Bridging semantics and pragmatics in information-theoretic emergent communication

Eleonora Gualdoni
Apple MLR*
Universitat Pompeu Fabra
e_gualdoni@apple.com

Mycal Tucker
MIT
mycal@mit.edu

Roger P. Levy
MIT
rplevy@mit.edu

Noga Zaslavsky
NYU
nogaz@nyu.edu

Abstract

Human languages support both semantic categorization and local pragmatic interactions that require context-sensitive reasoning about meaning. While semantics and pragmatics are two fundamental aspects of language, they are typically studied independently and their co-evolution is largely under-explored. Here, we aim to bridge this gap by studying how a shared lexicon may emerge from local pragmatic interactions. To this end, we extend a recent information-theoretic framework for emergent communication in artificial agents, which integrates utility maximization, associated with pragmatics, with general communicative constraints that are believed to shape human semantic systems. Specifically, we show how to adapt this framework to train agents via unsupervised pragmatic interactions, and then evaluate their emergent lexical semantics. We test this approach in a rich visual domain of naturalistic images, and find that key human-like properties of the lexicon emerge when agents are guided by both context-specific utility and general communicative pressures, suggesting that both aspects are crucial for understanding how language may evolve in humans and in artificial agents.

1 Introduction

Languages evolve through repeated interactions in rich contexts, where various communicative and non-communicative goals co-exist. The conveyed meaning is often shaped by the local conversational context of utterances (Figure 1), as captured by the *pragmatic* behavior of interlocutors [1]. For example, the word 'player' can be interpreted as a baseball batter, catcher, or a guitar player, depending on the conversational context that shapes the listener's beliefs about the speaker's state of mind. That is, understanding meaning in context requires pragmatic reasoning about other agents' intentions and beliefs [1–3]. At the same time, words are associated with non-contextualized meanings, as captured by *lexical semantics*. For example, we have a shared idea of what 'player' means, regardless of any specific context in which this word might appear, and we can use it to communicate with new conversational partners in new contexts. While semantics and pragmatics are widely studied, their interface and co-evolution is largely under-explored and not well understood. In this work, we focus on a key open question at the interface of lexical semantics and pragmatic reasoning: How can a shared human-like lexicon emerge from local context-sensitive pragmatic interactions?

To address this question, we build on a framework for information-theoretic emergent communication in multi-agent reinforcement learning systems [4]. This framework is particularly relevant for addressing our question because it integrates task-specific utility maximization — a central component in well-established models of human pragmatic reasoning [2, 3, 5], which has also been center-stage in the literature on emergent communication in artificial agents [6–8] — with general information-theoretic constraints that are believed to shape human lexical semantic systems [9–13] and have

*Work finished prior to joining Apple

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-0663

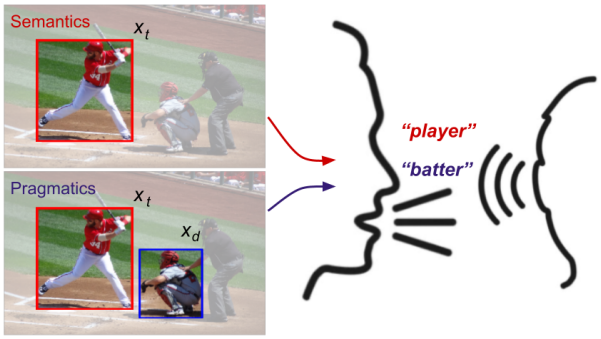

Figure 1: An example of an image from the ManyNames dataset illustrating both semantic and pragmatic communication. **Top** (semantic setting): A speaker communicates a target object x_t (red box) by naming it regardless of the context induced by the image. The ManyNames dataset includes such responses from native English speakers (in this example, 10 speakers produced 'man', 10 'batter', 5 'baseball player', 4 'player', and 3 'person'). **Bottom** (pragmatic setting): A speaker and a listener observe two objects in an image (red and blue boxes). Only the speaker knows which one is the target x_t and which one is the distractor x_d. The speaker's goal is to communicate x_t given this shared context, and the listener's goal is to discriminate the target from the distractor.

recently also been applied to human pragmatic reasoning [14]. These information-theoretic constraints are derived from rate-distortion theory [15, 16], and more specifically, the information bottleneck (IB) principle [17], characterizing semantic systems in terms of efficient compression of meanings into words by optimizing the IB tradeoff between the complexity and informativeness of the lexicon [9]. Tucker et al. [4] developed a framework for training agents in emergent communication settings by integrating utility maximization with the IB objective, yielding a multi-objective optimization problem that trades off maximizing task-specific utilities with maximizing task-agnostic communicative informativeness and minimizing communicative complexity.

Here, we extend that framework to explicitly model the co-evolution of semantics and pragmatics. In our setup, artificial neural agents learn to communicate in a pragmatic setting, i.e., in the presence of a shared conversational context that may alter the meaning of their communication signals (illustrated intuitively in the bottom image of Figure 1). Importantly, agents are trained via self-play, without any supervision or exposure to human languages. At test time, we assess the pragmatic competence (i.e., utility) of the agents as well as the 'human-likeness' of the shared emergent lexicon that they converged on, by invoking it in a lexical semantic setting (illustrated intuitively in the top image of Figure 1, and see Figure 2 for a full description of our model's architecture). Given the substantial empirical evidence that human semantic systems are pressured to optimize the IB tradeoff, we predict that a shared human-like lexicon will not emerge when agents are guided solely by utility maximization but rather when they are guided by a non-trivial tradeoff between optimizing utility, informativeness, and complexity. Our goal in this work is to test this prediction, and more generally, to characterize the landscape of emergent communication systems induced by our novel information-theoretic framework for the co-evolution of semantics and pragmatics.

