The Evolution of Metaphorical Language: A Formal Model of Figurative Communication

Augusto Antonio Basilico[0009−0009−2442−9796]

augustoantoniobasilico@gmail.com

Abstract. Metaphorical language is a pervasive feature of human communication, yet its evolutionary origins remain poorly understood. Why do speakers routinely convey meanings through figurative expressions that are literally false, rather than relying on available literal descriptions? In this paper, we present a formal model of the cultural evolution of metaphorical language, combining the Rational Speech Act framework with evolutionary game theory. Our model embeds pragmatic agents, who engage in recursive reasoning about goals and meanings, into a replicator–mutator dynamic where strategies are transmitted via Bayesian learning. Our central hypothesis is that metaphor emerges as an adaptive response to an evolutionary trade-off between communicative accuracy and cognitive cost: when speakers aim to convey complex clusters of features, repurposing simple conceptual categories metaphorically can be more efficient than encoding those clusters through literal strategies. To illustrate this, we analyze a highly simplified scenario where communicative goals are dispersive, requiring the transmission of feature clusters rather than isolated features. This case study shows that, under such conditions, metaphorical strategies can outcompete literal ones.

Keywords: metaphor, language evolution, evolutionary game theory, rational speech act framework, bayesian learning, replicator–mutator dynamics, formal pragmatics, evolutionary pragmatics

1 Introduction

Metaphor is one of the most striking and ubiquitous features of human language. Speakers routinely describe referents in one domain using expressions drawn from familiar categories in another domain—for example, calling a dangerous and unscrupulous person a "shark," or a wise and intelligent one an "owl." Such utterances are literally false, yet they reliably convey information about relevant features of the intended referent [9,23,28]. But why should a language community rely on these roundabout communicative devices, rather than sticking to literal descriptions?

Formal pragmatics has shed new light on this question. The Rational Speech Act (RSA) framework models communication as recursive reasoning between cooperative agents [6,7,11], and can be seen as a probabilistic formalization of

Grice's maxims of conversation [12,13]. When extended with the Question Under Discussion (QUD) framework [29,30,31], RSA captures a variety of non-literal uses of language—including metaphor, hyperbole, and irony—by treating them as informative with respect to shared communicative goals [20,21]. This line of work explains how listeners can make sense of figurative utterances in a given communicative exchange. However, RSA models are largely *synchronic*: they presuppose fixed lexica and are not concerned with how communicative strategies emerge, spread, and stabilize in a population over time.[1]

Evolutionary game theory (EGT) [17,18,24], by contrast, has been applied to study how communicative strategies spread and stabilize in populations through processes of learning and differential communicative success [26,27,32]. Within the emerging field of evolutionary pragmatics [10], EGT has been applied to phenomena such as scalar implicature [2,3] and hyperbole [22]. However, metaphor—despite its centrality to human communication—has not, to our knowledge, been given a formal evolutionary treatment. This leaves an important gap: we do not yet know whether there are trade-offs under which metaphorical strategies become evolutionarily viable, and if so, what their nature might be in relation to communicative accuracy and cognitive cost in both communication and learning.

This paper addresses this gap by presenting a formal model of the cultural evolution of metaphorical communication. The model integrates RSA-based pragmatic reasoning with *replicator–mutator dynamics* [25]—a standard tool within EGT—and Bayesian learning [14,15], building on the work of Brochhagen *et al.* [2,3] on the evolution of scalar implicatures. In our model, agents share a simple conceptual framework with *category terms* (e.g., *shark*, *owl*) and *feature terms* (e.g., *dangerous*, *clever*). Communicative strategies differ in how they express feature configurations: literal strategies express the features of the referent directly, whereas metaphorical strategies convey them indirectly by reusing category terms that do not literally apply to the referent. Strategies are transmitted across generations through Bayesian learning, and their evolutionary success depends jointly on communicative accuracy and complexity cost.

To illustrate the model, we then examine a highly simplified but revealing case study in which communicative goals are *dispersive*, i.e., they require agents to convey entire clusters of features at once. In this setting, metaphorical strategies dominate, since reusing simple category terms is more efficient than laboriously constructing complex literal descriptions from combinations of feature terms. This simple case study shows how metaphor *can* emerge as an adaptive solution to competing pressures for accuracy and simplicity in language evolution. At the same time, it does not exhaust the modeling power of our framework, which can be extended to alternative goal structures and richer and more complicated communicative scenarios.

[1] That is, RSA accounts are designed to model production and comprehension within a single communicative interaction, rather than the *diachronic* processes by which languages and strategies evolve across generations [3,2].

2 Model Description

2.1 Syntax, Semantics, and the Shared Conceptual Framework

We begin by specifying the representational resources available to all agents in the linguistic community. All individuals in the population share a *common language* defined by a fixed syntax and a shared probabilistic semantics. This language distinguishes between two types of terms:

- *category terms* $C = \{c_1, \ldots, c_n\}$, each corresponding to a cognitively salient prototype such as *shark*, *owl*, or *puppy*.
- *feature terms* $\Phi = \{\phi_1, \ldots, \phi_m\}$, each denoting a binary property such as *dangerous*, *clever*, or *strong*.

Well-formed formulas. The set of well-formed formulas WFF is generated inductively according to the following grammar in Backus–Naur Form:

$$\langle \text{Atom} \rangle ::= c_i = 1 \mid \phi_j = 1 \mid \phi_j = 0 \qquad (1 \le i \le n,\ 1 \le j \le m) \tag{1}$$

$$\langle \text{WFF} \rangle ::= \langle \text{Atom} \rangle \mid \neg\langle \text{WFF} \rangle \mid (\langle \text{WFF} \rangle \wedge \langle \text{WFF} \rangle) \mid (\langle \text{WFF} \rangle \vee \langle \text{WFF} \rangle) \tag{2}$$

Examples in Atom may include "$c_{shark} = 1$" ("the referent is a shark"), "$\phi_{dangerous} = 1$" ("the referent is dangerous"), and "$\phi_{dangerous} = 0$" ("the referent is harmless").[2] Instead, an example in WFF may be "$c_{puppy} = 1 \wedge \phi_{clever} = 1$", which states that the referent is a clever puppy.

