant portion of the training set remains unseen during training, blurring the clear distinction between train and test set, and forcing agents to develop generalizable communication strategies over their own training set. This experimental approach mirrors the conditions under which natural languages evolve, where frequency creates similar learning pressures.

5 Conclusion

This work extends Kirby (2001)'s research on frequency effects and confirms his finding, where frequent elements tend to develop less compositional structure than rare elements, mirroring patterns observed in natural languages (Pinker, 1999). Through our experiments at scale with neural agents, we provide quantitative evidence that the irregular verbs phenomenon exists in artificial communication systems.

Beyond confirming this phenomenon, our analysis reveals that compositionality is not an inherent property of frequency itself but rather emerges as a consequence of limited data exposure. Though not a requirement, frequency helps create a natural imbalance between frequent and rare elements, where rarity pushes toward the development of a compositional system.

Consequently, these findings provide compelling evidence that limited data exposure, which frequency distributions naturally create, drives the emergence of compositional structure in communication systems. This mechanism may help explain why human languages universally exhibit compositional properties, as they too must enable effective communication about rare or novel concepts with limited exposure.

6 Limitations

While this study provides significant insights into frequency-compositionality relationships in emergent communication, two important limitations should be acknowledged. First, our input space uses independent attributes, creating an artificial simplification of real-world objects where features are typically correlated and interdependent. This idealized structure may favor compositional strategies that wouldn't emerge with more naturalistic inputs. Future research should investigate how emergent communication adapts to input spaces with complex feature correlations and hierarchical structures. Secondly, by focusing primarily on topographic similarity metrics rather than analyzing message content and structure, we potentially overlook important qualitative aspects of the communication system. A deeper linguistic analysis of the emergent codes might reveal more nuanced patterns in how frequency affects not just the de-

gree of compositionality, but the specific encoding strategies agents develop for different attributes.

7 Ethics Statement

Our research exclusively analyses synthetic data posing minimal immediate ethical concerns. However, we acknowledge that advances in emergent communication systems could eventually contribute to the development of large-scale multi-agent systems with potential applications in defense contexts.

Acknowledgments

We would like to greatly thank Mathieu Rita for his insightful feedback and reviews. This work was granted access to the HPC resources of IDRIS under the allocation 2024-AD011015895 made by GENCI, and was supported in part by the Agence Nationale pour la Recherche (ANR-17-EURE-0017 Frontcog, ANR10-IDEX-0001-02 PSL*). ED in his EHESS role was funded by an ERC grant (InfantSimulator). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. (arXiv:1607.06450).

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

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

Joan Bybee. 2007. *Frequency of Use and the Organization of Language*. Oxford University Press, USA.

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*, page 4427–4442. 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 Altché, 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*.

Michael Cogswell, Jiasen Lu, Stefan Lee, Devi Parikh, and Dhruv Batra. 2020. Emergence of compositional language with deep generational transmission. (arXiv:1904.09067).

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Nelson Francis and Henry Kučera. 1982. *Frequency analysis of English usage: Lexicon and grammar*. Houghton Mifflin, Boston.

Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Charles F Hockett. 1960. The origin of speech. *Scientific American*, 203(3):88–97.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. (arXiv:1412.6980).

S. Kirby. 2001. Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110.

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. Association for Computational Linguistics.

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

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

David Kellogg Lewis. 1969. *Convention: A Philosophical Study*. Wiley-Blackwell.

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.

H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Paul Michel, Mathieu Rita, Kory Wallace Mathewson, Olivier Tieleman, and Angeliki Lazaridou. 2022. Revisiting populations in multi-agent communication.

Steven Pinker. 1999. *Words and rules: The ingredients of language*. Basic Books.

Yi Ren, Shangmin Guo, Matthieu Labeau, Shay B. Cohen, and Simon Kirby. 2020. Compositional languages emerge in a neural iterated learning model.

Mathieu Rita, Rahma Chaabouni, and Emmanuel Dupoux. 2020. "lazimpa": Lazy and impatient neural agents learn to communicate efficiently. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, page 335–343. Association for Computational Linguistics.

Mathieu Rita, Paul Michel, Rahma Chaabouni, Olivier Pietquin, Emmanuel Dupoux, and Florian Strub. 2024. Language evolution with deep learning.

