

# UC Merced

## Proceedings of the Annual Meeting of the Cognitive Science Society

### Title

Modeling Human-Like Color Naming Behavior in Context

### Permalink

<https://escholarship.org/uc/item/2040v6w5>

### Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

### Authors

Zhang, Yuqing

Ürker, Ecesu

Verhoef, Tessa

et al.

### Publication Date

2026

### Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

# Modeling Human-Like Color Naming Behavior in Context

Yuqing Zhang<sup>1</sup>, Ecesu Ürker<sup>2</sup>, Tessa Verhoef<sup>3</sup>, Gemma Boleda<sup>4</sup> & Arianna Bisazza<sup>1</sup>

<sup>1</sup>Center for Language and Cognition, University of Groningen

<sup>2</sup>Department of Translation and Language Sciences, Universitat Pompeu Fabra

<sup>3</sup>Leiden Institute of Advanced Computer Science, Leiden University

<sup>4</sup>Department of Translation and Language Sciences, Universitat Pompeu Fabra / ICREA

## Abstract

Modeling the emergence of human-like lexicons in computational systems has advanced through the use of interacting neural agents, which simulate both learning and communicative pressures. The NeLLCom-Lex framework (Zhang et al., 2025) allows neural agents to develop pragmatic color naming behavior and human-like lexicons through supervised learning (SL) from human data and reinforcement learning (RL) in referential games. Despite these successes, the lexicons that emerge diverge systematically from human color categories, producing highly non-convex regions in color space, which contrast with the convexity typical of human categories. To address this, we introduce two factors, upsampling rare color terms during SL and multi-listener RL interactions, and adopt a convexity measure to quantify geometric coherence. We find that upsampling improves lexical diversity and system-level informativeness of the color lexicon, while many-listener setups promote more convex color categories. The combination of moderate upsampling and multiple listeners produces lexicons most similar to human systems.

**Keywords:** lexical system; language efficiency; pragmatics; neural agents; referential game

## Introduction

The relationship between the use of a language and its properties as a linguistic system is highly dynamic, such that language use shapes language systems and vice versa (Beckner et al., 2009; Brochhagen & Boleda, 2022; Clark, 1996; Hawkins, 2004; Regier et al., 2015). Studying this interaction is challenging, as it emerges from system-level dynamics among large populations and unfolds over extended historical periods (Campbell, 2013; Hopper & Traugott, 2003). Computational modeling offers a powerful approach, enabling simulation of system-level pressures in a controlled way while varying parameters inaccessible in experiments with humans or field work (Brochhagen, 2018; Cangelosi & Parisi, 2002; De Boer, 2006; Steels, 1997; Steels & McIntyre, 1998). Recent advances in deep learning have substantially expanded what is possible in this domain through powerful and flexible neural-agent models (Chaabouni et al., 2021; Kharitonov et al., 2019; Lazaridou & Baroni, 2020; Lian et al., 2023).

In this paper, we use this method to investigate the relationship between language use and language system in the domain of the lexicon. We adopt the recently introduced NeLLCom-Lex framework (Zhang et al., 2025), aimed at analyzing how communicative factors shape lexicons by isolating interactional pressures from population-level transmission

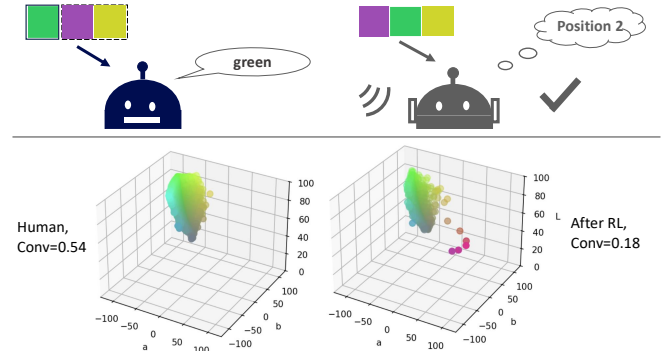


Figure 1: Agents develop non-convex color representations as a result of playing referential games. The top row illustrates the referential game setup, in which a target color chip (outlined with a solid line) is presented among distractors (dashed line). The bottom row compares human and agent color representations for the color term *green*.

dynamics. After first training agents on human data from a color discrimination task, we study how communicative use reshapes an already grounded lexicon. In this framework, agents are first pretrained via SL and subsequently interact through RL in referential games, as illustrated in Figure 1 (top). During these interactions, the speaker and the listener agents are presented with a target and two distractor color chips. The speaker generates a color term to convey the target to the listener (e.g. *green* in this example) and the listener agent selects a color chip based on the speaker's message. In Zhang et al. (2025), agents trained in this framework were shown to develop context-sensitive pragmatic behavior.

Despite these successes, the lexicons that emerge from NeLLCom-Lex agents show systematic divergences from human lexicons; in particular, they establish color categories that are often highly non-convex in color space, and agents seem to come up with partner-specific "conceptual pacts" (Brennan & Clark, 1996) (see Figure 1, bottom, for an illustration). Although such systems can support effective communication among agents, these properties contrast with well-established characteristics of human color categories, which typically correspond to convex regions in perceptual space (Gärdenfors, 2014; Jäger, 2007, 2009). This constraint, known as the *convexity universal*, states that the denotation of every color term is a convex region in the CIELAB color space that ap-

proximates human color vision (Brainard, 2003). Gärdenfors (2000, 2014) has argued more generally that convexity is a hallmark of human categories, and hence of word meaning.