Complexity. The complexity of a formula $\mathsf{C}(u)$ is defined recursively:

$$\begin{aligned} \mathsf{C}(u) &= 1 \text{ if } u \in \text{Atom}, \\ \mathsf{C}(\neg u) &= 1 + \mathsf{C}(u), \\ \mathsf{C}(u \wedge v) &= 1 + \mathsf{C}(u) + \mathsf{C}(v), \\ \mathsf{C}(u \vee v) &= 1 + \mathsf{C}(u) + \mathsf{C}(v). \end{aligned} \tag{3}$$

This measure will later serve as a proxy for cognitive cost, impacting both communication and learning.

Semantic framework. Semantics is defined relative to a shared conceptual framework $\mathcal{F} = (\Omega, 2^{\Omega}, P)$, which constitutes the *common ground* [5,33,34] among linguistic agents:

- Ω is the set of possible kinds of objects (referents). More specifically, each $\omega \in \Omega$ is a total assignment $\omega : C \cup \Phi \to \{0, 1\}$;

[2] For features, the two opposite values are symmetric: both can be expressed as atomic formulas of equal complexity (e.g., "$\phi_{strong} = 1$" for *strong* and "$\phi_{strong} = 0$" for *weak*). Categories, by contrast, are asymmetric: only positive membership is atomic (e.g., "$c_{shark} = 1$"), while negative membership must be expressed by a compound formula such as "$\neg(c_{shark} = 1)$", which increases complexity.

- 2^Ω is the σ-algebra of all subsets of Ω;
- P is a probability distribution over 2^Ω, encoding prior beliefs about the world that all agents share.

The semantic value of a formula $u \in \text{WFF}$ is the set of samples $\omega \in \Omega$ in which it holds. This is defined recursively:

$$[\![c_i = 1]\!]_\mathcal{F} = \{\omega \in \Omega \mid \omega(c_i) = 1\}, \tag{4}$$
$$[\![f_j = 1]\!]_\mathcal{F} = \{\omega \in \Omega \mid \omega(f_j) = 1\}, \tag{5}$$
$$[\![f_j = 0]\!]_\mathcal{F} = \{\omega \in \Omega \mid \omega(f_j) = 0\}, \tag{6}$$
$$[\![\neg u]\!]_\mathcal{F} = \Omega \setminus [\![u]\!]_\mathcal{F}, \tag{7}$$
$$[\![u \wedge v]\!]_\mathcal{F} = [\![u]\!]_\mathcal{F} \cap [\![v]\!]_\mathcal{F}, \tag{8}$$
$$[\![u \vee v]\!]_\mathcal{F} = [\![u]\!]_\mathcal{F} \cup [\![v]\!]_\mathcal{F}. \tag{9}$$

Communicative strategies. To capture cognitive and expressive limits, we assume a maximum complexity m, restricting agents to:

$$\text{WFF}_m = \{u \in \text{WFF} \mid \mathsf{C}(u) \leq m\}. \tag{10}$$

Each agent is endowed with a finite *strategy lexicon* $L \subseteq \text{WFF}_m$ of fixed size k, representing the set of expressions they are able and willing to use as speakers. While agents are assumed to be capable of interpreting any formula in WFF_m when acting as listeners, their productive vocabulary as speakers is restricted to L. This models the cognitive and expressive asymmetry between linguistic comprehension and production: humans are often capable of understanding expressions they would not naturally produce.[3] In our framework, this constraint allows us to formalize differences between strategies in terms of their expressive capacity and cognitive load, which in turn influence their evolutionary viability.

Let $\mathscr{L}_m^k \subseteq 2^{\text{WFF}_m}$ be the set of all size-k subsets of well-formed formulas with complexity at most m. In other words, we define the space of possible speaker lexicons as:

$$\mathscr{L}_m^k := \{L \subseteq \text{WFF}_m \mid |L| = k\}. \tag{11}$$

Each *communicative strategy* is defined as a pair $(L, n) \in \mathscr{L}_m^k \times \{0, 1, \ldots, N\}$, where n is the *depth* of recursive pragmatic reasoning [3,35]. We denote the full space of strategies as:

$$\mathscr{S}_{m,k,N} := \mathscr{L}_m^k \times \{0, 1, \ldots, N\}. \tag{12}$$

This space captures the key cognitive and expressive dimensions over which agents may differ: their available utterances as speakers (lexicon L) and their inferential/pragmatic sophistication (depth n). In evolutionary simulations, selection acts over this space of strategies, favoring those that optimize a trade-off between communicative success and cognitive cost.

[3] Such asymmetries are empirically well-attested—humans often understand more than they can fluently produce [1,4,19].

2.2 Pragmatic Reasoning and Communicative Strategies: The RSA–QUD Component

Communication in the model is framed as a cooperative inferential process between speakers and listeners. In each exchange, the speaker observes an object characterized by a complete feature configuration f and has a communicative goal g. Importantly, we assume that every object in the environment has a complete feature configuration, and that none of them literally instantiate the category terms of the language: for instance, no object is itself a "shark" or an "owl." Given the pair (f, g), the speaker selects a message $u \in L$ from their lexicon to convey the goal-relevant aspects of f to the listener. The listener, in turn, interprets the utterance by inferring both the intended goal and the corresponding projection of the feature configuration. This interaction follows the RSA paradigm, extended with the QUD framework [30,31].

Feature configurations. A *feature configuration* is a function

$$f : \Phi \to \{0, 1\}, \tag{13}$$

assigning to each feature $\phi_j \in \Phi$ a binary value 1 (present) or 0 (opposite feature present). The set of all possible feature configurations is denoted by $\mathfrak{F} = \{0, 1\}^{\Phi}$. Each configuration $f \in \mathfrak{F}$ induces an event in the underlying sample space Ω:

$$[\![f]\!]_{\mathcal{F}} := \{w \in \Omega \mid w \restriction_{\Phi} = f\} \tag{14}$$

where $w \restriction_{\Phi}$ denotes the restriction of a sample $w \in \Omega$ to the feature domain Φ.

Communicative goals as QUDs. Each interaction is oriented toward a communicative goal $g \subseteq \Phi$, which we interpret as a QUD [29]. Intuitively, a goal identifies *which features of the referent object are at stake in the current communicative exchange*, and thus which dimensions of meaning are relevant for successful communication. In other words, speakers do not aim to convey the entire feature configuration of a referent, but only the projection of that configuration onto the feature set singled out by the goal [30,31]. We denote by $\mathcal{G}$ the set of all possible communicative goals, that is, the space of feature subsets that can be pragmatically relevant in a given communicative context.