Mathieu Rita, Corentin Tallec, Paul Michel, Jean-Bastien Grill, Olivier Pietquin, Emmanuel Dupoux, and Florian Strub. 2022. Emergent communication: Generalization and overfitting in lewis games. *Advances in Neural Information Processing Systems*, 35:1389–1404.

Brian Skyrms. 2010. *Signals: Evolution, learning, and information*. OUP Oxford.

Kenny Smith, Simon Kirby, and Henry Brighton. 2003. Iterated learning: A framework for the emergence of language. *Artificial life*, 9(4):371–386.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

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

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press.

George Kingsley Zipf. 2013. *The psycho-biology of language: An introduction to dynamic philology*. Routledge.

A Message compression effect

We investigate whether message length compression affects the compositional structure differences observed between frequent and rare objects. Some researchers have proposed that structure loss in frequent messages might be due to a compression effect similar to how natural language phrases like "I do not know" become compressed to "don't know".

To test this hypothesis, we implemented the LazImpa mechanism from Rita et al. (2020). LazImpa introduces a novel communication system where a speaker agent is constrained by progressive "laziness" (implementing a scheduled length penalty) while a listener agent operates under "impatience" constraints (predicting message content incrementally at each symbol), jointly addressing neural networks' inherent bias toward verbose communication and enabling the emergence of near-optimal codes that exhibit human language-like efficiency conforming to Zipf's Law of Abbreviation.

Figure 8 compares topographic similarity and message length between systems with and without the LazImpa mechanism. The results show that while LazImpa successfully induces shorter messages as we could expect (approximately 14 vs. 30 characters for frequent objects, and 21 vs. 30 characters for rare objects), the compositional difference pattern between frequent and rare groups remains remarkably consistent.

Note that the Mann-Whitney test (Mann and Whitney, 1947), testing if the compositionality with LazImpa is significantly different than without it, confirms no significant difference in the compositionality gap between the two systems. Note also that even if not shown in the Figure, the accuracy is the same between the two settings for the two groups, matching the one shown in Figure 4.

These findings are particularly revealing because they demonstrate that the irregular verbs phenomenon persists independently of message length and is not accentuated with compression. This suggests that the frequency effect dominates compression in the structure loss phenomenon, and since frequency is just a way to create a natural limited data exposure, it reinforces the importance of the learning bottleneck in the emergence and loss of language's compositionality.

B Random messages baseline

To establish a proper baseline for comparison, we evaluate both topographic similarity and accuracy

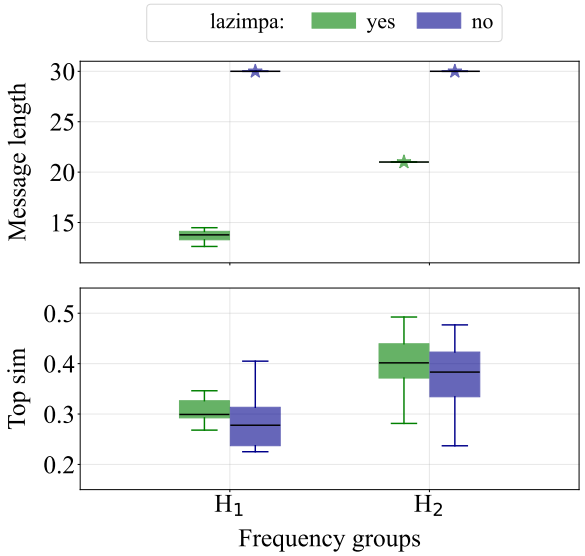

Figure 8: Topographic similarity evaluated from runs on the 6-attributes and 10-values settings for standard and LazImpa message compression variants. H_1 and H_2 correspond respectively to the frequent and the rare groups.