To this end, we consider a rich visual domain of naturalistic images provided by the ManyNames dataset [18], which also contains free-naming human data generated by native English speakers. This domain allows us to train agents across many different conversational contexts and then evaluate their emergent lexicon with respect to the English naming data. In support of our prediction, we find that human-like properties of the lexicon (its size, complexity, and alignment with English speakers), together with high pragmatic competence, emerge when agents are guided by both context-specific utility and general communicative pressures as derived from the IB principle. Interestingly, pressure for communicative informativeness, rather than (non-communicative) task utility, appears as the main driver of emergent communication, but weaker pressures to minimize complexity and maximize utility are still crucial for achieving human-like properties of the lexicon.

Our work is significant both from a cognitive science perspective and from a machine learning perspective. From a cognitive science perspective, our work provides a novel computational framework for studying the under-explored interface between lexical semantics and pragmatic reasoning, and our findings suggest that human languages may evolve under pressure to optimize a tradeoff

between task-specific utilities and general communicative constraints. From a machine learning perspective, our work demonstrates how cognitively-motivated optimization principles, implemented in neural network agents as intrinsic training objectives, can facilitate the emergence of interpretable human-like communication systems without any human supervision.

2 Related work

Our work integrates cognitively-motivated information-theoretic principles that are believed to underlie human language evolution, with deep learning tools for studying emergent communication in artificial agents, in order to develop a computational account of the co-evolution of lexical semantics and pragmatic reasoning. Semantics and pragmatics constitute two subfield in cognitive linguistics. While both focus on meaning in language, they capture largely complementary aspects of meaning. Lexical semantics is generally concerned with word meanings [19, 20], independent of any specific conversational context, whereas pragmatics is concerned with language use in context, typically assuming a known shared lexicon [1, 21, 3]. Our work focuses on the underexplored interface between these two aspects of language, departing from the traditional assumption that the lexicon is given a-priori and shared among pragmatic interlocutors.

Contemporary cognitive approaches to lexical semantics argue that word meanings are shaped by pressure for efficient communication [22, 23, 10]. Most relevant to our work, is the information-theoretic framework for semantic systems, proposed by Zaslavsky et al. [9], that predicts that human semantic systems evolved to optimize the information bottleneck (IB) trade-off [17] between the complexity of the lexicon (roughly, how many bits are allocated for communication) and its informativeness (roughly, how well a listener can understand a speaker, regardless of context). This framework has shown to account for the structure of semantic systems across hundreds of languages and multiple domains [9, 24, 11, 12], as well as semantic change over time [13]. Here, we leverage this empirically-supported theoretical framework to guide our interactive agents.

Research on pragmatic communication focuses on how speakers' lexical choices and listener's interpretations are affected by their local conversational context [1, 3, 25, 26]. Prominent models of pragmatics, such as the Rational Speech Act (RSA) framework [3], are grounded in game theory (see also [2]). In this view, pragmatic behavior is formulated within a cooperative reference game, where agents pragmatically reason about each other's communicative intentions with the goal of maximizing the game's utility. While these models enjoy broad empirical support [3], they assume that the underlying lexicon is given and shared across speakers and listeners, even when applied in reinforcement learning settings [27]. Recently, a theoretical link between the RSA framework and rate-distortion theory (RDT) has been derived [14], connecting pragmatic reasoning with general informational constraints that are closely related to the aforementioned IB trade-off. However, the RDT approach to pragmatics has also assumed a given lexicon [14], and more generally, the implied information-theoretic link between semantics and pragmatics has not yet been explored. Our work explored this potential link by relaxing the assumption that the lexicon is given to our agents, and instead studying how they may develop on their own a near-optimal human-like lexicon via a training objective that takes into account informational constraints.

Relatedly, Brochhagen et al. [28] developed a game-theoretic model for the evolution of the division of labor between semantics and pragmatics, and tested the model in a relatively small domain. In comparison, we consider here different learning dynamics and objective function, employ state-of-the-art deep-learning tools for training agents at scale, and evaluate our approach with respect to actual human data in a rich domain of naturalistic images [18]. McDowell and Goodman [29] considered the role of pragmatics in learning lexical meanings in deep-learning agents. However, they trained their agents in a supervised manner with respect to human-generated pragmatic data. In contrast, we are interested in how pragmatics and lexical semantic may emerge without any human supervision.

The emergent communication literature focuses on how agents may learn to communicate in reinforcement learning settings, without exposure to human-generated linguistic data [6, 7, 30]. While this body of literature largely focuses on utility (or reward) maximization, Chaabouni et al. [8] showed that utility-based emergent communication can lead to IB-efficient systems. In their settings, however, different complexity-informativeness tradeoffs can only be determined by external environmental factors. Tucker et al. [4] showed how to directly integrate the IB objective function with utility maximization in emergent communication and demonstrated the advantages of this framework for

 https://doi.org/10.52202/079017-0663

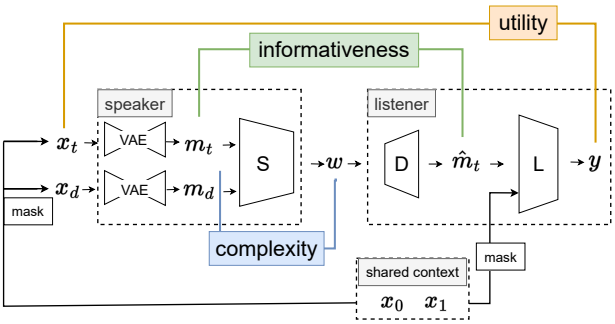

Figure 2: Communication model for the co-evolution of semantics and pragmatics (see Section 3). Agents are trained in a pragmatic setting, where both observe inputs x_0, x_1 as shared context; one input is randomly selected as target x_t for the speaker. After training, the agent's emergent communication systems are evaluated in a lexical semantics setting, without shared context; the speaker observes only x_t while x_d is masked, and both inputs are masked for the listener.

faster convergence rates and better out-of-distribution generalization [31]. However, they did not study the co-emergence of semantics and pragmatics. Our work directly builds on the framework of Tucker et al. [4, 31] and extends it for studying the interface between semantics and pragmatics.