Formally, given a full feature configuration f, its projection onto the goal $g \in \mathcal{G}$ is defined (with a slight abuse of notation) as

$$g(f) := f \restriction_{g}, \tag{15}$$

i.e. the restriction of f to the feature subset g. This projected configuration induces an event in the shared conceptual framework:

$$[\![g(f)]\!]_{\mathcal{F}} := \{w \in \Omega \mid w \restriction_{g} = g(f)\}. \tag{16}$$

Thus the goal g determines which features matters for communication in a specific exchange, and communication is successful just in case the listener can recover the goal-relevant projection of the feature configuration observed by the speaker.

Goal-equivalence. Two feature configurations $f, f' \in \mathfrak{F}$ are said to be *goal-equivalent* under goal $g \in \mathcal{G}$ if

$$g(f) = g(f'). \tag{17}$$

Intuitively, goal-equivalent configurations agree on the subset of features that matter for the communicative task.

RSA agent behavior. Given a strategy $(L, n) \in \mathscr{S}_m^k$, an agent following it behaves as follows: as a *speaker*, it selects utterances according to the distribution $\mathcal{S}_n(u \mid f, g, L)$, where it is constrained to produce only utterances $u \in L$; as a *listener*, it interprets utterances according to distribution $\mathcal{L}_n(f, g \mid u)$, which is defined for all utterances, not only those in L. Both the speaker distribution $\mathcal{S}_n$ and the listener distribution $\mathcal{L}_n$ are defined by recursively applying RSA-style equations up to depth n, starting from the literal speaker $\mathcal{S}_0$ and the literal listener $\mathcal{L}_0$:[4]

$$\mathcal{S}_0(u \mid f, g, L) \propto \mathbf{1}_{[u \in L]} \cdot \left[P\left([\![u]\!]_{\mathcal{F}} \,\middle|\, [\![g(f)]\!]_{\mathcal{F}} \cap \left(\bigcup_{i=1}^{n} [\![c_i = 1]\!]_{\mathcal{F}} \right)^{c} \right) \right]^{\lambda}, \tag{18}$$

$$\mathcal{L}_0(f, g \mid u) \propto P_{\mathcal{G}}(g) \cdot P\left([\![f]\!]_{\mathcal{F}} \mid [\![u]\!]_{\mathcal{F}} \right), \tag{19}$$

$$\mathcal{S}_k(u \mid f, g, L) \propto \mathbf{1}_{[u \in L]} \cdot \exp\left[\lambda \cdot \left(\log\left(\sum_{f'} \mathbf{1}_{[g(f) = g(f')]} \cdot \mathcal{L}_{k-1}(f', g \mid u) \right) - \mathsf{C}(u) \right) \right], \tag{20}$$

$$\mathcal{L}_k(f, g \mid u) \propto P\left([\![f]\!]_{\mathcal{F}} \right) \cdot P_{\mathcal{G}}(g) \cdot \sum_{L \in \mathscr{L}_m^k} P_{\mathscr{L}}(L) \cdot \mathcal{S}_k(u \mid f, g, L). \tag{21}$$

Equation (18) defines the behavior of the *literal speaker* $\mathcal{S}_0$, which selects utterances u when presented with a feature configuration f under communicative goal g. The indicator function $\mathbf{1}_{[u \in L]}$ ensures that the speaker produces only utterances contained in its lexicon L.

A central assumption of the model is that the *actual* environment contains objects that do not literally instantiate any of the category terms (such as *shark* or *owl*). In other words, speakers have to communicate about objects that never fall under the literal denotation of any of the modeled shared categories: in consequence, utterances of the form "$c_i = 1$" are always literally false and cannot be used by a literal speaker. To capture this restriction, the literal speaker's probability of producing an utterance $u \in L$ is conditioned on the event that the intended referent lies outside the extension of all known categories. Formally, this is represented by a conditional probability of the form:

$$P\left([\![u]\!]_{\mathcal{F}} \,\middle|\, [\![g(f)]\!]_{\mathcal{F}} \cap \left(\bigcup_{i=1}^{n} [\![c_i = 1]\!]_{\mathcal{F}} \right)^{c} \right). \tag{22}$$

[4] These equations draw on Kao *et al.* [20] for metaphor understanding and on the general RSA framework with QUD [30,31].

This quantity gives the probability that utterance u holds true of the intended referent, on the condition that the referent satisfies the goal-relevant feature projection $[\![g(f)]\!]_{\mathcal{F}}$ but lies outside the extension of all category terms $c_1, \ldots, c_n$. Finally, the exponential weighting in (18) is governed by the rationality parameter $\lambda > 0$.[5]

Equation (19) defines the behavior of the *literal listener* $\mathcal{L}_0$, who, upon hearing utterance u, infers both the feature configuration f and the communicative goal g. The literal listener is not constrained by any lexicon and interprets u solely based on its literal denotation. Interpretation is guided by a prior over goals $P_{\mathcal{G}}(g)$ and a conditional probability $P([\![f]\!]_{\mathcal{F}} \mid [\![u]\!]_{\mathcal{F}})$, which evaluates how likely the feature configuration f is, given the literal meaning of u.

Equation (20) defines the behavior of the *pragmatic speaker* $\mathcal{S}_k$, who chooses utterances by reasoning about how a listener of depth $k-1$ would interpret them. As in (18), the indicator function $\mathbf{1}_{[u \in L]}$ ensures that only utterances contained in the speaker's lexicon L receive non-zero probability. The choice of the utterance is guided by a compromise between *informativeness* and *cost* [6,8]. Informativeness is measured by the logarithm of the sum $\sum_{f'} \mathbf{1}_{[g(f)=g(f')]} \cdot \mathcal{L}_{k-1}(f', g \mid u)$, which ranges over all feature configurations f' but includes only those that are *goal-equivalent* to f under goal g, in the sense of (17), that is, those configurations whose goal-relevant projection coincides with $g(f)$. For these configurations, the term evaluates how likely a level-$(k-1)$ listener would recover them upon hearing u. The cost term $\mathsf{C}(u)$—as defined in (3)—penalizes utterances that are complex or otherwise costly to produce, so that the resulting distribution balances informativeness against cost.[6]

Equation (21) defines the behavior of the *pragmatic listener* $\mathcal{L}_k$. Upon hearing an utterance u, the listener jointly infers the feature configuration f and the communicative goal g. This inference is guided by three components. First, the prior $P([\![f]\!]_{\mathcal{F}})$ encodes expectations about feature configurations in the environment. Second, the prior $P_{\mathcal{G}}(g)$ captures the relative likelihood of different communicative goals. Finally, the summation marginalizes over possible lexica $L \in \mathscr{L}_m^k$, weighted by their prior probability $P_{\mathscr{L}}(L)$, and incorporates the behavior of a level-k speaker $\mathcal{S}_k(u \mid f, g, L)$, which specifies the likelihood that a speaker using strategy (L, k) would have produced the utterance u when trying to communicate about (f, g). Thus, the pragmatic listener of depth k is modeled as an agent who combines priors over configurations and goals with expectations about speaker strategies, integrating over lexical uncertainty.