If we are to use neural agents to model the dynamics between word use and lexical systems, we need to ensure that their lexicons exhibit key properties of human lexicons. Since convexity is one of the necessary characteristics of human lexical categories (Carlsson et al., 2024), this motivates a closer examination of the conditions that give rise to non-convex denotations in the first place. In this paper, we investigate factors that are crucial for convexity to arise, and focus on two factors that shape the geometry of emergent lexicons but were not central in prior analyses: sparsity of exposure to some color terms, and listener diversity. The listener-diversity manipulation is motivated by the hypothesis that greater interlocutor variety makes conceptual pacts more difficult to establish, thereby promoting more convex semantic categories. An alternative explanation, however, is that non-convexity arises from limited exposure to certain meanings, resulting in poorly learned denotations. Although this account is unlikely given that even high-frequency terms exhibit non-convexity. We therefore introduce upsampling to control for data sparsity, allowing us to disentangle the effects of listener diversity from those of general learning difficulty. Crucially, we also track other characteristics of human lexicons that have been described in the literature, including vocabulary size and system informativeness (Gualdoni & Boleda, 2024; Regier et al., 2015). This allows us to examine how convexity relates to these properties: for example, a one-word lexical system would be fully convex. We find that upsampling rare color terms during SL increases lexical diversity and informativeness while reducing semantic drift, whereas interacting with many listeners constrains vocabulary size and informativeness but promotes a more convex color space. The combination of moderate upsampling and a many-listener setup produces lexicons that are most human-like across multiple metrics, highlighting the importance of data exposure and population structure in shaping emergent lexical systems<sup>1</sup>.

## Related Work

### Emergent Communication in the Color Domain

Drawing on the World Color Survey (WCS) (Berlin & Kay, 1991; Cook et al., 2005), the color domain has become a widely used test case for agent-based models of language emergence. Early studies demonstrated that artificial agents (Steels, Belpaeme, et al., 2005) and even embodied robots (Bleys et al., 2009) could develop color categories from scratch through repeated language games. Subsequent work using similar language game paradigms produced increasingly human-like color naming systems by incorporating simple perceptual constraints (Baronchelli et al., 2010; Loreto et al., 2012). Comparable results have been achieved with more recent neural deep-learning agents that learn from inter-

actions through RL (Carlsson et al., 2024; Chaabouni et al., 2021; Kågebäck et al., 2020; Tucker et al., 2025), building on the information-theoretic model of Zaslavsky et al. (2018). In these set-ups, agents develop their own color naming systems from scratch. In contrast, NeLLCom-Lex (Zhang et al., 2025) initializes agents with a human color naming system to enable more interpretable analyses, and allow for more focused experiments on realistic scenarios of semantic change.

**Color Naming in Context** Prior approaches mainly model color lexicons as a partition of the perceptual color space, with a unique label for each color region, abstracting away from how speakers select among alternative terms in specific communicative situations. Natural languages typically provide multiple alternative labels for the same referent and a single color, for example, can be described as either purple or magenta (Gualdoni & Boleda, 2024). Recently, pragmatic modeling work employed color reference games to investigate agent color naming behavior in context. For instance, Monroe et al. (2017) model speakers and listeners using the Rational Speech Acts framework and show that speakers produce longer, more descriptive utterances in more difficult contexts. Similarly, White et al. (2020) demonstrate that agents trained to approximate pragmatic reasoning objectives adjust utterance length in response to communicative context. In contrast to these studies, NeLLCom-Lex does not rely on multi-word utterances or utterance length as a measure of effort, but rather accounts for variation in the informativeness of individual word choices, by adopting a measure developed by Gualdoni and Boleda (2024) that takes into account both context adaptation in language use and the structure of lexical systems. NeLLCom-Lex demonstrated that access to contextual information during RL is important for developing human-like pragmatic behavior. Our work builds directly on this framework, focusing on the geometric and semantic properties of the emergent lexicons.

**Geometry of Meaning Spaces** The geometric organization of lexical categories has long been a topic of interest in cognitive semantics. Conceptual space theory argues that natural language categories tend to form approximately convex regions in perceptual and conceptual spaces (Gärdenfors, 2014). Empirical studies of color naming across languages support this view, showing that color categories are typically compact and contiguous in perceptual color space. Recent computational work has formalized convexity as a measurable property of semantic systems. Steinert-Threlkeld and Szymanik (2020) show that color systems with more convex categories are easier for neural networks to learn, providing a learning-based explanation for semantic universals. Their convexity metric, based on the ratio between category size and its convex hull, offers a principled way to quantify the geometric coherence of lexical categories and has been used by Carlsson et al. (2024) to study the pressures underlying the emergence of efficient, human-like color naming systems. Building on this literature,

<sup>1</sup>Code and materials are available at [https://github.com/yuqing0304/NeLLCom\\_Lex\\_CogSci](https://github.com/yuqing0304/NeLLCom_Lex_CogSci).

we adopt this metric to analyze the convexity of the color categories acquired and developed by our agents.

