

# Evaluating Linguistic Creativity: Pragmatic Reasoning with a Distributional Semantics

Reuben Cohn-Gordon<sup>a</sup> and Leon Bergen<sup>b</sup>

<sup>a</sup>Stanford University; <sup>b</sup>University of California San Diego

This manuscript was compiled on May 15, 2019

Humans create and interpret novel metaphors, like *Time is a thief* or *My lawyer is a shark*, with relative ease, incorporating contextual knowledge to determine which aspects of the predicate (*thief*, *shark*) are true of the subject (*time*, *my lawyer*). An intuitive theory of the process underlying metaphor interpretation is that a listener reasons about an informative speaker (who in turn reasons about the listener) to update their beliefs about the subject and the relevant dimensional of meaning.

We present a Bayesian model of this reasoning process in which empirically learned word embeddings are used to provide an underlying representation of word meaning. This allows us to interpret open domain predicative and adjectival metaphors without manually stipulating the meanings of the words they contain. We find a significant preference in human judgments for our model over a system which uses word embeddings without a explicit representation of inter-agent reasoning, suggesting that this reasoning about an informative and relevant speaker is key to understanding non-literal language.

metaphor | informativity | Bayesian pragmatics | distributional semantics

Metaphor presents a compelling theoretical challenge for the understanding of meaning in natural language. On hearing (1) in a context where the subject, Jane, is known to be a consultant, a listener might infer that Jane is not literally a soldier, but rather that she shares certain attributes with soldiers (perhaps determination, endurance, or ruthlessness).

(1) Jane is a soldier

The *pragmatic* view of metaphor, proposed by Grice (1), takes the meaning conveyed by sentence (1) in a given context to be the result of a process of reasoning by a listener, about a speaker who is trying to communicate truthfully, informatively, and relevantly. That is, the listener attempts to jointly deduce what Jane must be like and what aspect of Jane is plausibly relevant, such that the speaker who wants to successfully communicate some aspect of Jane would have chosen the predicate *soldier* over other alternatives.

**Interpreting figurative language** Previous work in linguistics and cognitive science has modeled language interpretation in context as the product of back and forth reasoning between Bayesian agents (2). In models of this variety, language production corresponds to a speaker, whose goal is to communicate a state  $w$  by choosing an utterance  $u$ , and language interpretation to a listener, whose goal is to infer  $w$  having heard  $u$ . Agents of this sort are then nested, so that, for example, the listener can reason about its own model of the speaker who in turn reasons about its model of the listener.

$L_1^Q$ , a model of metaphor interpretation in this framework (3) relies on the mechanism of *projection functions* which

dictate the dimension of the world that the speaker cares about communicating.  $L_1^Q$  jointly determines the state of the world (e.g. what Jane is like) and a projection, corresponding to the aspect of the world the speaker cares to communicate (e.g. Jane's determination). As before, this listener assumes an informative speaker: one whose choice of utterance maximizes the probability of communicating the world  $w$  to the speaker's model of the listener, but now only up to a projection  $q$  which dictates the relevant dimension.

$L_1^Q$  provides an account of predicative metaphors (those of the form *A is a B*) and adjective-noun (AN) metaphors (like *fiery temper*). However, in order to generate predictions from the model, it is necessary to provide a semantics, specifying which utterances  $u$  and states  $w$  are compatible. Hand-constructing a semantics is possible, but restricts the model's scalability, and its evaluability on previously unseen metaphors.

**Our contribution** By adapting  $L_1^Q$ , a Bayesian model of non-literal meaning, to incorporate a word embedding semantics where words correspond to points in an abstract vector space (4, 5), we obtain a system capable of interpreting open domain predicative and adjectival metaphors without the need for hand-specified word meanings. This adaptation requires a generalization of projection functions to linear projections in a vector space, and a novel inference algorithm to calculate the now continuous posterior distribution of  $L_1^Q$ . Constructing this system permits what is to our knowledge the first open domain evaluation of a Bayesian model of pragmatic reasoning. We show that our model of metaphor interpretation significantly

## Significance Statement

Linguistic creativity — the ability to combine existing representations to create new meanings — is a distinctive trait of human cognition. Metaphor provides a general vehicle for this creativity: it allows people to communicate about one domain, using concepts from another. Here, we develop a system for open-domain interpretation of metaphor. Our system integrates world knowledge automatically induced from large text corpora, with reasoning about the social goals of the speaker. The approach provides a general architecture for composing domain-specific knowledge with social reasoning, providing insight into the origins of linguistic creativity.