[5] For $\lambda \to 0$, the literal speaker's behavior approaches randomness; as $\lambda \to +\infty$, the speaker deterministically chooses the utterance in L that maximizes the conditional probability given by (22)—that is, the utterance most likely to hold of the intended referent given the goal projection $g(f)$ and the restriction that the referent lies outside the extension of all category terms.

[6] Furthermore, the exponential form corresponds to a *softmax function*, where the rationality parameter $\lambda > 0$ determines how sharply the speaker concentrates on the best utterances, with $\lambda \to 0$ yielding near-random behavior and $\lambda \to +\infty$ yielding deterministic choice of the maximally informative, least costly utterance.

Strategies and prior biases. We define a prior over strategies $(L, n) \in \mathscr{S}_{m,k,N}$ as the product of a prior over lexica and a prior over reasoning depth:

$$P_{\mathscr{S}}((L, n)) = P_{\mathscr{L}}(L) \cdot P_{\mathscr{D}}(n). \tag{23}$$

The lexicon prior $P_{\mathscr{L}}(L)$ penalizes utterance complexity. Since each utterance $u \in \mathrm{WFF}_m$ has a well-defined complexity cost $\mathsf{C}(u)$ given by (3), we define:

$$P_{\mathscr{L}}(L) \propto \frac{1}{\sum_{u \in L} \mathsf{C}(u)}, \tag{24}$$

so that lexica composed of simpler utterances are favored.

The depth prior $P_{\mathscr{D}}(n)$ favors shallower pragmatic reasoning, reflecting cognitive and computational limitations. It is assumed to be a decreasing function of n, for example:

$$P_{\mathscr{D}}(n) \propto e^{-\alpha \cdot n}, \quad n = 0, \ldots, N, \tag{25}$$

for some fixed $\alpha > 0$.

Together, these priors model learning biases that promote low-complexity, cognitively accessible strategies, increasing their evolutionary viability: simpler lexica and shallower reasoning are more learnable and therefore more likely to spread.

2.3 Bayesian Learning of Communicative Strategies

Strategies are not inherited directly but transmitted culturally via learning. Each agent acquires a strategy $(L, n) \in \mathscr{S}_{m,k,N}$ by observing a finite dataset of communicative behavior produced by a teacher. This process is modeled as Bayesian inference over strategies [3,14,15].

Data observed by a learner. A learner is exposed to a k-length dataset

$$d = \langle (f_1, u_1), \ldots, (f_k, u_k) \rangle, \tag{26}$$

where each pair (f_i, u_i) consists of an observed feature configuration $f_i \in \mathfrak{F}$ and the corresponding utterance $u_i \in \mathrm{WFF}_m$ produced by the teacher.

Posterior inference. Given data d, the learner infers a posterior distribution over possible teacher strategies $(L, n) \in \mathscr{S}_{m,k,N}$:

$$P_{\mathscr{S}}((L, n) \mid d) \propto P_{\mathscr{S}}((L, n)) \cdot P_{\mathscr{S}}(d \mid (L, n)). \tag{27}$$

where, as we have seen in (23), the prior over strategies decomposes into independent components, i.e., $P_{\mathscr{S}}((L, n)) = P_{\mathscr{L}}(L) \cdot P_{\mathscr{D}}(n)$.

Likelihood. The likelihood of observing dataset d given (L, n) as the teacher's strategy is:

$$P_{\mathscr{S}}(d \mid (L, n)) \propto \prod_{i=1}^{k} \sum_{g} P_{\mathcal{G}}(g) \cdot \mathcal{S}_n(u_i \mid f_i, g, L), \tag{28}$$

where $\mathcal{S}_n$ is the speaker of depth n defined in Eqs. (18)-(21).

Learning precision. To capture variability in how faithfully learners adopt inferred strategies, we introduce a learning precision parameter ℓ [3]. The effective learning posterior, i.e. the probability that a learner acquires strategy (L, n) from dataset d, is given by:

$$P_{\text{learn}}((L, n) \mid d) \propto [P_{\mathscr{S}}((L, n) \mid d)]^{\ell}. \tag{29}$$

When $\ell \to 0$, learning becomes uniformly random; when $\ell \to +\infty$, the learner adopts the Maximum a Posteriori (MAP) strategy.

From individual learning to population mutation. Let a teacher use strategy $(L', n') \in \mathscr{S}_{m,k,N}$. The *mutation matrix* gives then the probability that a learner exposed to a teacher of type (L', n') acquires strategy (L, n):

$$Q((L', n'), (L, n)) \;=\; \sum_{d} P_{\text{learn}}((L, n) \mid d) \;\cdot\; P_{\mathscr{S}}(d \mid (L', n')). \tag{30}$$

This Q is the cultural transmission channel that will appear as the mutation term in the replicator–mutator dynamics.

2.4 Cultural Evolution: Replicator–Mutator Dynamics

The long-run distribution of communicative strategies in the population is governed by a replicator–mutator dynamic [25]. This framework combines two forces:

- *Replication:* strategies with a better trade-off between communicative success and communicative costs tend to increase in frequency.
- *Mutation:* due to imperfect learning, learners may acquire a strategy different from their teacher's. This is captured by the mutation matrix Q in (30).