3 A unifying model for the co-evolution of semantics and pragmatics

The information-theoretic framework for emergent communication that we build on [4, 31] consists of an objective function that integrates utility maximization with the IB principle, and a communication model that specifies the agents' architectures. Our model (Figure 2) builds on Tucker et al. [4]'s proposed method, called vector-quantized variational information bottleneck (VQ-VIB), which we modify in order to be able to model the co-evolution of semantic and pragmatics; that is, to be able to train agents in a pragmatic setting and evaluate their communication, at test time, in both pragmatic and lexical semantic settings. The initial VQ-VIB architecture includes a speaker and a listener tasked to jointly solve a reference game. In our case, the speaker is defined by (i) a pre-trained variational autoencoder [VAE, 32] that provides a visual representation module for mapping an input x to a 'mental' representation m; and (ii) an encoder module S that generates a communication signal w given the speaker's mental state. The listener is defined by (i) a decoder D that observes w and generates a reconstruction $\hat{m}$, and (ii) a policy L for solving a downstream task.

In our **pragmatic setting** both agents observe a shared context (x_0, x_1), while the speaker also observes which referent is the target t and which is a distractor d (more details about this step are provided in Section 4). The speaker then aims to communicate the target x_t. The VAE representation model is applied to x_t and x_d independently, generating m_t, the speaker's mental representations for the target, and m_d, the speaker's mental representations for the distractor. The listener's task is to guess the target based on $y = L(\hat{m}, (x_0, x_1))$. We wish to emphasize that in contrast to the standard approach in pragmatic models, in our setup the agents are not given a shared lexicon but rather implicitly learn it through their local, context-sensitive pragmatic interactions.

Agents are trained by optimizing a tradeoff between maximizing expected utility, maximizing informativeness, and minimizing complexity. The utility term, $U(x_t, y)$, is task-specific, and here we take it to be the accuracy of the listener's downstream decisions. Informativeness and complexity are task-agnostic communicative objectives, derived from the Information Bottleneck (IB) framework for semantic systems [9]. In this framework, the speaker's and listener's mental representations, m_t and $\hat{m}$, are defined as belief states (i.e., probability distributions) over features in environment. Informativeness is related to the distortion between m_t and $\hat{m}$, defined by the Kullback-Leibler (KL) divergence $D[m_t \| \hat{m}]$, such that low expected distortion amounts to high informativeness. Complexity corresponds to the mutual information between speaker's meanings and words, $I(m_t; w)$, which is roughly the number of bits that the agents will need for communication. Following [4, 31], we optimize a bound on the IB terms for practical reasons, as direct optimization of those terms does not scale. For informativeness, we treat m_t and $\hat{m}$ as the means of the agents' mental distributions and then approximate informativeness by their MSE. For complexity, we consider a known variational

bound on the mutual information, denoted here for simplicty by $\tilde{I}$ [for more details and full derivation of the objective function, see 4]. Overall, the training objective is to maximize

$$\lambda_U \mathbb{E}\left[U\left(x_t, y\right)\right] - \lambda_I \mathbb{E}\left[\|m_t - \hat{m}_t\|^2\right] - \lambda_C \tilde{I}(m_t, f(I); w)\,, \tag{1}$$

where the λs are non-negative tradeoff weights that sum to 1, and $f(I)$ can be seen as a function that extracts two objects (a target and a distractor) from an image.

Optimizing only the utility term in the objective function, without any pressure for alignment between the speaker's and listener's representations (i.e., high informativeness), is expected to lead to lexical systems that are biased towards success in the specific training downstream task, and such systems are likely to depart from human lexical systems. On the other hand, maximizing informativeness alone will lead to highly complex and task-agnostic systems, and minimizing complexity alone will lead to no communication. None of these extremes seems human-like, and we therefore expect that human-like systems will emerge when all tradeoff parameters are active (i.e., non-zero).

Our **semantic setting** is designed to evaluate the emergent lexicon at test time, lending a window into how the agents use their words irrespectively of any local context. In this setting, only the target is shown to the speaker (the distractor is masked) and then the listener reconstructs $\hat{m}_t$ based on the speaker's word w, without any additional context or downstream task. This task resembles in its nature the task with which annotations for ManyNames –our dataset of interest [18], see below– are gathered, i.e. a naming task where annotators were asked to freely produce names for target objects appearing in natural images, identified with bounding boxed. Thus, we can use these naming annotations at test time to evaluate the agents' emergent semantics.

4 Experimental setup

4.1 The ManyNames domain

We train and test our agents on the ManyNames dataset [18], which contains 25K naturalistic images (see Figure 1 for example), each with one target object, appearing in a bounding box, annotated with ~ 36 names provided by English native speakers asked to freely produce a name (one word) to describe the object.[2] The name produced by the majority of the annotators for a target object is called the *topname*. In the case of Figure 1, top image, 'man' and 'batter' are equally probable topnames. Objects in the ManyNames dataset are also annotated according to their high-level semantic domain, which can be *people*, *animals-plants*, *buildings*, *vehicles*, *clothing*, *food*, or *home*. Importantly, our agents only observe the images in the dataset. They are not trained on any of the linguistic labels, which are used only for evaluation.