Please provide details of author contributions here.

Please declare any conflict of interest here.

<sup>1</sup> A.O. (Author One) and A.T. (Author Two) contributed equally to this work (remove if not applicable).

<sup>2</sup> To whom correspondence should be addressed. E-mail: reubengc@stanford.edu

outperforms a baseline which uses a word embedding semantics without explicit pragmatic reasoning. This suggests that the information in word embeddings alone is not sufficient to capture the creativity of metaphorical language, but that an explicit model of pragmatic reasoning is also key.

## 1. Overview of metaphor

Metaphor exists in many syntactic forms (6). For present purposes, we focus on copular predicates (e.g. *Jane is a soldier*) and AN noun phrases (e.g. *fiery temper*). We refer to the predicated or modified noun (*Jane*, *temper*) as the *target* of the metaphor and the predicate or adjective (*soldier*, *fiery*) as the *source* (see (7) for the more general sense of these terms).

For a given metaphor, only certain properties of the target are described by the source, and which these are depend on the metaphor and the context. For instance, (2), said of a sleeping dog, could convey that it is unresponsive, but said of a large alert dog, could convey that it is heavy.

(2) The dog is a rock.

While certain metaphors are conventional - comparing someone to a lion tends to connote bravery - examples like (2) suggest that the interpretation of a metaphor is contextually dependent. The benefit of the pragmatic view of metaphor is the ability to explain this dependence on context, in a way which takes into account an underlying semantics (e.g. the conventional meanings of *dog* and *rock*).

## 2. A Bayesian model of metaphor interpretation

The Rational Speech Acts framework (RSA) provides an elegant and practical way of formalizing pragmatic reasoning as nested Bayesian inference (2).

In this framework, listeners and speakers are represented as conditional probability distributions. Speakers are distributions  $P(U|W)$  over possible utterances given worlds, and listeners distributions  $P(W|U)$  over possible worlds given utterances, where  $W$  is the set of possible states, and  $U$  is the set of utterances available to a speaker. The most basic version of RSA (2) is incapable of interpreting metaphors, due to the strict assumption that the speaker's utterances are literally true. To address this, Kao et al. (3) propose a model  $L_1^Q$ , shown in (5), which in turn is defined in terms of  $S_1$  (4) and  $L_0$  (3).

$$(3) \quad L_0(u|w) \propto \llbracket u \rrbracket(w) \cdot P_L(w)$$

$$(4) \quad S_1(u|w, q) \propto \sum_{w'} \delta_{q(w)=q(w')} \cdot L_0(w'|u)$$

$$(5) \quad L_1^Q(w, q|u) \propto S_1(u|q, w) \cdot P_L(w) \cdot P_{L_Q}(q)$$

**The literal listener**  $L_0$  represents a model of a listener that, given an utterance  $u \in U$  updates their belief about the world  $w \in W$  in accordance with the semantics  $\llbracket \cdot \rrbracket$  and their prior  $P_L$ . This semantics, usually a function  $U \rightarrow (W \rightarrow \{0, 1\})$ , represents the conventional association between states  $w$  and utterances  $u$  which the speaker and listener take as given.

**Projections** Functions  $q \in Q$  formalize the notion of picking a particular *aspect* or *dimension* of  $w$ . Formally, they are surjective functions out of  $W$ .

**The informative speaker**  $S_1$  has a state  $w$  they want to communicate, and reasons about  $L_0$ , preferring utterances  $u$  which maximize the  $L_0$  posterior probability on  $w$ , up to the aspect of  $w$  specified by  $q$ .  $\delta_{a=b}$  is the delta function, and is equal to 1 if and only if  $a = b$ . If  $q$  is the identity function,  $S_1(u|w) \propto L_0(w|u)$ , and is thus a model of a speaker who prefers to choose the most informative utterance available.

**The pragmatic listener** The full model,  $L_1^Q$ , jointly infers values for  $w$  and  $q$  by reasoning about  $S_1$ . The key dynamic is that the listener may hear an utterance  $u$  and infer a pair  $(w, q)$  where  $u$  is semantically incompatible with  $w$  (i.e.  $\llbracket u \rrbracket(w) = 0$ ) but where  $u$  conveys some aspect of  $w$  as determined by  $q$ .

$L_1^Q$  functions as a model of metaphor interpretation. For instance, using the metaphor in (6), the goal of the speaker is to communicate a state  $w$  (representing what John is like) along  $q$  (representing some *feature* of this state).