Population state. Let $x_{(L,n)}$ denote the proportion of agents using strategy $(L, n) \in \mathscr{S}_{m,k,N}$ in the population. The state of the population at time t is thus the simplex vector

$$x(t) = (x_{(L,n)}(t))_{(L,n)\in\mathscr{S}_{m,k,N}}. \tag{31}$$

Interaction model and success criterion. An interaction pairs a *speaker* of type (L, n) with a *listener* of type (L', n'). A feature configuration $f \in \mathfrak{F}$ and a goal $g \subseteq F$ are drawn independently from priors $P_{\text{env}}(f) = P([\![f]\!]_{\mathcal{F}} \mid (\bigcup_{i=1}^{n} [\![c_i = 1]\!]_{\mathcal{F}})^{c})$ and $P_{\mathcal{G}}(g)$.[7] The speaker samples an utterance $u \in L$ from the policy $\mathcal{S}_n(u \mid$

[7] P_{env} defines the distribution over observable feature configurations in the *actual* environment. We condition on the complement of the union of all category extensions, $\left(\bigcup_{i=1}^{n} [\![c_i = 1]\!]_{\mathcal{F}}\right)^{c}$, because—as discussed in Sect. 2.2—the model assumes that actual objects presented to speakers do not fall under any of the shared conceptual categories. The communicative game thus takes place in a literal space with respect to feature terms, and in a metaphorical space with respect to category terms: speakers must communicate features of objects that resemble known categories, but do not literally belong to them. Accordingly, categorical utterances such as "John is a shark" are assumed to be literally false, yet potentially informative. We shall investigate under which conditions metaphorical category labeling becomes more evolutionarily advantageous than literal feature ascription.

f, g, L); the listener forms a posterior $\mathcal{L}_{n'}(f', g' \mid u)$ over configurations f' and goals g'. *Communicative success* is defined at the level of goal-relevant projections:

$$\delta(g(f),\, g'(f')) \;=\; \begin{cases} 1 & \text{if } g(f) = g'(f'), \\ 0 & \text{otherwise.} \end{cases} \tag{32}$$

Thus a message succeeds if the listener recovers exactly the intended goal and the intended feature cluster under the intended goal.[8]

Expected communicative success and cost. Let F and G be random variables for the feature configuration presented to the speaker and the speaker's communicative goal, respectively. Let U be the random variable for the speaker's utterance, and (F', G') the random variables for the listener's reconstructed feature configuration and goal. The expected communicative success of a speaker of type (L, n) with a listener of type (L', n') is given by:

$$\begin{aligned} \mathrm{Succ}((L,n) \to (L', n')) &:= \mathbb{E}_{f,g}\, \mathbb{E}_{u \sim \mathcal{S}_n(\cdot \mid f,g,L)}\, \mathbb{E}_{(f',g') \sim \mathcal{L}_{n'}(\cdot \mid u)} \Big[\delta(g(f), g'(f')) \Big] \\ &= \sum_{f,g,u,f',g'} P_{\mathrm{env}}(f)\, P_{\mathcal{G}}(g)\, \mathcal{S}_n(u \mid f, g, L)\, \mathcal{L}_{n'}(f', g' \mid u)\, \delta(g(f), g'(f')). \end{aligned} \tag{33}$$

Communication incurs a per-utterance expected communicative cost:

$$\begin{aligned} \mathrm{Cost}((L, n) \to (L', n')) &:= \mathbb{E}_{f,g}\, \mathbb{E}_{u \sim \mathcal{S}_n(\cdot \mid f,g,L)} [\mathsf{C}(u)] \\ &= \sum_{f,g,u} P_{\mathrm{env}}(f)\, P_{\mathcal{G}}(g)\, \mathcal{S}_n(u \mid f, g, L)\, \mathsf{C}(u). \end{aligned} \tag{34}$$

Here, we measure the communicative cost in terms of the structural complexity $\mathsf{C}(u)$ of the utterances, as defined in (3): more complex formulas incur higher costs, reflecting greater cognitive or expressive effort. The expected cost is computed by averaging over the speaker's utterance distribution given each feature-goal pair sampled from the environment.

One-way payoff and sender–receiver symmetry. The *one-way payoff* of an interaction $(L, n) \to (L', n')$ is the payoff that arises when a speaker of type (L, n) communicates with a listener of type (L', n'). Treating communication as a cooperative process, we assume both agents receive the same payoff. The payoff is defined as expected communicative success minus expected communicative cost:

$$\mathrm{Payoff}((L, n) \to (L', n')) \;=\; \mathrm{Succ}((L, n) \to (L', n')) \; - b \cdot \mathrm{Cost}((L, n) \to (L', n')). \tag{35}$$

[8] Note that $\delta(g(f), g'(f')) = 1$ if and only if $g = g'$ and $g(f) = g(f')$, that is, if and only if the listener recovers both the intended goal g and a feature configuration f' that is *goal-equivalent* to the intended one under g, according to (17). Thus, a message is counted as successful only when the listener infers exactly the intended communicative goal and the relevant feature cluster under that goal.

where $b > 0$ controls the trade-off between communicative success and communicative cost. Higher values of b penalize complex utterances more heavily, thereby favoring strategies that communicate effectively with simpler messages.

Because agents alternate between speaker and listener roles, we define the *symmetrized payoff* $\Gamma((L,n),(L',n'))$ as the average of the two one-way payoffs. Concretely, it is the mean of the payoff when (L,n) acts as speaker to (L',n') and the payoff when (L',n') acts as speaker to (L,n):

$$\Gamma((L,n),(L',n')) = \frac{1}{2}\Big[\text{Payoff}((L,n)\to(L',n')) + \text{Payoff}((L',n')\to(L,n))\Big]. \tag{36}$$

This symmetrized payoff is the value assigned to both strategies (L,n) and (L',n') when they interact with each other, reflecting their joint performance across both communicative roles [26,27].

The fitness of strategy (L,n) in population state x is then

$$F(L,n) = \sum_{(L',n')\in\mathscr{S}_{m,k,N}} x_{(L',n')}\cdot\Gamma((L,n),(L',n')). \tag{37}$$

Dynamics. The rate of change of $x_{(L,n)}$ is given by the replicator–mutator equation [25]:

$$\frac{dx_{(L,n)}}{dt} = \sum_{(L',n')\in\mathscr{S}_{m,k,N}} x_{(L',n')}\,Q((L',n'),(L,n))\,F(L',n') - x_{(L,n)}\cdot\bar{F}, \tag{38}$$

where:

- $Q((L',n'),(L,n))$ is the mutation probability defined in (30),
- $F(L',n')$ is the fitness of strategy (L',n') defined in (37),
- $\bar{F}$ is the mean fitness of the population:

$$\bar{F} = \sum_{(L,n)\in\mathscr{S}_{m,k,N}} x_{(L,n)}\cdot F(L,n). \tag{39}$$

Interpretation. The first term on the right-hand side of (38) captures the inflow of learners who adopt strategy (L,n) after being trained by teachers of type (L',n'). The second term subtracts the baseline replication proportional to the average fitness $\bar{F}$, which ensures that population shares remain normalized, i.e. $\sum_{(L,n)\in\mathscr{S}_{m,k,N}} x_{(L,n)}(t) = 1$ for all t. At equilibrium, the population distribution x^* reflects a balance between selective pressures—favoring strategies that optimize the trade-off between communicative success and cost—and mutational noise induced by imperfect learning. We should note that complexity costs shape both dimensions: in learning, prior biases—as shown in (23)—favor simpler strategies, making them more likely to be acquired; in communication, simpler strategies reduce the expected cost of expression, as shown in (35), thereby offering a direct advantage in use.