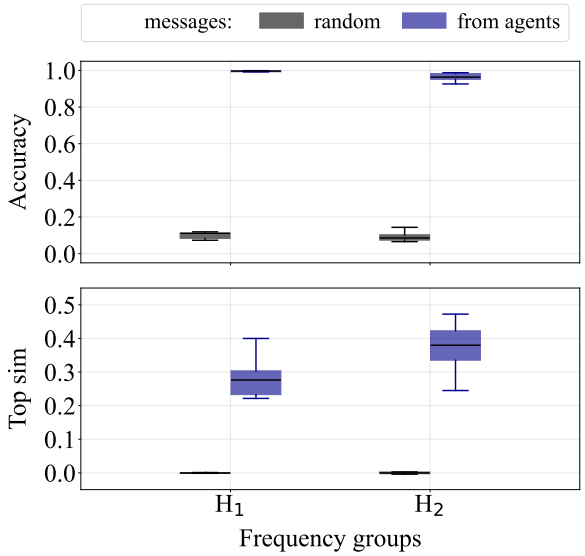

Figure 9: Topographic similarity evaluated from runs on the 6-attributes and 10-values settings for messages generated with random tokens and from trained agents. H_1 and H_2 correspond respectively to the frequent and the rare groups.

of a randomly generated message. For each object in the object space, we generate a random message by sampling each token uniformly from the entire vocabulary. Message generation stops when either the EOS token is sampled or when the maximum message length limit is reached.

Compositional measures are then computed directly from these object-message pairs. To assess accuracy, we provide these random messages to a randomly initialized and untrained listener agent. The messages produced by trained agents in our main experiments serve as the regular condition for comparison.

As shown in Figure 9, the topographic similarity of the random messages approximates 0.0, confirming that randomly generated messages lack compositional structure. The accuracy measurement yields approximately 0.1, which aligns with theoretical expectations since accuracy is calculated as the average number of correctly guessed attribute values. In our 6-attributes & 10-values setting, a listener would have a 10% chance of correctly predicting the right attribute value by random guessing.

Since topographic similarity and accuracy of the random baseline remain consistent across all experimental conditions and to maintain clarity in our visualizations, we have omitted these baseline measures from the plots in the core of the article.

C Listener reset benefits

Following the methodology proposed by Rita et al. (2022), we implement an "Early Stopping" technique wherein the listener is reset and retrained from scratch until an early stopping criterion is met on a validation set. The complete alternating training procedure is formalized in Algorithm 2.

This approach yields two significant benefits for our experimental framework.

First, as illustrated in the right panel of Figure 10, this technique substantially enhances the convergence properties of the training process. By employing listener resetting, all experimental runs converge within fewer iterations, allowing us to avoid selective reporting of only successful runs that could introduce methodological bias in our analysis.

Second, as shown on the left side of Figure 10, resetting the listener increases the learning bottleneck pressure since the listener is exposed to fewer data during its limited lifetime. This constraint pushes the system toward the development of stronger compositional structure. As a consequence, for the same pattern of accuracy decrease along the hypercube frequency gradient, the compositional measures are consistently higher across all hypercubes when using listener reset compared to experiments without this technique.