**Frequency Effects in Learning** Frequency plays a central role in the learning process across both humans and computational models. In human language acquisition, high-frequency words are typically learned earlier, show greater semantic stability, and are more resistant to attrition (Ellis, 2002). Similar biases arise in computational models trained on imbalanced data. Neural networks are known to overrepresent frequent categories while under-differentiating rare ones, a problem widely studied under class imbalance (Henning et al., 2023). Data-level interventions such as upsampling minority classes have been shown to partially mitigate these biases by increasing effective exposure during training (Chawla et al., 2002).

**Listener Diversity** Most emergent communication models focus on a single speaker-listener pair, which is also the case in the original NeLLCom-Lex. This setup allows lexical conventions to overfit to a particular partner, potentially encouraging idiosyncratic and non-convex category boundaries. While there is robust evidence for partner-specific accommodation in human communication (Brennan & Clark, 1996; Pickering & Garrod, 2004), research also suggests that when partner-specificity offers no communicative advantage, speakers converge on partner-independent, aggregate statistics across their interlocutors (Ostrand & Ferreira, 2019). Thus, communicating with diverse listeners may favor more systematic and listener-agnostic linguistic systems, as also demonstrated in laboratory experiments where larger groups of participants developed more systematic and compositional lexicons than smaller groups (Raviv et al., 2019). Agent-based simulations similarly show that when a speaker interacts with multiple listeners sequentially, languages that are more compositional and easier to teach to newcomers emerge (Li & Bowling, 2019). In addition, representations learned in populations are less idiosyncratic and more shareable than those learned in fixed pairs (Tieleman et al., 2019), and promote compositionality (Michel et al., 2023). Here, we therefore introduce a *many-listeners setup* in NeLLCom-Lex, in which one speaker communicates with multiple listeners to test how it affects convexity.

## Experimental Setup

In the NeLLCom-Lex framework (Zhang et al., 2025), both speaker and listener are implemented as feed-forward neural networks operating on CIELAB color representations. The speaker encodes the target and its context (when context-aware) and produces a discrete message via a classifier, while the listener selects a referent by comparing the message embedding with candidate color embeddings. During SL, agents are trained on human-produced labels using cross-entropy loss, whereas during RL they are jointly optimized based on communicative success in referential games. Building on

this framework, we introduce two main changes to test their effects on the properties of the resulting lexicon.

In our new *many-listeners setup* one speaker communicates with 5 or 30 listeners<sup>2</sup>, each of which is an independently initialized listener network trained for 30 SL epochs on the human data. During the subsequent 30-epoch RL phase, referential game training is distributed across listeners for controlled comparison. Specifically, listeners are randomly ordered without replacement, and the speaker interacts with each listener for a fixed number of consecutive RL epochs before moving to the next (e.g., with 5 listeners, each listener is trained for 6 consecutive RL epochs).

Second, severe frequency imbalance in the SL data, with a mean frequency of 328 and a maximum frequency of 2585, biases the initial lexical representations learned by the speaker. High-frequency color terms receive strong grounding, whereas rare colors are weakly represented. In addition, RL tends to amplify initially successful mappings, causing rare terms to be forgotten or absorbed into broader categories. To address this issue, we introduce an *upsampling setup* where rare color samples are duplicated during SL until each reaches a target count  $N$ , while leaving the RL phase unchanged<sup>3</sup>. We set the target upsampling levels  $N$  to 100 and 200 to represent moderate and strong rebalancing of the highly skewed frequency distribution. These values increase exposure to rare terms while still preserving relative frequency differences.

Training follows Zhang et al. (2025)’s SL+RL+ pipeline (SL with context followed by RL with context), which yielded the lexicon with the most human-like properties. All architectural parameters and hyperparameters are kept identical.<sup>4</sup>

## Datasets

For SL, we use the English **Colors** dataset processed by Gualdoni and Boleda (2024), originally collected by Monroe et al. (2017). The data come from a dyadic reference game in which a speaker describes a target color chip among two distractors, and the listener needs to select the correct target. Context difficulty depends on visual similarity: **FAR** (both distractors distinct), **SPLIT** (one distinct, one similar), and **CLOSE** (both similar). We use all successful single-word trials, containing 15,434 instances (9,309 far, 3,886 split, 2,239 close). For SL, 3K instances are held out for testing ( $\text{TEST}_{\text{hum}}$ ), and the remaining 12.4K are used for training ( $\text{TRAIN}_{\text{hum}}$ ), similar to Zhang et al. (2025)’s original setup. For RL, where human labels are not required, we generate synthetic far/close color

<sup>2</sup>In NeLLCom-Lex terms, a speaker (or listener) refers to a randomly initialized speaker (listener) network. Multiple speakers (listeners) then refer to multiple networks initialized with different random seeds prior to training.

<sup>3</sup>More complex augmentation strategies, such as slightly corrupting or modifying existing rare color samples to generate color variants, are an interesting direction, which we leave for future work.