As an example in a totally hand-constructed setting, we could take John to be fully characterized by two features, whether he is vicious, and whether he is aquatic, so that a state  $w$  is a value (true or false) for both of these predicates. The projections  $q \in Q$  are then the functions mapping a state to its value on viciousness ( $q_{vicious}$ ) or aquaticness ( $q_{aquatic}$ ) respectively. Further, we assume that *shark* is compatible only with the state in which John is both vicious and aquatic.

(6) John is a shark.

The key property of  $L_1^Q$  is that, on hearing (6), the prior belief that John is not literally an aquatic animal leads  $L_1^Q$  to conclude that the speaker cares about conveying the viciousness dimension (i.e. has projection  $q_{vicious}$ ), and that John is vicious. See (3) for quantitative examples.

Importantly,  $L_1^Q$  can do more than simply using prior knowledge to interpret literally false statements in a flexible way. It is also capable of reasoning about alternative utterances: for instance, suppose we add a third property, *quickness*, so that *shark* is compatible only with the state in which John is quick, aquatic and vicious, and also add a third utterance, *goldfish*, compatible only with John being quick, aquatic and *not* vicious.

In this second example, when  $L_1^Q$  hears *shark*, it infers that John is more likely vicious than quick. This is because a speaker who wanted to communicate that John is vicious would only be able to use the utterance *shark*, whereas a speaker who wanted to communicate that John is quick would be able to choose between either *shark* or *goldfish*. The utterance *shark* is therefore more likely to have been produced by the speaker trying to communicate John's viciousness.

$L_1^Q$  can model AN metaphors in a similar way. For a phrase like *fiery temper*, we say that the goal of a listener is to decide what is true of the temper in question given that the speaker has modified it with *fiery*.

## 3. Distributional Semantics

From a linguistic corpus, it is possible to obtain a mapping from words to points in a high-dimensional vector space that has the property that semantic similarity of a pair of words  $a$  and  $b$  corresponds to a metric, such as cosine distance, between the vectors  $\vec{a}$  and  $\vec{b}$ .

Mappings of this sort, commonly referred to as *word embeddings* or *distributional models of word meaning*, can be obtained

either by dimensionality reduction of a co-occurrence matrix (5), or by extracting the weights of a statistical model (4, 8, 9) trained on a separate task. In either case, word embeddings provide a way to empirically obtain fine grained connotations of lexical items (4), and have been used effectively in a number of NLP tasks (10, 11). They have also been used to compute vectorial representations of phrases and sentences (12, 13).

Metaphor is an obvious candidate for approaches that use distributional semantics: a wide variety of attempts have been made to leverage the information inherent in pre-trained word vectors for the detection, interpretation and paraphrase of metaphor (see (14) for an overview of proposed systems.).

Our hypothesis is that, while the information in high quality word embeddings captures important aspects of meaning, a cognitively realistic model of metaphor interpretation should also incorporate pragmatic reasoning, of the sort formalized in the RSA framework. We now explain how  $L_1^Q$  can be combined with a distributional model of word meaning.

#### 4. Bayesian pragmatics with a distributional semantics

We now introduce a *vectorial* interpretation of  $L_1^Q$ . Importantly, this requires no modification to equations (3-5). The crucial difference is that our state space  $W$  is now not just a set, but a vector space determined by a word embedding  $E : U \rightarrow W$ , so that elements  $\vec{w} \in W$  are vectors. For our application of the model, we assume  $U$  is a set of adjectives.

**The listener's prior** In a setting with a finite set  $W$ , a discrete prior  $P_L$  over  $W$  sufficed. In the present case, where  $W$  is necessarily infinite (ranging over real-valued vectors), we use a multivariate spherical Gaussian distribution, which can be parametrized by a vector  $\vec{\mu}$  for the mean and a single scalar  $\sigma$  (the value of every diagonal entry of the covariance matrix). We define the prior over projections  $P_{L_Q}$  to be uniform.

$$(7) \quad P_L(w) = P_{\mathcal{N}}(w | \mu = E(\text{target}), \sigma = \sigma_1)$$

We can view  $P_L$  as representing uncertainty over the position of the entity or concept that the target noun (e.g. *man* in “The man is a shark”) represents. The goal of the speaker is to convey a position in the space to the listener, and the goal of the listener is to infer what this position is. In this sense, a spatial reference game is being played (15), in an abstract word embedding space. Our vectorial semantics bears comparison to the *conceptual space* semantics of (16), as well as the proposal for metaphor comprehension of (17).