4.2 Data selection for pragmatic training

With respect to the choice of the target-distractor pairs for our pragmatic training setup (Figure 1, bottom image), our intuition is that identifying competing objects appearing in the same image, e.g., a batter and a baseball player in the same field, or a car and a taxi on the same street, instead of using random objects cropped from different images, should encourage the emergence of more natural semantic partitions, since agents will need to create lexical entries to solve naturally occurring ambiguities. For each image in the dataset, we consider the target object highlighted in the ManyNames annotation; and an additional object that we individuate through automatic object detection and a few filtering filtering steps. We used the Bottom-Up object detection model proposed by Anderson et al. [33], which incorporates a Faster R-CNN architecture [34] for object detection, and a ResNet-101 architecture [35] for feature extraction. Since Anderson et al. [33]'s model is fine-tuned on the VisualGenome image dataset [36], of which ManyNames images are a sample, we are guaranteed that this model can make good quality predictions on ManyNames. After running the object detector model on our images, we filter the detected bounding boxes keeping only those with Intersection over Union smaller than 0.1 with the ManyNames target, and that did not have size smaller than 10% of the target size.[3] We then computed the similarity between the candidates' visual features –automatically

[2] Creative Commons Attribution 4.0 International License.

[3] For $2.5K$ images, typically depicting only one large object as the ManyNames target, we do not find any detected object with the desired properties. Being unable to use these images in the pragmatic setting, we exclude them from our data sample.

https://doi.org/10.52202/079017-0663

Model	ΔCompl.	ΔLexSize	NID	Utility	MSE
$\lambda_U = 1$	1.76 ($\pm$0.31)	559 ($\pm$382)	0.84 ($\pm$0.02)	0.95 ($\pm$0.00)	20 ($\pm$7.18)
$\lambda_I = 1$	2.80 ($\pm$0.06)	1687 ($\pm$107)	0.60 ($\pm$0.00)	0.83 ($\pm$0.00)	0.13 ($\pm$0.00)
$\lambda_C = 1$	5.21 ($\pm$0.01)	381 ($\pm$0.98)	0.99 ($\pm$0.00)	0.58 ($\pm$0.00)	0.32 ($\pm$0.00)
λ^*_{Eng}	**1.48** ($\pm$0.08)	**22** ($\pm$35)	**0.55** ($\pm$0.00)	0.72 ($\pm$0.01)	0.23 ($\pm$0.00)

Table 1: Evaluation of the model that is best aligned with English, λ^*_{Eng} ($\lambda_U = 0.005$, $\lambda_I = 0.98$, $\lambda_C = 0.015$), in comparison with three baselines: $\lambda_U = 1$, corresponding to utility maximization without any informational constraints, which is the most common objective function in the emergent communication literature; and $\lambda_I = 1$ and $\lambda_C = 1$ which correspond to the other two extremes of either maximizing informativeness or minimizing complexity. ΔCompl. and ΔLexSize are the absolute value of the differences between the complexity or lexicon size of the emergent system and English. NID measures the misalignment between the emergent system and English. For all three measures, lower values reflect a better fit to the human data. Utility and MSE correspond to the agent's pragmatic competence and reconstruction error, respectively. Each cell shows the mean value across three random seeds $\pm$ SEM.

extracted by the ResNet-101 incorporated in Anderson et al. [33]'s model– and the ManyNames target's features, choosing the detected object with the highest similarity value. This final selection step based on visual similarity aims at providing our agents with some challenging cases to solve, like the one shown in Figure 1, where communication needs to allow the discrimination between two objects from the same semantic category, e.g. a batter and another baseball player.

At the end of this procedure, we obtained two objects per image: the ManyNames target; and the additional object detected by us. We used these two objects to train our agents, choosing our target's identity randomly, making sure that in 50% of the cases the target would be the larger object, and in the other 50% the distractor would be larger.

4.3 Multi-agent simulations

We trained 270 agent pairs with a range of combinations of λ_U, λ_C and λ_I. We used parameter annealing which, besides being a computationally efficient alternative to training agents from random initializations, has been shown to capture at least some aspects of the evolution of human semantic systems [9]. We used a pretrained ResNet18 [35] to extract 512 dimensional features from target and distractor objects before passing them to the agents.[4] In all our experiments, we trained agents on 70% of the images (randomly sampled) in self-play, with no human supervision, batch size of 128, and codebook initialized with 3000 trainable communication vectors (see Appendix A for further details). [5]

5 Results

5.1 Quantitative evaluation

We report quantitative results for the 5 evaluation metrics shown in Figure 3, for each set of λs, at test time. The top panel shows the three components of the agents' training objective: (a) the agents' pragmatic competence, as measured by their test-time expected utility on unseen images, in the pragmatic setting; (b) the reconstruction loss between the speaker's and listener's mental representations, measured by the negative MSE (values closer to 0 corresponds to higher informativeness); and (c) the complexity of the emergent lexicon. The bottom panel of Figure 3 shows two measures for assessing the human-likeness of the emergent systems: (d) the lexicon size of the artificial systems,

[4]We extract visual features with a ResNet18 model, that is a standard choice for object classification, and choose not use the visual features from Anderson et al. [33]'s model, which we employ for the object detection phase. This is because this model, being specialized for object detection, produces visual features for objects in bounding boxes that carry some degree of information about the image context. This is a desirable property when performing object detection, but not for our experiments.

[5]Our code is available at `https://github.com/InfoCogLab/info-sem-prag-neurips2024`.