<sup>4</sup>We increase the granularity of the color space by representing CIELAB values with one decimal place instead of integers. We also note that Zhang et al. (2025)’s color sampling excludes greyish colors, which may limit generalization to pale shades; accordingly, we adjust the sampling criteria to include all saturation levels.

triplets following the sampling procedure of Monroe et al. (2017). Both RL training and evaluation are conducted on generated data whose context distribution matches that of the human **Colors** dataset used for SL. In total, we use 12.4K generated triplets for training ( $\text{TRAIN}_{\text{gen}, \text{distH}}$ ) and 15.4K for evaluation ( $\text{TEST}_{\text{gen}, \text{distH}}$ ).

## Evaluation

Following Zhang et al. (2025), agent performance is evaluated using communication accuracy and pragmatic adaptation, while properties of the emergent lexicon are characterized using several system-level measures. In addition, we introduce convexity as a key metric to more closely examine the geometric structure of lexical categories.

**Communication accuracy** ( $\text{Acc}_{\text{comm}}$ ) is defined as the proportion of trials in which the listener correctly identifies the target color given the input from the speaker.

**Pragmatic adaptation** is assessed by testing whether agents adapt their naming choices to task difficulty, quantified by the link between word informativeness and context ease. Word informativeness ( $I_w$ ) is computed from the spread of a word’s denotation in CIELAB space (Gualdoni & Boleda, 2024). For a color term  $w$ , spread ( $S_w$ ) is defined as the average pairwise Euclidean distance between all color chips denoted by  $w$ , that is  $S_w = (\sum_i \sum_{j \neq i} d(o_i, o_j))/P$ , where  $P$  is the number of unique object pairs and  $d(o_i, o_j)$  is the Euclidean distance between objects  $o_i$  and  $o_j$ . Informativeness is defined as  $I_w = 1/S_w$ , such that more compact denotations correspond to higher  $I_w$ . Context ease ( $E_{\text{ctx}}$ ) refers to the distance between the target color chip and the hardest distractor in the context. A linear mixed-effects model is fitted to predict the informativeness of the produced word ( $I_w$ ) from  $E_{\text{ctx}}$ , with agent seeds and target chips as random effects. For each setting, we report  $\beta(E_{\text{ctx}})$ , the estimated effect of context ease on word informativeness.

Our system-level metrics are also taken from Zhang et al. (2025): **System-level informativeness** (Gualdoni & Boleda, 2024) is defined as the average informativeness  $I_w$  of the words used to solve  $N$  interactions,  $I_L = (\sum_{i=1}^N I_w^i)/N$ . Human-like lexicons exhibit intermediate  $I_L$  values, supporting meaningful distinctions without being overly specific (Gualdoni & Boleda, 2024). **Lexical diversity** is defined as the number of different color words used by the speaker. **Semantic drift** quantifies how agents’ lexicons diverge from the human lexicon by computing the Euclidean distance between their respective prototypes for each word (Gualdoni et al., 2023).

In addition to these measures, we introduce a **convexity** measure to examine the geometric structure of the agents’ conceptual space. Specifically, we adopt the procedure proposed by Steinert-Threlkeld and Szymanik (2020), whereby the degree of convexity of a color system is defined as its proximity to the closest convex color system. Following this approach, for each color term  $c$  in a lexicon  $L$ , we compute the convex hull of  $c$ , defined as the smallest convex region that contains all CIELAB points associated with  $c$ . We then calcu-

late the proportion of this convex hull that is already covered by  $c$ , which constitutes the convexity degree of  $c$ . To obtain the overall degree of convexity of  $L$ , we deviate from the original measure in Steinert-Threlkeld and Szymanik (2020) by averaging the convexity scores of individual color terms without applying any weights<sup>5</sup>:

$$\text{Convexity}(L) = \frac{1}{T} \sum_{i=1}^T \frac{|c^i|}{|\text{ConvexHull}(c^i)|} \quad (1)$$

where  $T$  denotes the number of color terms in the lexicon.

All experiments are repeated with 10 random seeds, and agent production is evaluated after 30 epochs of RL training.

## Results and Discussion

Results are reported in Table 1. Overall, agents achieve high communication accuracy across conditions, comparable to the accuracy levels reported by Monroe et al. (2017) (far: 97%, split: 90%, close: 83%). However, accuracies below 90% before RL indicate that SL alone does not fully support robust communication on the test set. Incorporating RL further improves accuracy, suggesting that communication-based training facilitates more effective language use, and confirming the findings of Zhang et al. (2025). With respect to the newly added factors, we observe that accuracy is slightly lower when speakers interact with a larger number of listeners, but remains stable under increased upsampling for both SL and RL.

As for pragmatic naming behavior, neural agents exhibit context-sensitive pragmatic adaptation under all conditions (all  $\beta(E_{\text{ctx}})$  values are significant, with  $p < 0.001$ , and standard errors for all estimates below 0.001). In addition, pragmatic adaptation is already strong after SL and slightly exceeds that observed after RL, consistently with Zhang et al. (2025).

Having assessed that agents broadly succeed at context-sensitive language use, we turn to the properties of the learned lexicons, shown in Table 1 and illustrated in Figure 2, focusing on lexical diversity, informativeness, and convexity.