The multidimensional Gaussian distribution weights most heavily those points nearest to its mean. By setting the mean of the prior as  $E(\text{target})$ , we encode the listener's assumption that the meaning the speaker wishes to communicate is in the neighborhood of the source noun.  $\sigma_1$  is a hyperparameter of the model.

**The semantics** A word embedding space has no explicit representation of truth. That is to say, while we can compare the similarity of a noun and an adjective according to a variety of metrics, we do not have a means of categorically determining the compatibility of that adjective and noun. As far as our model is concerned, this is not a problem since the definition of  $L_0$  in (3) requires only that the semantics  $\llbracket \cdot \rrbracket$  be a function  $U \rightarrow (W \rightarrow \mathbb{R})$ . We can define such a function as follows, with  $\sigma_2$  as a hyperparameter:

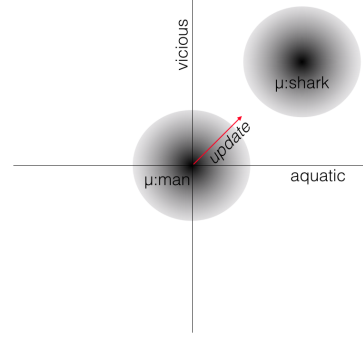


Fig. 1. 2D depiction of vectorial  $L_0$ , for  $\vec{man} = (0,0)$  and  $\vec{shark} = (1,1)$ .

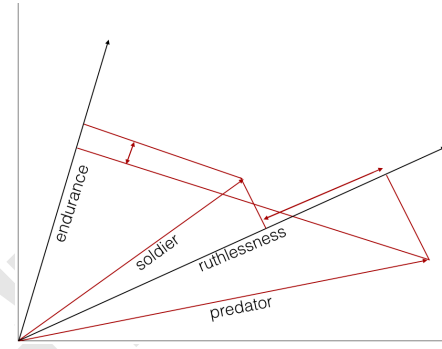


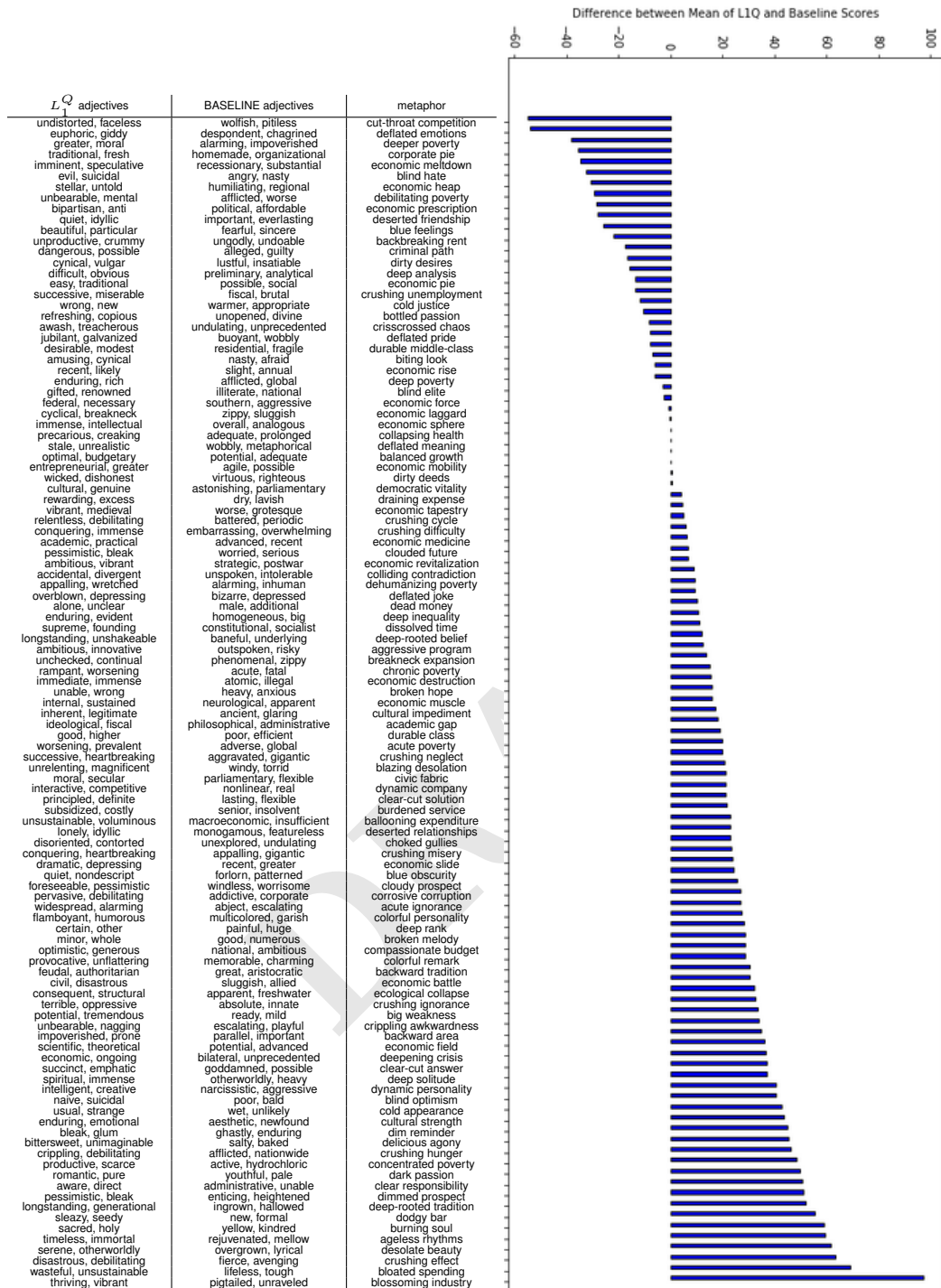
Fig. 2. In this hand-constructed 2D example, vectors for *soldier* and *predator* are mapped onto subspaces given by *endurance* and *ruthlessness*.