It is important to emphasize that all phenomena

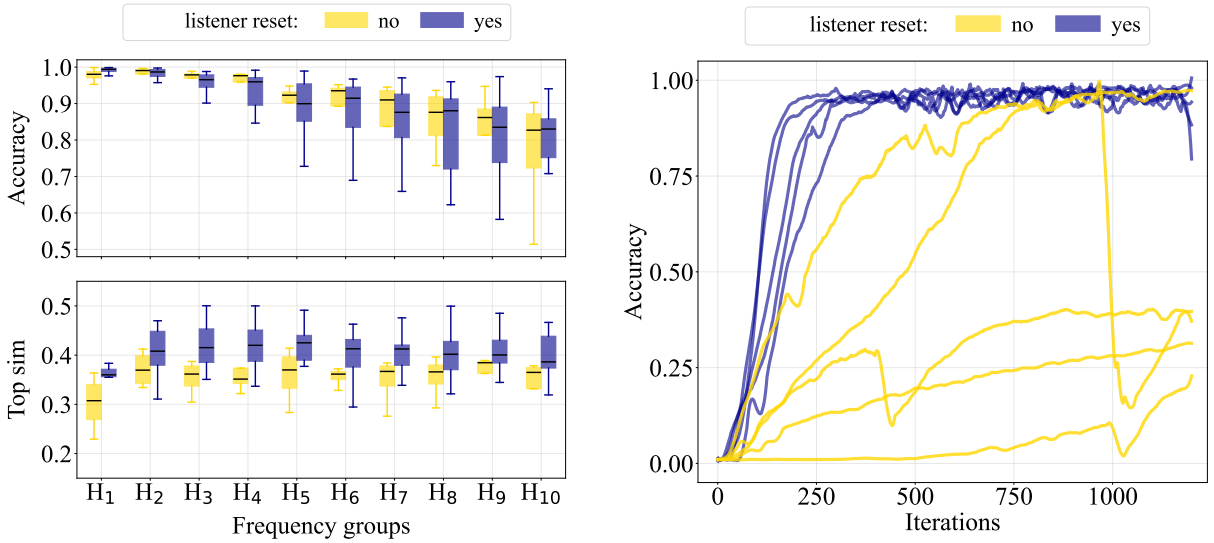

Figure 10: Left: Topographic similarity evaluated from runs on the 3-attributes 100-values settings with Zipfian marginals for algorithm versions with and without listener reset, respectively at 250 and 750 iterations. Frequency groups are arranged from most frequent (H_1, leftmost) to least frequent (H_{10}, rightmost). Right: Algorithm performance over training iterations. Listener reset improves the properties of the convergence.

Algorithm 2 Early stopping algorithm

for `NUM_ITERATIONS` **do**

// Stage 1: Listener training
Reset $\theta_{listener}$
Freeze $\theta_{speaker}$
while not stopping_criterion **do**
for `NUM_STEPS` **do**
Sample one batch from $\mathcal{D}_{train} \sim P_{joint}$
Make one forward pass through the netwrks
Perform one gradient descent step
end for
Evaluate loss on $\mathcal{D}_{validation}$
Update stopping_criterion with validation loss
end while

// Stage 2: Speaker training
Freeze $\theta_{listener}$
Unfreeze $\theta_{speaker}$
for `NUM_STEPS` **do**
Sample one batch from $\mathcal{D}_{train} \sim P_{joint}$
Make one forward pass through the networks
Perform one gradient descent step
end for

end for

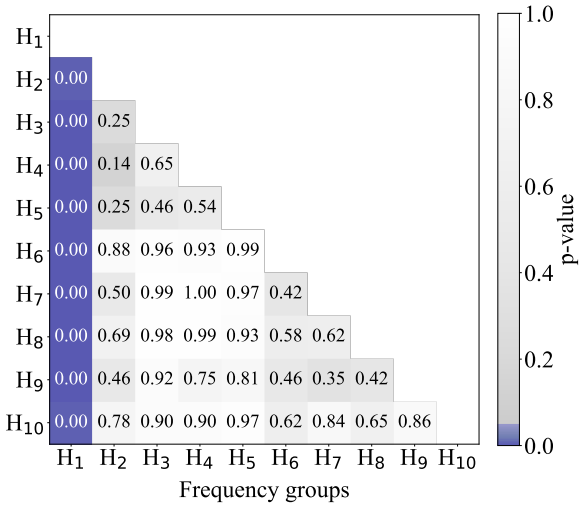

Figure 11: p-values of the unilateral Wilcoxon signed-rank test evaluated from runs on the 3-attributes and 100-values settings, pair-wisely testing if the compositionality on a group in a column is significantly smaller than one in a row. Frequency groups are arranged from most frequent (H_1, leftmost) to least frequent (H_{10}, rightmost).

documented in this paper are robust to variations in training technique. While listener reset enhances compositionality, the irregular verbs phenomenon emerges independently of this technique, as clearly demonstrated in the left panel of Figure 10, where the compositinality on the hyperfrequent group H_1 is significantly lower than on the other groups, regardless of whether listener reset is employed.