**Lexical Diversity.** Comparing SL and SL+RL, we observe that RL training overall increases lexical diversity, replicating the patterns in Zhang et al. (2025). The number of listeners affects lexical diversity. With no upsampling,  $|W|$  decreases with fewer listeners. Even with higher upsampling, many-listener setups still limit vocabulary usage compared to few-listener setups. This suggests that interacting with multiple listeners constrains the agents to use a more consistent and focused set of words. For a fixed listener setup, upsampling consistently increases lexical diversity. For instance, in the SL-only condition,  $|W|$  increases with higher levels of upsampling. Similarly, in the SL+RL condition with one listener,  $|W|$  increases across the same upsampling levels. This indicates that upsampling allows the agents to expand their vocabulary and use a wider range of words. Overall, the max-

<sup>5</sup>We use an unweighted average because our production data contain many high-frequency general terms, which would otherwise dominate the score and obscure rarer specific terms.

Table 1: Agents’ production properties under different listener setups and upsampling strategies. Numbers show the mean across 10 seeds.  $\beta(E_{ctx})$ : effect of context ease on word informativeness.  $|W|$ : lexical diversity;  $I_L$ : system-level informativeness;  $D_L$ : semantic drift. Human values are computed using the full 15.4K **Colors** dataset (Gualdoni & Boleda, 2024).

| Conditions   | Setup     |            | Communication |                  | Lexicon Properties |       |           | $D_L$ |
|--------------|-----------|------------|---------------|------------------|--------------------|-------|-----------|-------|
|              | Listeners | Upsampling | $Acc_{comm}$  | $\beta(E_{ctx})$ | $ W $              | $I_L$ | Convexity |       |
| <b>SL</b>    | N/A       | 0          | 0.87          | -0.008           | 13.6               | 3.18  | 0.60      | 10.31 |
|              | N/A       | 100        | 0.86          | -0.010           | 28.6               | 3.21  | 0.60      | 12.24 |
|              | N/A       | 200        | 0.87          | -0.013           | 35.9               | 3.36  | 0.42      | 12.15 |
| <b>SL+RL</b> | 1         | 0          | 0.94          | -0.004           | 31.8               | 2.62  | 0.24      | 39.69 |
|              | 1         | 100        | 0.94          | -0.004           | 39.0               | 2.68  | 0.22      | 35.11 |
|              | 1         | 200        | 0.94          | -0.004           | 42.6               | 2.74  | 0.20      | 27.11 |
|              | 5         | 0          | 0.93          | -0.005           | 23.6               | 2.48  | 0.28      | 41.98 |
|              | 5         | 100        | 0.93          | -0.005           | 31.4               | 2.49  | 0.30      | 37.42 |
|              | 5         | 200        | 0.93          | -0.005           | 37.6               | 2.54  | 0.26      | 28.51 |
|              | 30        | 0          | 0.92          | -0.005           | 21.7               | 2.38  | 0.31      | 49.19 |
|              | 30        | 100        | 0.92          | -0.005           | 29.7               | 2.39  | 0.37      | 37.72 |
|              | 30        | 200        | 0.93          | -0.006           | 37.2               | 2.48  | 0.26      | 27.74 |
| <b>Human</b> | —         | —          | 1.00          | -0.008           | 49.0               | 2.78  | 0.32      | —     |

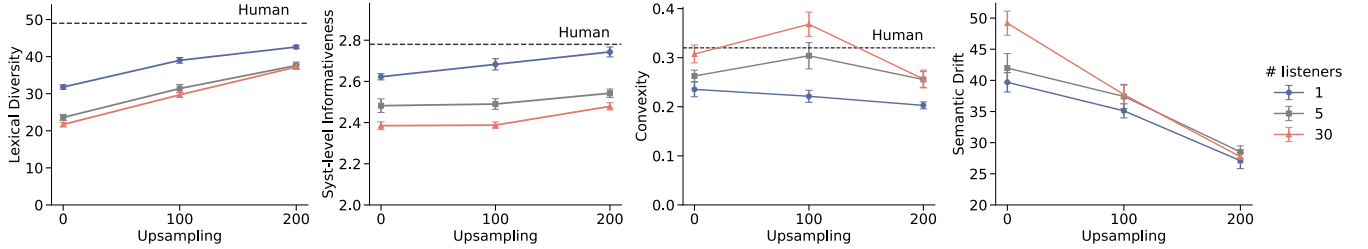


Figure 2: Lexicon properties across listener setups and upsampling levels. Each marker represents the mean across 10 seeds, with bars showing standard error. The human reference values are indicated by dashed horizontal lines.

imum lexical diversity is achieved with a single listener and the highest level of upsampling, yielding the richest vocabulary.