3 Case Study: Dispersive Communicative Goals

While we leave to future work the task of exploring the full potential of the model just presented, here we restrict our attention to a highly simplified yet revealing test case. Our aim is not to exhaust the model's expressive resources, but to establish a *basic result*: metaphorical strategies can become evolutionarily advantageous when communicative goals are *dispersive*, that is, when agents aim to transmit information about clusters of features at once.

3.1 Case Study Description

Features and categories. The common language of our case study includes three feature terms, each of which can be either true (= 1) or false (= 0):

- $\phi_{dangerous}$, true if the object is dangerous and false if harmless;
- ϕ_{clever}, true if the object is clever and false if stupid;
- ϕ_{strong}, true if the object is strong and false if weak.

Together, these terms define a space of 8 feature configurations $f : \Phi \to \{0, 1\}$, where each configuration assigns a binary value to the three feature dimensions: dangerousness, cleverness, and strength.

In parallel with feature terms, the language also contains a set of eight category terms intended to evoke familiar prototypes:

$$C = \{c_{tiger}, c_{snake}, c_{shark}, c_{goose}, c_{elephant}, c_{owl}, c_{ox}, c_{puppy}\}.$$

Each category is associated with a prototypical feature vector configuration, shown in Table 1. For instance, the prototype for *tiger* is dangerous, clever, and strong, while the prototype for *puppy* is harmless, stupid, and weak. Formally, we define a mapping $\pi : C \to \{0, 1\}^3$, s. t. $\pi(c) = (\pi(c)_{dangerous}, \pi(c)_{clever}, \pi(c)_{strong})$ gives the canonical values of $(\phi_{dangerous}, \phi_{clever}, \phi_{strong})$ for category c.

Shared conceptual framework. Each sample $\omega \in \Omega$ is a total assignment $\omega : C \cup \Phi \to \{0, 1\}$, specifying both the categorical labels and the feature values of a hypothetical kind of object. The sample space Ω divides into three disjoint regions.

First, *multi-categorical samples* are those for which there exist distinct i, j s.t. $\omega(c_i) = \omega(c_j) = 1$. These are excluded by an assumption of *category exclusivity*, hence $P(\omega) = 0$.

Second, *uni-categorical samples* are those where exactly one category holds, i.e. $\exists!\ c \in C$ s.t. $\omega(c) = 1$. For these samples we define

$$c^*(\omega) = \text{the unique category } c \in C \text{ s.t. } \omega(c) = 1. \tag{40}$$

The probability of ω is then determined by the distance between its projected feature configuration $\omega|_\Phi$ and the prototype $\pi(c^*(\omega))$ of its associated category:

$$P(\omega) \propto K(d(\omega|_\Phi, \pi(c^*(\omega)))), \tag{41}$$

Table 1: Category terms and their prototypical feature configurations $(\phi_{dangerous}, \phi_{clever}, \phi_{strong})$.

Category	$\phi_{\text{dangerous}}$	ϕ_{clever}	ϕ_{strong}
c_{tiger}	1	1	1
c_{snake}	1	1	0
c_{shark}	1	0	1
c_{goose}	1	0	0
$c_{elephant}$	0	1	1
c_{owl}	0	1	0
c_{ox}	0	0	1
c_{puppy}	0	0	0

where d is the Hamming distance [16] and K is a decreasing function (e.g. $K(d) = e^{-\beta d}$). Normalization ensures that uni-categorical samples carry total probability mass $1/2$.

Third, *non-categorical samples* are those for which $\forall c \in C,\ \omega(c) = 0$. In this case feature values are drawn from a base distribution P_{N} on $\{0,1\}^{\Phi}$. For simplicity, we take P_{N} to be uniform, so $P_{\mathsf{N}}(f) = 1/8$ for every $f \in \{0,1\}^{\Phi}$. We assign this slice total probability $1/2$. Since there are eight possible non-categorical samples, each has probability $P(\omega) = 1/2 \cdot P_{\mathsf{N}}(\omega|_{\Phi}) = 1/2 \cdot 1/8 = 1/16$.

Observational environment. In the communicative game, speakers are only presented with non-categorical objects. The effective distribution over feature configurations is therefore $P_{\text{env}}(f) = P([\![f]\!]_{\mathcal{F}} \mid \text{non-categorical samples}) = P_{\mathsf{N}}(f) = 1/8$, for every $f \in \{0,1\}^{\Phi}$.

Communicative goal and strategies. In our simplified case study, speakers always aim to transmit the entire feature vector: ergo $\mathcal{G} = \{\Phi\}$ and $P_{\mathcal{G}}(\Phi) = 1$. Furthermore, we compare only two strategies:

- *Literal strategy* $(L_{\text{lit}}, n_{\text{lit}})$: its lexicon contains one conjunction for each of the 8 feature triples, $L_{\text{lit}} = \{(\phi_d{=}i) \wedge (\phi_c{=}j) \wedge (\phi_s{=}k) \mid i,j,k \in \{0,1\}\}$, with reasoning depth $n_{\text{lit}} = 0$.
- *Metaphorical strategy* $(L_{\text{met}}, n_{\text{met}})$: its lexicon contains the 8 category atoms $L_{\text{met}} = \{c{=}1 \mid c \in C\}$, with $n_{\text{met}} = 1$.

Replicator Dynamics. In this restricted comparison, the two lexica are disjoint $(L_{\text{lit}} \cap L_{\text{met}} = \emptyset)$. Consequently, learning cannot map one lexicon into the other, so the mutation matrix Q defined in (30) is the identity matrix I, meaning $Q((L',n'),(L,n)) = \delta((L',n'),(L,n))$. The replicator–mutator dynamics thus collapses to the pure *replicator equation* [17,18,25]:

$$\frac{dx_{\text{met}}}{dt} = x_{\text{met}}(F_{\text{met}} - \bar{F}), \tag{42}$$

where x_{met} denotes the proportion of agents using the metaphorical strategy $(L_{\text{met}}, n_{\text{met}})$ in the population, and F_{met} its associated fitness.

