Pragmatics of Metaphor Revisited: Modeling the Role of Degree and Salience in Metaphor Understanding

Alexandra Mayn (amayn@lst.uni-saarland.de)
Department of Language Science and Technology, Saarland University
Campus C7.2, 66123 Saarbrücken, Germany

Vera Demberg (vera@coli.uni-saarland.de)
Department of Language Science and Technology;
Department of Computer Science, Saarland University
Campus C7.2, 66123 Saarbrücken, Germany

Abstract

One of the advantages of using metaphorical expressions over literal ones might be that speakers can convey not only the intended property, but also its degree. For example, when hearing "John is a shark", the listener might infer that the speaker aims to communicate that John is as mean as a typical shark. We present experimental findings supporting this hypothesis, along with a novel metaphor interpretation model, which is implemented within the Rational Speech Act framework. We compare our model's predictions to those of an existing RSA model of metaphor understanding, within which the listener infers just the presence or absence of a feature as opposed to its degree, and find that our model produces a significantly better fit.

Introduction

Metaphors abound in natural language and are therefore an important phenomenon to capture when modeling discourse processing. Metaphor comprehension has been approached from various perspectives (see Tendahl & Gibbs Jr, 2008 for a review), notably that of cognitive linguistics, which views metaphor as a reflection of people's mental conceptual organization (i.e., we think in metaphors) (Lakoff, 1993), and that of pragmatics, which views metaphor as an apt means to communicate some intended meaning (Wilson & Carston, 2006).

A prominent representative of the pragmatics approach, relevance theory (Wilson & Carston, 2006; Wilson & Sperber, 2002) argues that metaphorical utterances, just like literal language, are produced and interpreted in terms of their relevant aspects. There is experimental evidence supporting relevance theory's notion that property attribution is a principle underlying metaphor understanding (Glucksberg & Manfredi, 1997; Oka & Kusumi, 2020).

Kao, Bergen, and Goodman (2014) were the first to formalize metaphor understanding using the Rational Speech Act (RSA) Framework (Frank & Goodman, 2012), within which the speaker and listener recursively reason about each other's knowledge and intent to arrive at the intended meaning. RSA has been used to model a variety of nonliteral language uses in context, such as politeness (Yoon, Tessler, Goodman, & Frank, 2016) and puns (Kao, Levy, & Goodman, 2013).

Kao et al. (2014)'s approach builds on the ideas of relevance theory and property attribution. The authors argue that metaphor processing can be explained through basic pragmatic principles of communication. The crux of their idea is that when using an animal metaphor, the speaker aims to communicate features which are characteristic of this animal and are relevant to humans. For instance, when interpreting, "John is a shark", the listener is unlikely to think that John actually has fins and lives in the ocean. Instead, she assumes that the speaker aims to communicate a feature of a shark relevant to the referent, probably scariness or meanness.

We build on Kao et al. (2014)'s model but take a graded approach and represent features of the metaphor in terms of their typicality. We hypothesized that when hearing a metaphor, the listener infers not only the presence or absence of a certain feature, but also its *degree*. For instance, when hearing "John is a shark", the listener might interpret that the speaker wants to convey that John is as mean as a typical shark.

We also incorporate feature salience into our model. There is experimental evidence that feature salience plays a role in metaphor processing, and that high typicality is an indicator of salience (Katz, 1982). Our model captures that effect. In addition, it follows from our model that typicality of a feature is a spectrum and both ends of that spectrum are salient, in line with the notion of salience of the extreme as it has been used in the context of gradable adjectives (Franke, 2012). Therefore, we expect that an animal is also likely to be referred to if the feature in question is very *atypical* of it. Average typicality, on the other hand, we expect to not be salient and therefore less likely to be uttered and, if uttered, more confusing to interpret.

We test these predictions of our model experimentally and compare it to Kao et al. (2014).

Dataset

In Kao et al. (2014)'s model, an animal is defined as a vector of three binary features specific to that animal, disallowing comparison between animals. In our model, an animal has every feature to some, possibly very low, degree. Therefore, we collected a new dataset to serve as priors for our model.

Experiment 1a: Free-Response Feature Elicitation

In this experiment, preliminary adjective and animal lists were created.

Materials We compiled a preliminary list of 20 adjectives describing human personality traits. 14 of the adjectives were selected from Kao et al. (2014)'s list, and 6 additional ones

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

were included[1]. Out of that list of 20 adjectives, two lists were created containing 10 of the original adjectives and opposites of the remaining 10 (e.g., “dishonest” for “honest”.)

Methods 20 native English speakers, average age 32 (*sd*=11), 9 males, 11 females and one unspecified, where recruited on Prolific and received a compensation of £1.00. The participants read each of the two lists of 20 characteristics (10 participants per list) and were asked to type in any animals they associate with the trait in question. They were required to fill out at least 10 of the 20 fields in to determine which adjectives collected more responses and were therefore useful to keep.

Results Based on the obtained responses, the list was edited in the following way: if the participants’ responses had low agreement, i.e. many responses with little overlap, the corresponding adjective or animal was excluded. If both adjective opposites (e.g., “loyal”-”disloyal”) had high agreement, the one with relatively lower agreement was removed from the list.

Experiment 1b: Closed-Set Feature Elicitation

In this follow-up closed-set experiment, adjective and animal lists were finalized.

Materials We used the lists of animals and adjectives obtained in the free-response feature elicitation. Two lists of adjectives were compiled using the adjectives obtained in Experiment 1a. Each of the two lists contained 14 adjectives. Additionally, a list of 28 animals was compiled using the most common responses from the previous experiment.