**System-level Informativeness.** Comparing SL and SL+RL, RL training is associated with lower  $I_L$ , consistent with the patterns reported by Zhang et al. (2025), suggesting that SL-trained agents may develop lexicons that encode redundant distinctions, and the emergent lexicon gets optimized through interactive communication. The number of listeners affects  $I_L$ . Similar to lexical diversity,  $I_L$  decreases as the number of listeners increases, regardless of the upsampling level. This likely reflects the fact that, with a fixed number of interactions, increasing the number of listeners makes it more difficult for each agent to learn a well-formed lexicon. As a result, communicating with multiple listeners favors more general and conservative lexical choices, leading to lower diversity and  $I_L$ . For a fixed listener setup, upsampling leads to a moderate but consistent increase in informativeness. In the SL-only condition,  $I_L$  rises as upsampling increases. A similar trend is observed in the SL+RL condition with one listener, where  $I_L$  increases for the same upsampling settings. This suggests that exposure to more balanced training data enables agents to construct lexicons that convey meanings more informatively. Overall, system-level informativeness is most human-like in

the SL+RL setting with a single listener and the highest upsampling level. These results suggest that optimizing input for training can shift agents away from overly fine-grained lexical distinctions toward more efficient lexicons, bringing the agents’ informativeness closer to the human level.

**Convexity.** SL-trained agents exhibit very high convexity across conditions, suggesting that they acquire structured semantic representations, likely due to low lexical diversity. In contrast, RL training reduces convexity to levels closer to or slightly below those observed in humans. Interacting with more listeners consistently increases convexity. Under SL+RL with no upsampling, convexity rises from 0.24 (1 listener) to 0.28 (5 listeners) and 0.31 (30 listeners), a trend that persists under higher upsampling levels. Thus, interacting with multiple listeners may encourage agents to adopt more compact and consistent color representations. The effect of upsampling on convexity is non-monotonic. Since upsampling tends to increase lexical diversity, excessive upsampling makes convexity harder to maintain, given that lexicons learned after SL alone are already highly convex. Consistent with this intuition, upsampling-200 setting leads to a decline in convexity across conditions. For instance, in SL, convexity drops from 0.60 to 0.42 as upsampling increases from 0 to



200, alongside a vocabulary increase from  $|W|=13.6$  to 35.9. A similar downward trend is observed in SL+RL. However, intriguingly, moderate upsampling improves convexity under many-listener setups. For example, in the SL+RL condition with five listeners, convexity increases from 0.28 to 0.30 at an upsampling level of 100, but decreases to 0.26 when upsampling is further increased to 200. This pattern, also visible in the general trend in Figure 2, suggests a potential Goldilocks principle<sup>6</sup>, named after the fairy tale Goldilocks and the Three Bears, in which the optimal choice is "just right" rather than too extreme (Brochhagen & Boleda, 2022). In our case, moderate upsampling yields a medium-sized effective lexicon (around 30 words), and multiple interlocutors ensure convexity, together helping agents learn coherent partitions of the color space. In contrast, excessive upsampling yields a larger-sized lexicon (around 37 words) that likely introduces finer-grained distinctions that may fragment semantic regions, making it more difficult to maintain high convexity. Notably, SL+RL settings with 30 listeners and moderate upsampling level (i.e., 100) achieve the highest convexity across conditions and, though not closest to the human value, provide the best overall balance across all metrics, suggesting that interacting with many listeners combined with moderate data augmentation promotes more human-like semantic organization. Such an effect is consistent with prior work showing that increasing the number of interacting agents reduces idiosyncratic co-adaptation (Graesser et al., 2019) and aligns with findings that greater input variability in groups can drive languages toward simpler and more systematic systems in human experiments (Raviv et al., 2019).

To illustrate the effect of the optimal condition, Figure 3 compares denotation examples from the baseline setting (single listener, no upsampling) with those from the 30-listener, upsampling-100 setting. Under the latter, the denotations of most color terms, including both specific color names such as *aqua* and more general terms like *green* and *red*, become more spatially compact and convex in CIELAB space.

**Semantic Drift.** We finally examine semantic drift, which measures how much the learned lexicon deviates from the human lexicon learned during SL. Importantly, higher semantic drift does not necessarily indicate poor agent performance; the representations may still be internally consistent and convex. As in other neural-agent-based simulations of language evolution (Lian et al., 2024, 2025; Zhang et al., 2024), RL provides interactive communication settings, and semantic drift in RL provides a useful lens for exploring how different interaction conditions shape the properties of the emergent agent lexicon.

Interacting with many listeners tends to increase the drift. Combined with the observation that more listeners increase convexity, this suggests that group communication amplifies pressure toward internally consistent but potentially less human-like semantic conventions. Upsampling under SL

<sup>6</sup>Establishing a strict Goldilocks optimum would require a more systematic exploration of intermediate upsampling values, which we leave for future work.