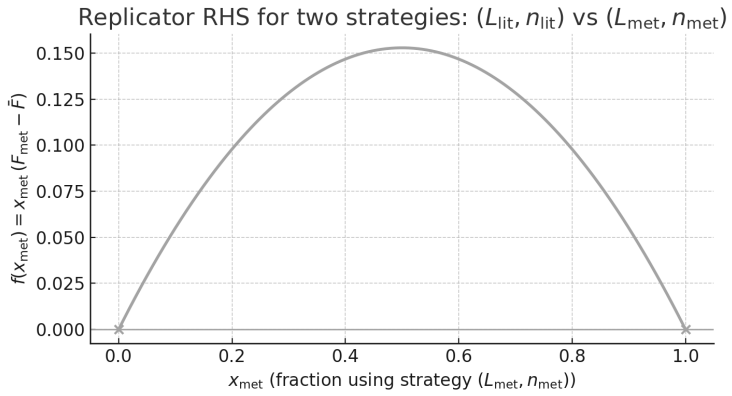

Fig. 1: Graph of the function $f(x_{\text{met}}) = x_{\text{met}}(F_{\text{met}} - \bar{F})$ in the RHS of the replicator equation (42) for the test case, generated using Python.

3.2 Results

With parameter values: $\beta = 1$, $\lambda = 1$, $b = 0.5$, the resulting symmetrized payoff matrix Γ is:

$$\Gamma = \begin{pmatrix} -1.500 & -0.888 \\ -0.888 & -0.277 \end{pmatrix},$$

where rows/columns are ordered as $((L_{\text{lit}}, n_{\text{lit}}), (L_{\text{met}}, n_{\text{met}}))$.

Figure 1 shows the right-hand side function of the replicator equation:

$$f(x_{\text{met}}) = x_{\text{met}}(F_{\text{met}} - \bar{F}). \tag{43}$$

The function is positive throughout the interior $(0, 1)$, with $f(0) = f(1) = 0$.

Equilibrium and stability. Since $\Gamma_{met,met} = -0.277 > \Gamma_{lit,met} = -0.888$, the metaphorical strategy $(L_{\text{met}}, n_{\text{met}})$ is a *strict Nash equilibrium* (NE). Every strict NE is also an *evolutionarily stable strategy* (ESS), hence $(L_{\text{met}}, n_{\text{met}})$ is an ESS [24,25]. Moreover, because $f(x_{\text{met}}) = x_{\text{met}}(F_{\text{met}} - \bar{F})$ is strictly positive for all $x_{\text{met}} \in (0, 1)$, the replicator dynamics guarantees global convergence: $x_{\text{met}}(t) \to 1$ as $t \to +\infty$. Thus, under this highly simplified scenario with one (dispersive) goal and two strategies, the metaphorical strategy robustly dominates the literal strategy.

3.3 Discussion

The analysis of dispersive goals under our simplified setup shows a clear result: metaphorical strategies *can* strictly dominate literal ones and converge to fixation under replicator dynamics. This provides a proof-of-concept that the use of metaphor can be evolutionarily advantageous when communication requires conveying entire clusters of features at once. At the same time, these results are obtained under highly idealized assumptions. We considered only:

- a single communicative goal (the full feature vector);
- two strategies with disjoint lexica, making the mutation matrix trivial (identity matrix);
- uniform priors over non-categorical objects.

In this setting, imperfect learning does not occur, and no partially overlapping strategies were admitted.

A natural extension is to allow for richer sets of communicative goals, including both *dispersive* goals (full or multi-feature transmission) and *specific* goals (communication about a single or a few features). This would enable direct competition between strategies specialized for different communicative demands. It would also be important to introduce a wider range of strategies: not only purely literal or purely metaphorical ones, but also strategies with partly literal (feature-based) and partly metaphorical (category-based) lexica, and especially explore strategy spaces with partially overlapping lexica, where imperfect Bayesian learning (i.e., mutation) becomes central to the dynamics.

Finally, the parameter space itself remains unexplored. The rationality parameter λ, which regulates the sharpness of pragmatic reasoning, and the evolutionary trade-off parameter b between communicative success and cost are especially promising levers. Varying these parameters may shift the balance between literal and metaphorical strategies, alter the stability of equilibria, or generate coexistence dynamics.

The present paper thus merely sets the stage, by illustrating the framework in its general form and showcasing its potential with a single, tractable yet highly revealing case study. A more systematic exploration will be the subject of future work.

4 Conclusion

We have presented a formal model of the cultural evolution of metaphorical language by embedding RSA [6,7] reasoning into a replicator–mutator dynamic [25] with Bayesian learning [14,15], following the approach of Brochhagen *et al.* [2,3] on scalar implicatures. The model represents communicative strategies as lexica paired with pragmatic depth, and evaluates them under selective pressures balancing communicative success and cognitive cost.

In a simplified test case with a single, dispersive, communicative goal, we found that metaphorical strategies form a strict Nash equilibrium, driving the entire population toward metaphorical communication. This initial and "small" exploration begins to provide a principled explanation of why figurative language may be adaptively favored in certain communicative environments.

Although highly idealized, this case study highlights the power of the presented modeling framework and points the way toward richer analyses involving multiple goals, increased strategy variation, imperfect learning, and systematic parameter sweeps. The broader aim is to integrate formal pragmatics with EGT, opening a novel research program on the adaptive dynamics of metaphorical communication.