$$(8) \quad \llbracket u \rrbracket(w) = P_{\mathcal{N}}(w | \mu = E(\text{source}), \sigma = \sigma_2)$$

The result of this definition is that the value of  $\llbracket u \rrbracket(w)$  is a real number which decreases with the Euclidean distance between  $u$  and  $w$ . The advantage of defining a semantics in this way is that both the prior of  $L_0$ , shown in (7) and the likelihood, namely the semantics shown in (8), have the form of Gaussian distributions, which allows for a closed form solution of  $L_0$ .

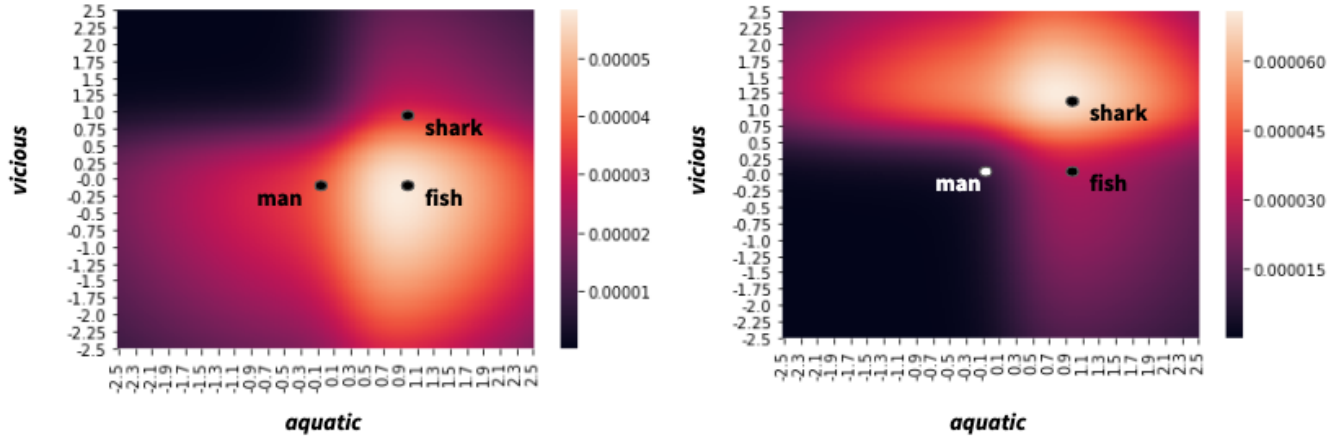
**Projections** Finally, we need to supply an notion of a projection function  $q$  that is defined on our vector space, and to specify a set  $Q$  of such projections. For this, we use linear projections along a vector (or hyperplane)  $\vec{v}$  capturing the degree to which each  $\vec{w}$  extends along  $\vec{v}$ , ignoring orthogonal dimensions. Geometrically, it can be thought of as dropping a line from an input vector  $\vec{w}$  at a right angle onto  $\vec{v}$ , as depicted in figure 2. These linear projections exploit the linear structure of the word embedding space (5), although see (18, 19) for a discussion of the degree of this linearity.