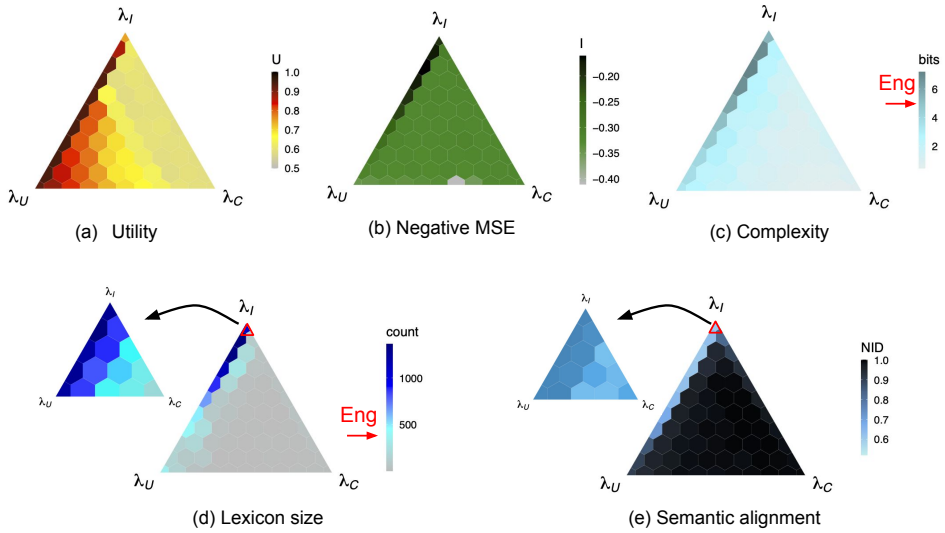

Figure 3: **The information-theoretic landscape of emergent communication**. Each simplex shows a test-time evaluation metric using either pragmatic (a) or semantic (b-e) settings, for the range of models spanned by the values of $\lambda = (\lambda_U, \lambda_I, \lambda_C)$. Overall, none of the extremes yield human-like communication, and the best alignment with English is achieved for $\lambda > 0$ near the top of the simplex. **(a)** Utility, reflecting the agents' pragmatic competence in solving the discrimination task. **(b)** Reconstruction loss, measured by the negative MSE between the speaker's and listener's target representations; values closer to 0 correspond to more informative communication (for ease of visualization, the model trained with $\lambda_U = 1$, with negative MSE -20, is ignored in this plot). **(c)** Complexity, measured by the mutual information between the speaker's inputs and communication signals. The red arrow indicates the complexity of English as estimated from the ManyNames dataset. **(d)** The effective lexicon size of the emergent systems (i.e., the number of signals that are used with non-zero probabilities). The red arrow indicates the number of unique English terms in the ManyNames dataset. **(e)** Semantic (mis)alignment between the emergent systems and English, measure by Normalized Information Distance (NID). Lower values correspond to better alignment. **(d-e) Insets** show the top of the simplex with higher resolution, revealing the importance of $\lambda_C > 0$ in attaining human-like lexicon sizes and better alignment with English.

in comparison to the number of English words used in the ManyNames dataset; and (e) the semantic (mis)alignment between the emergent systems and the English naming system, measured by the Normalized Information Distance [NID 37]. NID takes into account for full meaning-to-form mappings and is bounded between $[0, 1]$, with lower values corresponding to better alignment. The measures in (b-e) are evaluated in the semantic setting, which was not used during training, across the full dataset.

As expected, agents trained with large weights on one of the three λs develop communication systems skewed towards one component of the objective, yielding non-human like solutions at the extremes of the λ simplex. Next, we characterize the landscape of the emergent systems and discuss the human-like tradeoffs that best capture the linguistic behavior of English speakers.

High λ_U regime. High values of λ_U (bottom left corner) allow agents to achieve very high pragmatic competence (utility > 0.9, Figure 3a), maximizing the listener's success in solving the downstream task (i.e., distinguishing the target from the distractor), but result in very poor alignment between the emergent lexicon and the English lexicon (NID values > 0.8, Figure 3e). This finding resonates well with the cognitive science literature on the role of informational constraints in the evolution of human semantic systems [10], as well as with empirical findings that suggest that humans do not achieve maximal utility in our task when restricted to using only lexicalized items [26].[6] In other words, in this regime, which focuses primarily on utility maximization without information

[6]In our settings, agents can communicate only using lexicalized items. The usage of more complex structures, such as syntactic constructions, is beyond the scope of the present work and is left for future research.

https://doi.org/10.52202/079017-0663

constraints, agents are likely to develop communication strategies that do not align with human-like lexical semantics as they are pressured to over-compensate for the lack of syntactic constructions.

High λ_I regime. High values of λ_I (top corner) lead to highly informative communication (negative MSE > -0.20, Figure 3b). However, as expected, this comes at the cost of large lexicon sizes and high complexity. Agents trained with λ_I values close to 1 learn to use thousands of words for the ManyNames domain (> 1.5K vs. ~ 400 topnames used by English speakers for $\lambda_I = 1$, Figure 3d), and too complex communication systems (> 6 bits vs. ~ 5 bits in English, Figure 3c). Compared to λ_U, high values of λ_I seem to generally favour semantic alignment (NID ~ 0.6, Figure 3e), although it is important to note that the best NID is not achieved with $\lambda_I = 1$, but rather with a non-trivial combination of all tradeoffs, as discussed below. Finally, it is noteworthy that highly informative lexical systems also yield high utility (e.g., utility > 0.9 and negative MSE > -0.18 for $\lambda_U < 0.3$, $\lambda_I > 0.7$, $\lambda_C = 0.0$). This is in line with findings from Tucker et al. [38]. Indeed, in our framework, utility and informativeness are not in complete competition, but rather capture complementary aspects of successful communication that are only partially aligned.

High λ_C regime. High values of λ_C (bottom right corner) encourage minimal complexity and lead to unsuccessful communication, both in terms of utility and informativeness, as well as to very small lexicon sizes and low semantic alignment. In general, λ_C seems to act at a very small scale, with important effects on the lexicon already perceivable at very small values (see, for instance, the large decrease in lexicon size for $\lambda_C \sim 0.02$ in Figure 3d, as well as the corresponding increase in semantic alignment in Figure 3e).

Human-like tradeoffs ($\lambda^*_{\text{Eng}} > 0$). A combination of pressures fosters the emergence of natural solutions. Table 1 summarizes key properties of the landscape that we have explored. As expected, the three extremes on the simplex lead to unnatural, non-human-like solutions. Moving away from the extremes, we identify a non-trivial tradeoff, $\lambda = (\lambda_U = 0.005, \lambda_I = 0.98, \lambda_C = 0.015)$, that achieves the best fit with respect to the English data. We thus refer to this model as λ^*_{Eng}. This model achieves the best semantic alignment with English (Figure 3e), and roughly matches the English lexicon size (Figure 3d) and complexity rate (Figure 3c). It also achieves good reconstruction (Figure 3b) and high pragmatic competence (Figure 3a).