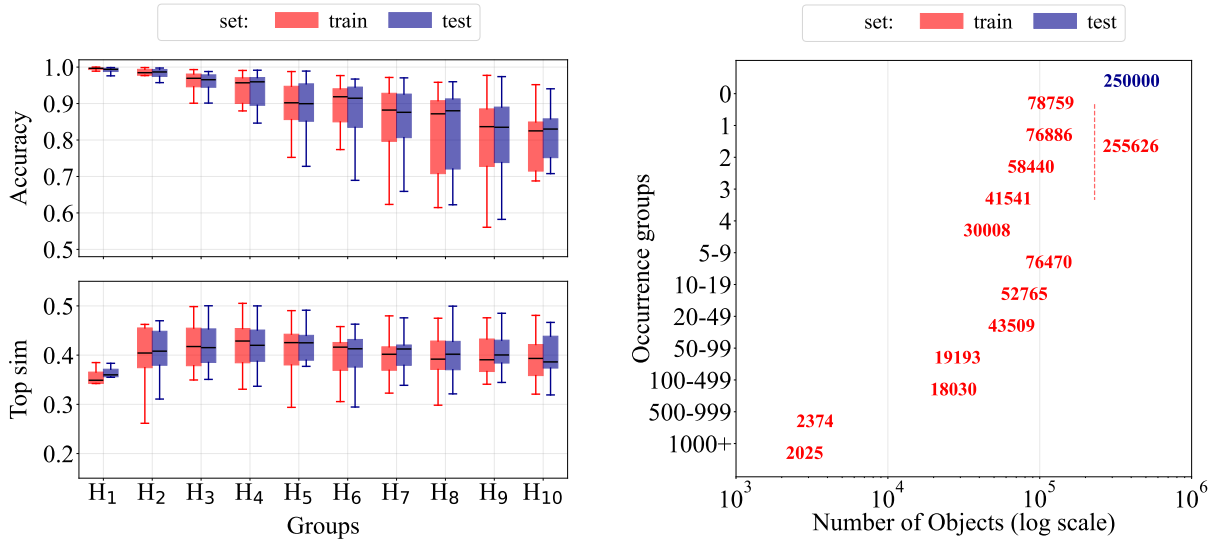

Figure 12: Left: Topographic similarity evaluated from runs on the 3-attributes and 100-values settings with Zipf marginals on train and test set. Frequency groups are arranged from most frequent (H_1, leftmost) to least frequent (H_{10}, rightmost). Right: average number of times each data has been seen in the speaker training loop for the related runs

D Pair-wise statistical analysis

To rigorously assess the statistical significance of compositional differences between frequency groups, we conducted pairwise comparisons using the unilateral Wilcoxon signed-rank test (Wilcoxon, 1945). Figure 11 presents the p-values resulting from these tests, examining whether the compositionality of hypercubes in columns is significantly smaller than those in rows.

The matrix reveals several key patterns. First, the consistently low p-values (0.00) in the H_1 column demonstrate that the hyper-frequent hypercube (H_1) exhibits significantly lower compositionality than all other hypercubes, confirming our irregular verbs phenomenon observation with strong statistical evidence. Second, the generally high p-values (> 0.40) in comparisons between mid-frequency hypercubes (H_2-H_{10}) indicate no statistically significant differences in compositionality among these groups, suggesting that compositional structure remains relatively consistent across the non-hyper-frequent portions of the object space.

This statistical analysis supports our conclusion that frequency effects on compositionality primarily manifest at the extreme high-frequency end of the distribution, with the most frequent objects developing distinct, less compositional communication patterns while the remainder of the object space maintains a relatively uniform level of compositionality.

E Train and test performance similarity

As described in Section 3.4, we split our dataset into non-overlapping train, validation, and test sets.

By construction, the test set is only composed of never-seen objects with unknown combinations of values.

However, in our experiments, as we observe in Figure 12, we note that there is only a tiny difference between train and test performances.

On the right side of Figure 12, at the time of convergence (250 iterations) on the speaker side, we see that 15% (78,759) of the training data will never be seen, as a direct consequence of frequency: while frequent elements are overly sampled (seen 1000+ times for the 2025 most frequent), the rarest are barely sampled once. Furthermore, more than 50% (255,626) of them will be seen 3 times or less (not influencing model parameters too much): that amount is greater than the data the test set contains (250,000), meaning that the boundary between the training and test sets is becoming blurred.

Agents couldn't memorize the entire training set since they're only exposed to a part of it, and have to generalize over their own training set in order to reduce their error on rare or even new upcoming training data: this learning bottleneck (Kirby, 2001) pushes towards the emergence of a compositional language and explains why accuracy and topographic similarity on training and test set is similar.