In practice, we restrict ourselves to projections along a vector, rather than a larger subspace. To obtain a set  $Q$  of projections, we first note that since the denotations of words are vectors in  $W$ , any word parametrizes a linear projection  $q$ . For instance, we can think of the word *vicious* as parametrizing a *viciousness* projection, which measures how far the denotations of all other points in the space fall along *vicious*. We choose  $Q$  as a set of gradable adjectives, so that the projection of a noun  $n$  onto  $\vec{v}$  amounts to asking: to what extent is  $n$  *v*? Figure 4 provides a visualization of the  $L_1^Q$  posterior in a simple 2D case corresponding to the example discussed in section 2.



**Fig. 3.** The 109 metaphors used in the experiment, and baseline and  $L_1^Q$  interpretations. Bar positions indicate difference between judgments of  $L_1^Q$  and baseline proposals, averaged across participants and across both proposals of each model. Bars right of center indicate a preference for the pragmatic model, showing that for roughly 75% of the metaphors, the  $L_1^Q$  interpretation is preferred.





**Fig. 4.** Heatmaps visualizing the inferred  $L_1^Q$  marginal posterior over states in a simple 2D case given *fish* (left) and *shark* (right), with  $U = \{man, shark, fish\}$ , hand-chosen denotations overlaid, and  $\sigma_1 = 5.0$ ,  $\sigma_2 = 0.5$ .

**A. Interpreting results from  $L_1^Q$ .** We now have an algorithm for approximating the joint posterior distribution of  $L_1^Q$  over  $W$  and  $Q$  after hearing a metaphor  $u$ . Unlike points  $w \in W$ , projections  $q \in Q$  are readily interpretable, since they correspond to adjectives, describing the aspect of the metaphorical adjective or predicate that is inferred to be relevant. For this reason, we use the marginal posterior over  $Q$  to generate predictions from the model. For instance, the top two  $L_1^Q$  marginal posterior projections  $q$  for each metaphor, which we use in our experiment, are shown in figure 3.

## 5. Experimental Evaluation

In order to test our model on human judgments, we design an experiment which compares the  $L_1^Q$  interpretations of metaphors to a baseline model which uses word embeddings but no pragmatic reasoning.

**Experimental Design.** In our experiment, each participant is shown a series of 12 metaphors, selected randomly from a total 109. For each metaphor, they are asked to rate on a slider four adjectives representing interpretations of the metaphor, which are the best and second best adjective generated by  $L_1^Q$ , and from a baseline model which selects adjectives by their distance in the word embedding space to the mean of the target and source (see *Materials and Methods*). An example is shown in figure 5.

**Analysis.** The results, shown in figure 3, were analyzed using a mixed-effects model with random slopes and intercepts for items and participants. The target interpretations were rated significantly higher than the baseline interpretations ( $\beta=13.8$ ,  $t=5.3$ ,  $p<0.001$ ).

## 6. Discussion

We have shown that it is possible to scale Bayesian pragmatic reasoning to a distributional semantics and by so doing to obtain a model of metaphor interpretation. Our evaluation, the first such open domain test of a Bayesian model of pragmatic language interpretation, indicates that the principles of pragmatic reasoning continue to operate at this scale, and are key to obtaining human-like interpretations of metaphors. We

see this as an important step towards a cognitively accurate and computationally tractable model of pragmatic language interpretation and production in general.

## Materials and Methods

**Inference in vectorial setting.** Because  $P_L(w)$  is a continuous distribution in the vectorial interpretation of  $L_1^Q$ , inference by enumeration is not possible, and either analytic or approximate methods are required. We employ a mix of the two; the  $L_0$  and  $S_1$  posteriors can be calculated analytically, while  $L_1^Q$  requires us to develop an approximate inference algorithm. We describe this algorithm in parts, working up from the  $L_0$ . Implementations, written in Tensorflow will be made publicly available.

**$L_0$  Inference** Intuitively, the vectorial interpretation of  $L_0$  amounts to the process shown in figure 1, where a ball, corresponding to the prior, is moved in the direction of the point corresponding to the received utterance. To calculate  $L_0$  analytically, we make use of Gaussian conjugacy. When the prior  $P_L$  is defined as in Equation 7, and the semantic interpretation is defined as in Equation 8, then conjugacy implies that the listener posterior is given by:

$$(9) \quad L_0(w|u) = P_{\mathcal{N}}(w|\mu = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2} (\frac{E(target)}{\sigma_1^2} + \frac{E(source)}{\sigma_2^2}), \sigma = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2})$$

**$S_1$  Inference** The speaker is defined by Equation 4, which in the continuous case can be rewritten as:

$$(10) \quad S_1(u|w, q) \propto \int_{w'} \delta_{q(w)=q(w')} \cdot L_0(w'|u)$$

This integral is computing the marginal probability of  $w_q$ , the projection of state  $w$  onto QUD vector  $q$ . From Equation 9,  $L_0(\cdot|u)$  is a normally distributed random variable, and therefore projection of this random variable onto a linear subspace is also normally distributed, providing a closed-form solution to  $S_1$ .