These quantitative findings support our predictions and demonstrate how our framework can advance our understanding of the co-evolution of semantics and pragmatics. To further understand the communication strategies learned by our agents, we next turn to a qualitative exploration of the emergent communication systems.

5.2 Qualitative evaluation

Figure 4 offers a visualization of the agents' communication. We plot as dots the visual features (reduced to 2D via PCA) of 500 objects randomly sampled from 3 categories in ManyNames ('woman', 'giraffe', and 'train', i.e. the most frequent names in the semantic domains of 'people', 'animals/plants', and 'vehicles'), identified by color. White crosses correspond to the listeners' reconstructions ($\hat{m}_t$, see Figure 2) in the semantic setting, which roughly represent word meanings.

Figure 4a, b, and c illustrate the solutions learnt by the models at the edges of our simplex. When trained with $\lambda_U = 1$ (Fig. 4a), agents are only driven by the task-related utility, i.e., maximizing success in pragmatic interactions. In this scenario, the listener does not learn to reconstruct a mental representation of the object, and the solution lacks a robust, non-contextual semantics. When trained with $\lambda_I = 1$ (Fig. 4b), in order to maximize informativeness, agents learn highly complex solutions, with large lexicons mapping words to small sets of objects, and not identifying human-like categories. When trained with $\lambda_C = 1$ (Fig. 4c), agents compress their lexicons at the cost of losing important distinctions, and achieve a solution where the same word describes all the objects. This solution does not enable successful communication.

In contrast, our λ^*_{Eng} model (Fig. 4d), trained with a tradeoff between utility, complexity, and informativeness, starts approaching a natural solution, learning word meanings that roughly map to the human categories. This solution is simple, yet it allows for informative communication and successful pragmatic interactions. Still, the agents seem to have learnt additional words, capturing spurious distinctions, especially for the peripheral areas of each category cluster, and for the category

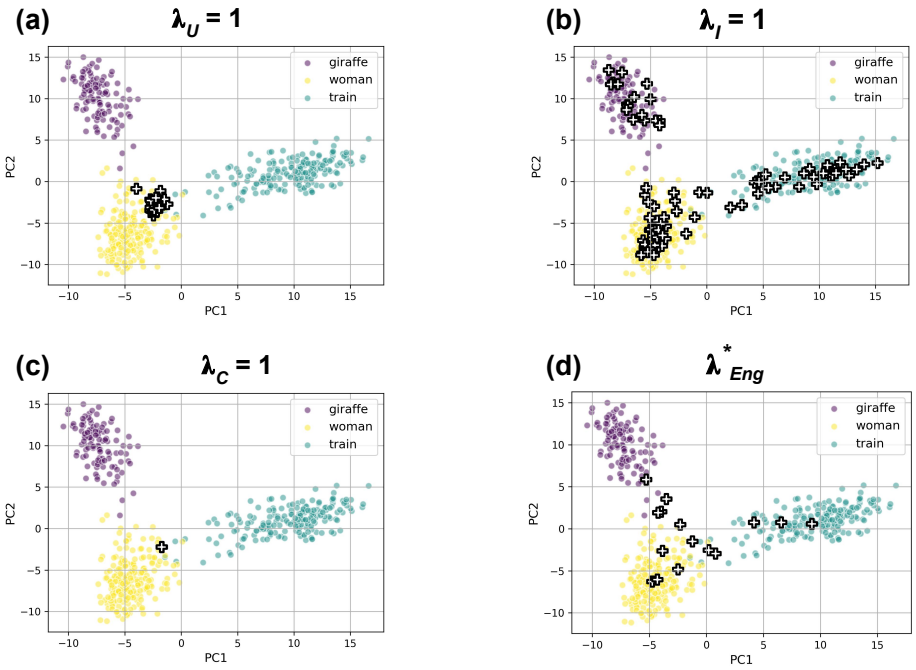

Figure 4: Visualization of the emergent communication systems for different combinations of λs (best seed for each). In each plot, the colored dots correspond to a 2D PCA of the input features of 500 randomly sampled objects from 3 distinct high-frequency categories in the ManyNames dataset ('woman', 'train', 'giraffe'). The white crosses correspond to the listener's reconstructions ($\hat{m}_t$, see Figure 2) given the speaker's communication signal, which roughly captures the meaning of each signal. **(a)** When agents are trained only with respect to utility, the listener does not learn to reconstruct a mental representation of the object. **(b)** When agents are trained only to maximize informativeness, the emergent communication system is highly complex, essentially assigning unique signals to much of the input space. **(c)** When agents are trained only to minimize complexity, a trivial non-informative system emerges that employs only one signal. **(d)** When agents are trained with respect to a non-trivial tradeoff between utility, complexity, and accuracy, a more human-like communication system emerge. These systems are as simple as possible while maintaining sufficient utility and informativeness.

'woman' (the most frequent one in ManyNames). Besides reasons related to the data distribution, our hypothesis is that the agents may have developed human-like categorizations for prototypical members of the human categories (i.e. those with visual features near the center of the category cluster), but also additional, non human-like categories for atypical objects. Indeed, atypical objects are harder to categorize for humans as well, and may not clearly belong to one single semantic category [39, 40]. This hypothesis may also explain why our λ^*_{Eng} model has an NID score of 0.55, showing moderate, but not perfect, alignment with human semantics. We gather support for this explanation by computing the NID score only with objects prototypical for their category: on this set of objects, the NID for λ^*_{Eng} decreases to 0.45 (± 0.02), suggesting higher alignment –see Appendix B for further details.