References

[1] Bornstein, M.H., Hendricks, C.: Basic language comprehension and production in >100,000 young children from sixteen developing nations. J. Child Lang. **39**(4), 899–918 (2011), doi: 10.1017/S0305000911000407

[2] Brochhagen, T., Franke, M., van Rooij, R.: Learning biases may prevent lexicalization of pragmatic inferences: A case study combining iterated (bayesian) learning and functional selection. In: Proc. 38th Annu. Conf. Cogn. Sci. Soc. pp. 2081–2086. Cogn. Sci. Soc., Austin, TX (2016)

[3] Brochhagen, T., Franke, M., van Rooij, R.: Coevolution of lexical meaning and pragmatic use. Cogn. Sci. **42**(8), 2757–2789 (2018), doi: 10.1111/cogs.12681

[4] Clark, E.V., Hecht, B.F.: Comprehension, production, and language acquisition. Annu. Rev. Psychol. **34**(1), 325–349 (1983), doi: 10.1146/annurev.ps.34.020183.001545

[5] Clark, H.H.: Using language. Cambridge Univ. Press, Cambridge (1996), doi: 10.1017/CBO9780511620539

[6] Degen, J.: The rational speech act framework. Annu. Rev. Linguist. **9**(1), 519–540 (2023), doi: 10.1146/annurev-linguistics-031220-010811

[7] Frank, M.C., Goodman, N.D.: Predicting pragmatic reasoning in language games. Science **336**(6084), 998 (2012), doi: 10.1126/science.1218633

[8] Frank, M.C., Goodman, N.D.: Supplementary materials for predicting pragmatic reasoning in language games. Science Online Suppl. Materials (2012), doi: 10.1126/science.1218633

[9] Gentner, D., Bowdle, B.: Metaphor as structure-mapping. In: Gibbs, R.W.J. (ed.) The Cambridge handbook of metaphor and thought, pp. 109–128. Cambridge Univ. Press, Cambridge (2008), doi: 10.1017/CBO9780511816802.008

[10] Geurts, B., Moore, R. (eds.): Evolutionary pragmatics: Communicative interaction and the origins of language. Oxford Univ. Press, Oxford (2025), doi: 10.1093/9780191967566.001.0001

[11] Goodman, N.D., Frank, M.C.: Pragmatic language interpretation as probabilistic inference. Trends Cogn. Sci. **20**(11), 818–829 (2016), doi: 10.1016/j.tics.2016.08.005

[12] Grice, H.P.: Logic and conversation. In: Cole, P., Morgan, J.L. (eds.) Syntax and semantics, vol. 3: Speech acts, pp. 41–58. Academic Press, New York (1975), doi: 10.1163/9789004368811˙003

[13] Grice, H.P.: Studies in the way of words. Harvard Univ. Press, Cambridge, MA (1989)

[14] Griffiths, T.L., Kalish, M.L.: A bayesian view of language evolution by iterated learning. In: Proc. 27th Annu. Conf. Cogn. Sci. Soc. pp. 827–832. Lawrence Erlbaum, Mahwah, NJ (2005)

[15] Griffiths, T.L., Kalish, M.L.: Language evolution by iterated learning with bayesian agents. Cogn. Sci. **31**(3), 441–480 (2007), doi: 10.1080/15326900701326576
[16] Hamming, R.W.: Error detecting and error correcting codes. Bell Syst. Tech. J. **29**(2), 147–160 (1950), doi: 10.1002/j.1538-7305.1950.tb00463.x
[17] Hofbauer, J., Sigmund, K.: Evolutionary games and population dynamics. Cambridge Univ. Press, Cambridge (1998), doi: 10.1017/CBO9781139173179
[18] Hofbauer, J., Sigmund, K.: Evolutionary game dynamics. Bull. Am. Math. Soc. **40**(4), 479–519 (2003), doi: 10.1090/S0273-0979-03-00988-1
[19] Huttenlocher, J.: The origins of language comprehension. In: Solso, R.L. (ed.) Theories in cognitive psychology: The Loyola symposium, pp. 331–368. Lawrence Erlbaum, Hillsdale, NJ (1974), doi: 10.4324/9781032722375
[20] Kao, J., Bergen, L., Goodman, N.D.: Formalizing the pragmatics of metaphor understanding. In: Proc. 36th Annu. Conf. Cogn. Sci. Soc. pp. 719–724. Cogn. Sci. Soc., Austin, TX (2014)
[21] Kao, J., Goodman, N.D.: Let's talk (ironically) about the weather: Modeling verbal irony. In: Proc. 37th Annu. Conf. Cogn. Sci. Soc. pp. 1051–1056. Cogn. Sci. Soc., Austin, TX (2015)
[22] Krieger, M.S.: Evolutionary dynamics of hyperbolic language. PLoS Comput. Biol. **19**(2), e1010872 (2023), doi: 10.1371/journal.pcbi.1010872
[23] Lakoff, G., Johnson, M.: Metaphors we live by. Univ. Chicago Press, Chicago (1980), doi: 10.7208/chicago/9780226470993.001.0001
[24] Maynard Smith, J.: Evolution and the theory of games. Cambridge Univ. Press, Cambridge (1982), doi: 10.1017/CBO9780511806292
[25] Nowak, M.A.: Evolutionary dynamics: Exploring the equations of life. Harvard Univ. Press, Cambridge, MA (2006), doi: 10.2307/j.ctvjghw98
[26] Nowak, M.A., Komarova, N.L., Niyogi, P.: Evolution of universal grammar. Science **291**(5501), 114–118 (2001), doi: 10.1126/science.291.5501.114
[27] Nowak, M.A., Krakauer, D.C.: The evolution of language. Proc. Natl. Acad. Sci. USA **96**(14), 8028–8033 (1999), doi: 10.1073/pnas.96.14.8028
[28] Ortony, A. (ed.): Metaphor and thought. Cambridge Univ. Press, Cambridge, 2 edn. (1993), doi: 10.1017/CBO9781139173865
[29] Roberts, C.: Information structure in discourse: Towards an integrated formal theory of pragmatics. Semant. Pragmat. **5**(6), 1–69 (2012), doi: 10.3765/sp.5.6
[30] Scontras, G., Tessler, M.H., Franke, M.: Probabilistic language understanding: An introduction to the rsa framework. `https://www.problang.org/` (2018), accessed 2025-09-13
[31] Scontras, G., Tessler, M.H., Franke, M.: A practical introduction to the rational speech act modeling framework (2021), doi: 10.48550/arXiv.2105.09867
[32] Skyrms, B.: Signals: Evolution, learning, and information. Oxford Univ. Press, Oxford (2010), doi: 10.1093/acprof:oso/9780199580828.001.0001
[33] Stalnaker, R.: Assertion. In: Cole, P. (ed.) Syntax and semantics, vol. 9: Pragmatics, pp. 315–332. Academic Press, New York (1978), doi: 10.1163/9789004368873˙013

[34] Stalnaker, R.: Common ground. Linguist. Philos. **25**(5), 701–721 (2002), doi: 10.1023/A:1020867916902
[35] Zaslavsky, N., Hu, J., Levy, R.P.: A rate–distortion view of human pragmatic reasoning (2020), doi: 10.48550/arXiv.2005.06641