**$L_1^Q$  Inference** The  $L_1$  posterior is a joint distribution over one continuous and one discrete random variable. Because of the linear structure of the problem, we are able to devise a near-exact inference algorithm for the marginal distribution over QUDs in  $Q$ , derived as follows:

$$\begin{aligned}
L_1(q|u) &= \int_{\mathbb{R}^n} L_1(w, q|u) dw \\
&= \frac{1}{K} P_{L_Q}(q) \int_{\mathbb{R}^n} P_L(w) S_1(u|w, q) dw \\
&= \frac{1}{K} P_{L_Q}(q) \int_{\mathbb{R}^n} P_L(w_q, w^\perp) S_1(u|w_q, q) dw \\
&= \frac{1}{K} P_{L_Q}(q) \int_{\mathbb{R}^n} P_L(w_q) P_L(w^\perp) S_1(u|w_q, q) dw \\
&= \frac{1}{K} P_{L_Q}(q) \int_{Q^\perp} P_L(w^\perp) dw^\perp \int_Q P_L(w_q) S_1(u|w_q, q) dw_q \\
&= \frac{1}{K} P_{L_Q}(q) \int_Q P_L(w_q) S_1(u|w_q, q) dw_q
\end{aligned}$$

Here  $K$  is a normalizing constant,  $w, q \in \mathbb{R}^n$ , and  $w_q$  is the projection of  $w$  onto the vector  $q$ .  $Q$  is the subspace of  $\mathbb{R}^n$  spanned by the vector  $q$ , and  $Q^\perp$  is the orthogonal complement of  $Q$ . The vector  $w^\perp$  is the projection of vector  $w$  onto the subspace  $Q^\perp$ . The final equation is a one-dimensional integral, and can be computed using a discrete approximation. We use a Gaussian approximation, which easily generalizes to the setting of multi-dimensional QUDs. The constant  $K$  can be found from the constraint  $\sum_q L_1(q|u) = 1$ .

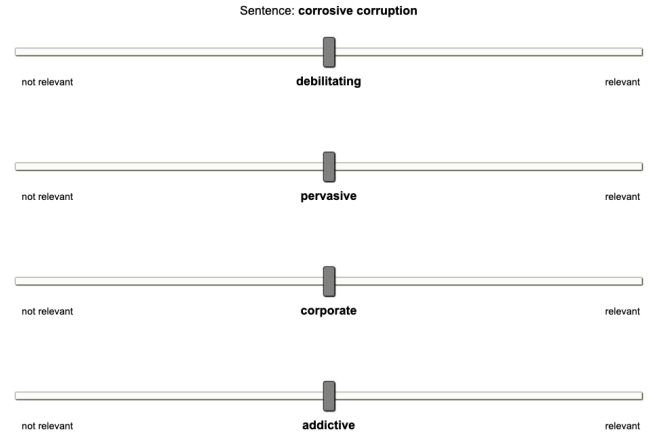
**Experiment.** The aim of our experiment is to determine whether pragmatic reasoning results in better interpretations of metaphors, according to human judgments. As such, a natural baseline model to compare against is a model with a distributional semantics that does not make use of the pragmatic reasoning process inherent in  $L_1^Q$ .

**Baseline model** Our baseline model is defined as follows: for a given metaphor of the form  $(a\ n)$ , we take the mean of the adjective  $a$  and noun  $n$ . The two nearest (measured by cosine distance) adjectives  $q$  to this mean are our baseline interpretations for the metaphor. We use the mean (a weighted sum) in light of the effectiveness of vector addition in deriving representations of phrasal and sentence meanings from constituent words (12, 20, 21). Cosine distance is a standard metric of similarity used for word embeddings (5).

**$L_1^Q$  hyperparameters** We use the largest available (300 dimensional) GloVe vectors, as our word embedding  $E$ . For each AN metaphor  $(a\ n)$ , we specify  $U$  as a set of 101 alternative utterances, consisting of  $a$  and 100 of the nearest adjectives (by cosine distance) to  $n$ . These adjectives are chosen from the set of the 1425 adjectives with concreteness ranking  $> 3.0$  in the concreteness corpus of (22), to exclude abstract nouns. Similarly, we select a set  $Q$  of projections corresponding to the hundred closest adjectives to the mean of the subject and predicate (the method of adjective choice in the baseline model), and take  $P_{L_Q}$  to be a uniform distribution over  $Q$ .