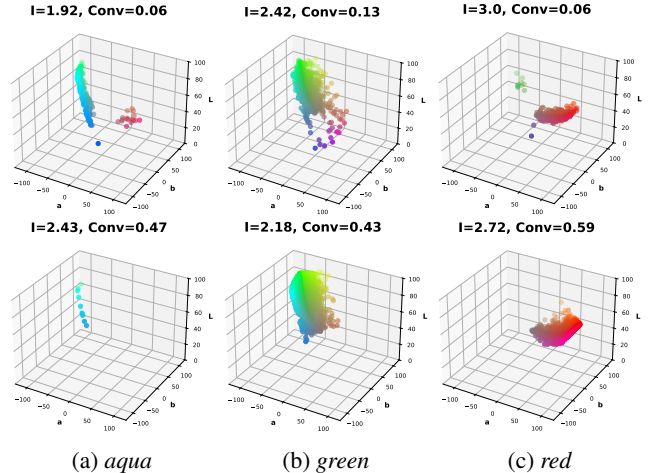


Figure 3: Denotations of three color terms in CIELAB space under two training conditions (top: 1 listener, no upsampling; bottom: 30 listeners, upsampling 100; both are after RL).

leads to a slight increase in drift, likely because a larger vocabulary introduces additional lexical distinctions. Nevertheless, upsampling substantially reduces semantic drift in SL+RL conditions for a fixed listener setup. For example, with one listener,  $D_L$  decreases from 39.69 (no upsampling) to 35.11 (upsampling 100) and further to 27.11 (upsampling 200). A similar decreasing trend is observed with 5 and 30 listeners. This pattern suggests that upsampling promotes more stable lexicons, making it more difficult for agents to shift denotations or establish idiosyncratic conceptual pacts over time.

**Discussion.** Overall, our results show that lexical diversity and informativeness are shaped by both the number of listeners and upsampling, with single-listener and moderate-to-high upsampling setups promoting richer and more informative lexicons. Convexity is generally high under SL, but RL combined with multiple listeners and moderate upsampling yields more human-like color representations. Semantic drift highlights the trade-off between alignment with human lexicons and internally consistent group communication.

## Conclusion

We investigated why neural agents trained in the NeLLComLex framework develop lexicons that diverge from human color naming systems, despite achieving high communicative accuracy. Our experiments identify two factors that could influence the properties of emergent lexicons: frequency imbalance in the supervised training data and listener diversity during the referential game. Upsampling rare color terms increases lexical diversity, informativeness, and reduces semantic drift, while interacting with multiple listeners constrains idiosyncratic productions and promotes more convex semantic categories. Neither factor alone is sufficient; the most human-like lexicons emerge from their combination, balancing convexity, informativeness, vocabulary size, and drift.

## Acknowledgments

Arianna Bisazza is funded by the Talent Programme of the Dutch Research Council (NWO) under project VI.Vidi.221C.009. Gemma Boleda is funded by the Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación (Spain; grant PID2020-112602GB-I00/MICIN/AEI/10.13039/501100011033).

## References

- Baronchelli, A., Gong, T., Puglisi, A., & Loreto, V. (2010). Modeling the emergence of universality in color naming patterns. *Proceedings of the National Academy of Sciences*, 107(6), 2403–2407. <https://doi.org/10.1073/pnas.0908533107>
- Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language Is a Complex Adaptive System: Position Paper [eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9922.2009.00533.x>]. *Language Learning*, 59(s1), 1–26. <https://doi.org/10.1111/j.1467-9922.2009.00533.x>
- Berlin, B., & Kay, P. (1991). *Basic color terms: Their universality and evolution*. University of California Press, Berkeley.
- Bleys, J., Loetzsch, M., Spranger, M., & Steels, L. (2009). The grounded colour naming game. *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication (RoMan 2009)*.
- Brainard, D. H. (2003). Color appearance and color difference specification. In *The science of color* (2nd ed., pp. 191–216). Elsevier.
- Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6), 1482.
- Brochhagen, T. (2018). *Signaling under uncertainty* [Doctoral dissertation, University of Amsterdam].
- Brochhagen, T., & Boleda, G. (2022). When do languages use the same word for different meanings? the goldilocks principle in colexification. *Cognition*, 226, 105179.
- Campbell, L. (2013). *Historical linguistics*. Edinburgh University Press.
- Cangelosi, A., & Parisi, D. (2002). Computer simulation: A new scientific approach to the study of language evolution. In *Simulating the evolution of language* (pp. 3–28). Springer.
- Carlsson, E., Dubhashi, D., & Regier, T. (2024). Cultural evolution via iterated learning and communication explains efficient color naming systems. *Journal of Language Evolution*, 9(1-2), 49–66.
- Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2021). Communicating artificial neural networks develop efficient color-naming systems. *Proceedings of the National Academy of Sciences*, 118(12), e2016569118.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321–357.
- Clark, H. H. (1996). *Using language*. Cambridge university press.
- Cook, R. S., Kay, P., & Regier, T. (2005). The world color survey database. In *Handbook of categorization in cognitive science* (pp. 223–241). Elsevier.
- De Boer, B. (2006). Computer modelling as a tool for understanding language evolution. In *Evolutionary epistemology, language and culture: A non-adaptationist, systems theoretical approach* (pp. 381–406). Springer.
- Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition*, 24(2), 143–188.
- Gärdenfors, P. (2000). Conceptual spaces: The geometry of thought.
- Gärdenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces*. MIT Press.
- Graesser, L. H., Cho, K., & Kiela, D. (2019). Emergent linguistic phenomena in multi-agent communication games. *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 3700–3710.
- Gualdoni, E., & Boleda, G. (2024, November). Why do objects have many names? a study on word informativeness in language use and lexical systems. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 18150–18163). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.1009>
- Gualdoni, E., Brochhagen, T., Mädebach, A., & Boleda, G. (2023). What's in a name? A large-scale computational study on how competition between names affects naming variation [Publisher: Elsevier]. *Journal of Memory and Language*, 133, 104459.
- Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. OUP Oxford.
- Henning, S., Beluch, W., Fraser, A., & Friedrich, A. (2023). A survey of methods for addressing class imbalance in deep-learning based natural language processing. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 523–540.
- Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization*. Cambridge university press.
- Jäger, G. (2007). The evolution of convex categories. *Linguistics and philosophy*, 30(5), 551–564.
- Jäger, G. (2009). Natural color categories are convex sets. *Logic, Language and Meaning: 17th Amsterdam Colloquium, Amsterdam, The Netherlands, December 16-18, 2009, Revised Selected Papers*, 11–20.