6 Conclusion

Words in the human lexicon are associated with non-contextual meanings, as well as shaped by the local conversational context. In this work, we have addressed a key open question for language evolution: How can a shared lexicon emerge from local context-sensitive interactions? We modeled the semantics-pragmatics interface by building on a framework for information-theoretic emergent communication in neural agents. We trained agents to interact in self-play in the presence of a shared

https://doi.org/10.52202/079017-0663

conversational context, guiding them with combinations of cognitively-motivated pressures. We then tested their pragmatic competence, as well as the human-likeliness of their emergent semantics. By exploring the landscape of emerging artificial languages, we demonstrate that, if trained with pressures for both context-specific utility and general communicative constraints, agents learn systems with key human-like properties and that allow for successful pragmatic interactions. Our findings inform current theories of language evolution, and show that cognitively-motivated optimization principles can facilitate the emergence of human-like communication strategies in neural networks.

7 Limitations

Our work aims to better understand the computational principles that underlie language evolution, in humans and artificial agents, with focus on the interface between semantics and pragmatics. While we presented an important first step towards this goal, we were only able to evaluate our model on English data and focused on the lexicon only. An important direction for future work is the evaluation of our model on a larger, more diverse set of languages. In addition, our work has focused only on the use of lexical items in communication. Therefore, another important direction for future research is to extend our framework to more complex communication structures, such as syntax, morphology, and compositional meaning.

One potential concern about our model is that it employs a pre-trained object classification model [35] to extract visual features. This pre-trained model was trained with classification labels. Thus, one might worry that our agents were implicitly exposed to some linguistic knowledge. We have several reasons to believe that this exposure is negligible. First, the classification labels are coarse while we evaluate the agents with respect to fine-grained naming data. Second, our agents are trained in a pragmatic setting, whereas the classification labels are non-contextual. Third, the majority of our agents develop non-human-like lexical systems, suggesting that the pretrained vision component is not sufficient for alignemnt with English. Having said that, in further work we intend to explore the influence of other types of visual features.

Finally, the NID score achieved by our agents, while encouraging, suggests that even our best-performing agents are not yet fully aligned with (English speaking) humans. Therefore, further research is needed to understand how to close this gap and guide our agents toward more human-like communication systems.

Acknowledgements

This research was partially by supported by grant PID2020-112602GB-I00/MICIN/AEI/10.13039/501100011033 from the Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación (Spain) and the European Union's Horizon 2020 research and innovation programme (grant agreement No. 715154). We thank the COLT group from Universitat Pompeu Fabra for feedback on this work.

References

[1] Herbert P. Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Speech Acts*, volume 3, pages 41–58. Academic Press, 1975. URL `http://www.ucl.ac.uk/ls/studypacks/Grice-Logic.pdf`.

[2] Michael Franke. Signal to act: Game theory in pragmatics. 01 2009.

[3] Noah D. Goodman and Michael C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829, 2016.

[4] Mycal Tucker, Roger P. Levy, Julie Shah, and Noga Zaslavsky. Trading off utility, informativeness, and complexity in emergent communication. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=O5arhQvBdH`.

[5] Anton Benz and Jon Stevens. Game-theoretic approaches to pragmatics. *Annual Review of Linguistics*, 4, 02 2018. doi: 10.1146/annurev-linguistics-011817-045641.

[6] Serhii Havrylov and Ivan Titov. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[7] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6382–6393, Red Hook, NY, USA, 2017. Curran Associates Inc.

[8] Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. Communicating artificial neural networks develop efficient color-naming systems. *Proceedings of the National Academy of Sciences*, 118(12):e2016569118, 2021. doi: 10.1073/pnas.2016569118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2016569118.

[9] Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. Efficient compression in color naming and its evolution. *PNAS*, 115(31):7937–7942, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1800521115. URL http://www.pnas.org/content/115/31/7937.

[10] Noga Zaslavsky. *Information-Theoretic Principles in the Evolution of Semantic Systems*. Ph.D. Thesis, The Hebrew University of Jerusalem, 2020.

[11] Noga Zaslavsky, Mora Maldonado, and Jennifer Culbertson. Let's talk (efficiently) about us: Person systems achieve near-optimal compression. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, 2021.

[12] Francis Mollica, Geoff Bacon, Noga Zaslavsky, Yang Xu, Terry Regier, and Charles Kemp. The forms and meanings of grammatical markers support efficient communication. *Proceedings of the National Academy of Sciences*, 118(49), 2021. ISSN 0027-8424. doi: 10.1073/pnas.2025993118.

[13] Noga Zaslavsky, Karee Garvin, Charles Kemp, Naftali Tishby, and Terry Regier. The evolution of color naming reflects pressure for efficiency: Evidence from the recent past. *Journal of Language Evolution*, 04 2022. ISSN 2058-458X. lzac001.

[14] Noga Zaslavsky, Jennifer Hu, and Roger Levy. A Rate–Distortion view of human pragmatic reasoning. 2020. URL https://arxiv.org/abs/2005.06641.

[15] Claude E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142-163):1, 1959.

[16] Toby Berger. *Rate distortion theory; a mathematical basis for data compression*. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[17] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999. URL https://arxiv.org/abs/physics/0004057.

[18] Carina Silberer, Sina Zarrieß, and Gemma Boleda. Object naming in language and vision: A survey and a new dataset. In *Proceedings of LREC*, pages 5792–5801, Marseille, France, 2020. European Language Resources Association.

[19] Eleanor Rosch and Carolyn B Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4):573–605, 1975. ISSN 0010-0285.

[20] Gregory Murphy. *The Big Book of Concepts*. MIT Press, 2004.

[21] Stephen C. Levinson. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press, Cambridge, 2000.

[22] Terry Regier, Charles Kemp, and Paul Kay. *Word Meanings across Languages Support Efficient Communication*, pages 237–263. 01 2015. ISBN 9781118301753. doi: 10.1002/9781118346136.ch11.

 https://doi.org/10.52202/079017-0663

[23] Charles Kemp, Yang Xu, and Terry Regier. Semantic typology and efficient communication. *Annual Review of Linguistics*, 4:109–128, 01 2018.