By tuning on an independent validation set of hand-selected metaphors, we choose  $\sigma_1 = \sigma_2 = 0.1$ . We select the adjectives corresponding to the two projections with highest marginal posterior mass under  $L_1^Q$  as the interpretations provided from our model in the experiment. We choose two rather than one since the model tends to distribute most of its marginal posterior mass over  $Q$  to at least two projections, intuitively reflecting the fact that there is usually more than one good interpretation of a metaphor.

**Experimental Methods** Tsvetkov et al. (23) provide a corpus of  $\sim 800$  AN metaphors, gathered by human annotators, from which we select  $\sim 100$  of the least frequent by bigram count, with  $n$ -gram data from the Corpus of Contemporary American English (24), in order to filter out conventionalized metaphors. Our full set of 109 metaphors is shown in figure 3. The experiment was run on Mechanical Turk, with 99 participants, all of whom are native English speakers. Participants who failed to follow instructions on a test item were excluded, leaving 60 participants (analyses remain significant if all participants are included). In the experiment, participants are shown a metaphor, as in figure 5 and asked to judge



**Fig. 5.** An item in the experiment. Item order, and in-item order of the 4 adjectives from  $L_1^Q$  and baseline models is randomized.

how relevant each proposed adjective (here, *debilitating*, *pervasive*, *corporate*, *addictive*) is to the metaphorical meaning of the AN phrase. In an example, they are told to rate *intense* as relevant to *fiery temper* “because a fiery temper is an intense temper” but *warm* as irrelevant.

- Grice HP (1975) Logic and conversation. 1975 pp. 41–58.
- Frank MC, Goodman ND (2012) Predicting pragmatic reasoning in language games. *Science* 336(6084):998–998.
- Kao JT, Bergen L, Goodman N (2014) Formalizing the pragmatics of metaphor understanding. in *CogSci*.
- Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality in *Advances in neural information processing systems*. pp. 3111–3119.
- Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. in *EMNLP*. Vol. 14, pp. 1532–1543.
- Black M (1954) Metaphor.
- Lakoff G, Johnson M (1980) Metaphors we live by. *Chicago, IL: University of*.
- Peters ME, et al. (2018) Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
- Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dai AM, Le QV (2015) Semi-supervised sequence learning in *Advances in neural information processing systems*. pp. 3079–3087.
- Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding with unsupervised learning, (Technical report, OpenAI), Technical report.
- Socher R, et al. (2013) Recursive deep models for semantic compositionality over a sentiment treebank in *Proceedings of the 2013 conference on empirical methods in natural language processing*. pp. 1631–1642.
- Coecke B, Sadrzadeh M, Clark S (2010) Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*.
- Shutova E (2016) Design and evaluation of metaphor processing systems. *Computational Linguistics*.
- Golland D, Liang P, Klein D (2010) A game-theoretic approach to generating spatial descriptions in *Proceedings of the 2010 conference on empirical methods in natural language processing*. (Association for Computational Linguistics), pp. 410–419.
- Gärdenfors P (2004) *Conceptual spaces: The geometry of thought*. (MIT press).
- Kintsch W (2000) Metaphor comprehension: A computational theory. *Psychonomic bulletin & review* 7(2):257–266.
- Linzen T (2016) Issues in evaluating semantic spaces using word analogies. *arXiv preprint arXiv:1606.07736*.
- Finley G, Farmer S, Pakhomov S (2017) What analogies reveal about word vectors and their compositionality in *Proceedings of the 6th joint conference on lexical and computational semantics (\*SEM 2017)*. pp. 1–11.
- Mitchell J, Lapata M (2010) Composition in distributional models of semantics. *Cognitive science* 34(8):1388–1429.
- Grefenstette E (2013) Category-theoretic quantitative compositional distributional models of natural language semantics. *arXiv preprint arXiv:1311.1539*.
- Brysbaert M, Warriner AB, Kuperman V (2014) Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods* 46(3):904–911.
- Tsvetkov Y, Boytsov L, Gershman A, Nyberg E, Dyer C (2014) Metaphor detection with cross-lingual model transfer in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1, pp. 248–258.
- Davies M (2011) Word frequency data: Corpus of contemporary american english. *Provo, UT: COCA*.