- Kågeback, M., Carlsson, E., Dubhashi, D., & Sayeed, A. (2020). A reinforcement-learning approach to efficient communication. *PLOS ONE*, 15(7), 1–26. <https://doi.org/10.1371/journal.pone.0234894>
- Kharitonov, E., Chaabouni, R., Bouchacourt, D., & Baroni, M. (2019, November). EGG: A toolkit for research on emergence of lanGuage in games. In S. Padó & R. Huang (Eds.), *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* (pp. 55–60). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-3010>
- Lazaridou, A., & Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*.
- Li, F., & Bowling, M. (2019). Ease-of-teaching and language structure from emergent communication. *Advances in neural information processing systems*, 32.
- Lian, Y., Bisazza, A., & Verhoef, T. (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. [https://doi.org/10.1162/tac1\\_a\\_00587](https://doi.org/10.1162/tac1_a_00587)
- Lian, Y., Bisazza, A., & Verhoef, T. (2025). Simulating the emergence of differential case marking with communicating neural-network agents. *Proceedings of the 47th Annual Conference of the Cognitive Science Society (CogSci)*. <https://doi.org/10.48550/arXiv.2502.04038>
- Lian, Y., Verhoef, T., & Bisazza, A. (2024, November). NeLLCom-X: A comprehensive neural-agent framework to simulate language learning and group communication. In L. Barak & M. Alikhani (Eds.), *Proceedings of the 28th conference on computational natural language learning* (pp. 243–258). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.conll-1.19>
- Loreto, V., Mukherjee, A., & Tria, F. (2012). On the origin of the hierarchy of color names. *Proceedings of the National Academy of Sciences*, 109(18), 6819–6824.
- Michel, P., Rita, M., Mathewson, K. W., Tieleman, O., & Lazaridou, A. (2023). Revisiting populations in multi-agent communication. *The Eleventh International Conference on Learning Representations*.
- Monroe, W., Hawkins, R. X., Goodman, N. D., & Potts, C. (2017). Colors in context: A pragmatic neural model for grounded language understanding (L. Lee, M. Johnson, & K. Toutanova, Eds.). *Transactions of the Association for Computational Linguistics*, 5, 325–338. [https://doi.org/10.1162/tac1\\_a\\_00064](https://doi.org/10.1162/tac1_a_00064)
- Ostrand, R., & Ferreira, V. S. (2019). Repeat after us: Syntactic alignment is not partner-specific. *Journal of memory and language*, 108, 104037.
- Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2), 169–190.
- Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Larger communities create more systematic languages. *Proceedings of the Royal Society B*, 286(1907), 20191262.
- Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. *The handbook of language emergence*, 237–263.
- Steels, L. (1997). The synthetic modeling of language origins. *Evolution of communication*, 1(1), 1–34.
- Steels, L., Belpaeme, T., et al. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and brain sciences*, 28(4), 469–488.
- Steels, L., & McIntyre, A. (1998). Spatially distributed naming games. *Advances in complex systems*, 1(04), 301–323.
- Steinert-Threlkeld, S., & Szymanik, J. (2020). Ease of learning explains semantic universals. *Cognition*, 195, 104076.
- Tieleman, O., Lazaridou, A., Mourad, S., Blundell, C., & Precup, D. (2019). Shaping representations through communication: Community size effect in artificial learning systems. *arXiv preprint arXiv:1912.06208*.
- Tucker, M., Shah, J., Levy, R., & Zaslavsky, N. (2025). Towards human-like emergent communication via utility, informativeness, and complexity. *Open Mind*, 9, 418–451.
- White, J., Mu, J., & Goodman, N. D. (2020). Learning to refer informatively by amortizing pragmatic reasoning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 42.
- Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31), 7937–7942.
- Zhang, Y., Ürker, E., Verhoef, T., Boleda, G., & Bisazza, A. (2025, November). NeLLCom-lex: A neural-agent framework to study the interplay between lexical systems and language use. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Findings of the association for computational linguistics: Emnlp 2025* (pp. 10929–10945). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.findings-emnlp.580>
- Zhang, Y., Verhoef, T., van Noord, G., & Bisazza, A. (2024, May). Endowing neural language learners with human-like biases: A case study on dependency length minimization. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 5819–5832). ELRA; ICCL. <https://aclanthology.org/2024.lrec-main.516/>