[24] Noga Zaslavsky, Terry Regier, Naftali Tishby, and Charles Kemp. Semantic categories of artifacts and animals reflect efficient coding. In Ashok K. Goel, Colleen M. Seifert, and Christian Freksa, editors, *Proceedings of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019: Creativity + Cognition + Computation, Montreal, Canada, July 24-27, 2019*, pages 1254–1260. cognitivesciencesociety.org, 2019. URL https://mindmodeling.org/cogsci2019/papers/0229/index.html.

[25] Caroline Graf, Judith Degen, Robert D. Hawkins, and Noah D. Goodman. Animal, dog, or dalmatian? level of abstraction in nominal referring expressions. *Cognitive Science*, 2016. URL https://api.semanticscholar.org/CorpusID:9066747.

[26] Andreas Mädebach, Torubarova Ekaterina, Eleonora Gualdoni, and Gemma Boleda. Effects of task and visual context on referring expressions using natural scenes. In J. Culbertson, A. Perfors, and V. Rabagliati, H. & Ramenzoni, editors, *Proceedings of the 44th Annual Conference of the Cognitive Science Society*. Cognitive Science Society, 2022.

[27] Julia White, Jesse Mu, and Noah D. Goodman. Learning to refer informatively by amortizing pragmatic reasoning. *CoRR*, abs/2006.00418, 2020. URL https://arxiv.org/abs/2006.00418.

[28] Thomas Brochhagen, Michael Franke, and Robert van Rooij. Coevolution of lexical meaning and pragmatic use. *Cognitive Science*, 42(8):2757–2789, 2018. doi: https://doi.org/10.1111/cogs.12681. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12681.

[29] Bill McDowell and Noah D. Goodman. Learning from omission. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 619–628, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1059.

[30] Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[31] Mycal Tucker, Roger P. Levy, Julie Shah, and Noga Zaslavsky. Generalization and translatability in emergent communication via informational constraints. In *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems*, 2022. URL https://openreview.net/forum?id=yf8suFtNZ5v.

[32] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[33] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of CVPR*, 2018.

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 39, 06 2015.

[35] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[36] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32 – 73, 2016. URL https://api.semanticscholar.org/CorpusID:4492210.

[37] Nguyen Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *JMLR*, 11:2837–2854, 10 2010.

[38] Mycal Tucker, Roger P. Levy, Julie Shah, and Noga Zaslavsky. Trading off utility, informativeness, and complexity in emergent communication. In *RSS Workshop: Social Intelligence in Humans and Robots*, 2022. URL `https://social-intelligence-human-ai.github.io/docs/camready_4.pdf`.

[39] Joan Gay Snodgrass and Mary Vanderwart. A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology. Human learning and memory*, 6 2:174–215, 1980.

[40] Eleonora Gualdoni, Thomas Brochhagen, Andreas Mädebach, and Gemma Boleda. What's in a name? A large-scale computational study on how competition between names affects naming variation. *Journal of Memory and Language*, 133:104459, 2023.

 https://doi.org/10.52202/079017-0663

A Details about agents' training

We started by training our agents with $\lambda_U = 1$ until convergence, i.e. for $10K$ epochs. After that, we first annealed models by keeping λ_C fixed at 0, while gradually decreasing the value of λ_U and increasing the value of λ_I, until reaching $\lambda_I = 1$. Then, for each trained value of λ_U, we gradually annealed λ_C. For each annealing step, we trained until reaching variance in the training objective lower than 0.0001 for the latest $1K$ epochs (as a criterion for convergence). We followed a non-uniform annealing schedule, re-fined after an initial exploration phase, aimed at identifying regions in the simplex showing interesting patterns with respect to our metrics. In these regions, we sampled the combinations of parameters more densely, e.g. with steps of 0.002, while in the other regions we sampled with steps of 0.1. Agents were trained with batch size 128, hyperparameter β of the VQ architecture set at 1. See Tucker et al. [38] for further details about the architecture and hyperparameters.

Experiments were run on a cluster with 12 nodes with 5 NVIDIA A30 GPUs and 48 CPUs each. Training the $\lambda_U = 1$ model took around 30 minutes. Training one annealed model could take up to 15 minutes, often less. Computing evaluation metrics took a total of 10 hours. We estimate the overall time required to run this analysis to be around 4 days. Considering our exploration phase and failed experiments, we estimate the total runtime required by this paper to have been around 20 days.

B Identifying prototypical objects

To identify in ManyNames objects that are prototypical members of their categories, we took the following approach: for each topname appearing at least 30 times in ManyNames, we selected the 15 most probable images based on the human annotations (in general, visual typicality for a name correlates with name probability [39, 40]). This process resulted in 69 words and a total of 1035 images. These images are, at worst, the 50% most typical images for their human name. On this set, λ^*_{Eng} achieves NID of 0.45 (± 0.02).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state our research question, and how our findings contribute to answer that question.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: see Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

https://doi.org/10.52202/079017-0663

Justification: Our paper does not contain theoretical results in terms of theorems and proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our code is available at `https://github.com/InfoCogLab/info-sem-prag-neurips2024`. We describe in detail our model (Section 3 and Appendix A), and reference the original paper where the model's core components were first presented. We provide information about our data, data selection strategies, and models' training.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Our code is available at `https://github.com/InfoCogLab/info-sem-prag-neurips2024`

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: see Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: yes, when we report the numerical results reported in Table 1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

https://doi.org/10.52202/079017-0663

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: see Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer:[Yes]

   Justification: No potential harms are identified.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: we consider this paper foundational research, with no risk of malicious use.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not train models at high risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: see Section 4.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

https://doi.org/10.52202/079017-0663

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: we do not release any new dataset. The experiments' code will be released upon acceptance.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer:[NA]

    Justification: The paper does not involve crowdsourcing

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.