Methods A group of 20 native English speakers who did not participate in the previous experiment, average age 35.7 (*sd*=11), 8 males and 12 females, where recruited on Prolific and received a compensation of £1.50. They were presented with the adjectives one at a time and a list of animals (obtained in Experiment 1a; 10 participants per list) and were asked to select the animals they associate with the adjective. Additionally, a text box was available to optionally type in animals which were not on the list.

Results The participants’ responses were further edited by excluding animals and adjectives with little overlap in participant responses. As a result, 4 animals and 4 adjectives were removed at this stage[2]; no new items were added.

Experiment 2: Typicality Elicitation

Next, we collected typicality priors for each animal-feature combination.

Materials Using the feature and animal lists from Experiment 1b, we created four balanced lists of adjectives, such that each adjective appeared on only one list.

Methods A group of 124 native English speakers, 31 per list, who did not participate in the previous experiment, average age 38 (*sd=12*), 54 males and 70 females, where recruited on Prolific and received a compensation of £1.25. They were presented with an animal along with the 20 features. For each feature, the participants used a slider bar to answer the question: “How typical is this feature for this animal?”, from “extremely atypical” to “extremely typical”. Typicality of each of these features for human males was also elicited. We use only males to be consistent with Kao et al. (2014).

Results The typicality ratings for each feature for each animal category (e.g., friendliness for a dolphin) were averaged, yielding a typicality rating for each animal category, including human. The obtained ratings were then examined individually to ensure that every animal used in final experiment had distinctive (i.e., rated very high or very low) features and every feature was distinctive for at least one animal. We removed 4 adjectives and 1 animal this way[3], resulting in a final list of 20 adjectives and 21 animals including human.

Model

In this section, we describe our model of metaphor understanding, highlighting the differences from Kao et al. (2014)’s approach.[4] Kao et al. (2014) extended the classical RSA model (Frank & Goodman, 2012) to include communicative goals in order to incorporate literally false utterances into the model, which we also adopt. We limit the scope of the types of metaphors to the type “X is a Y”, where X is a male name and Y is an animal category; in this way, the present work forms a natural extension to Kao et al. (2014).

In contrast to Kao et al. (2014), who represent an animal with a vector of typical binary features of length 3, we define an animal as a vector of size [total number of features], where each of the values is the feature’s typicality for a given animal. In this way, we aim to capture relative typicality of features both between and within animals.

The literal listener L_0 in our model is defined as follows:

$$L_0(c, deg(f) = d|u) = \begin{cases} P(deg(f) = d|c) \\ \quad \text{if } \; c = u, \\ 0 \quad \text{otherwise} \end{cases} \tag{1}$$

When hearing the utterance “John is a fox”, L_0 interprets it as John literally belonging to that category and having one of the corresponding features to some degree. $P(deg(f) = d|c)$ is the acceptability of naming animal c to convey the degree *deg(f)* of feature *f*. The closer the interpreted degree to the

[1] brave, calm, patient, reliable, shy, stubborn.

[2] camel, hippo, rhino, toad; disloyal, flexible, foolish, impatient.

[3] tiger; funny, happy, patient, reliable.

[4] Model code and collected priors are available at `https://github.com/sashamayn/metaphor_rsa_cogsci22`.

elicited typicality prior of this animal-feature combination, the higher the acceptability:

$$P(deg(f) = d|c) \propto \\ 1 - abs(typ(c,f) - d) - \varepsilon \tag{2}$$

So, when hearing *John is a fox*, the literal listener is more likely to interpret that John is a fox and is 0.2 loyal than that he is 0.8 loyal, since foxes are typically not very loyal. The literal listener considers all 20 features, 21 degrees for each: these 21 degrees are the typicality values elicited for the all animals including human for that feature. In contrast, the literal listener in Kao et al. (2014)'s model interprets a binary feature vector, e.g. that John is disloyal, graceful and cunning, when hearing that John is a fox; their model does not include degrees. We are not the first to use typicality for literal listener's word meanings. Graf, Degen, Hawkins, and Goodman (2016)'s model of reference levels uses typicality to represent the degree of acceptability of a label for an object.

The pragmatic speaker selects her utterance to fulfill her communicative goal g, which is to talk about a particular feature, and seeks to maximize her utility:

$$U(u|g, deg(f) = d) = \\ log \sum_{c,f} \delta_{g=f} L_0(c, deg(f) = d|u) \cdot P(f,c) \tag{3}$$

Our definition of the utility is equivalent to Kao et al. (2014)'s with one important difference: the inclusion of the salience term $P(f,c)$. This term denotes the probability of the speaker using a specific animal-feature combination, e.g. loyalty in combination with foxes. We assume that the speaker is unlikely to talk about average values of features – i.e., saying "John is a giraffe" to mean "John is averagely loyal" because giraffes are not known for being either loyal or disloyal, so despite high acceptability, *giraffe* in this case has low salience and is unlikely to be used by the speaker. We test this assumption experimentally in Section 4. We define the probability $P(f,c)$ as the KL-divergence between a normal distribution representing this animal-feature combination, where the mean and *sd* are the typicality priors for this animal-feature combination from Experiment 2, and a neutral normal distribution centered around 0.5 with *sd*=0.15[5]. Therefore, the further removed the animal-feature combination's typicality is from 0.5, in either direction, the more likely the speaker is to want to refer to this animal to convey the feature.

The pragmatic speaker chooses an utterance which maximizes her utility. α is the speaker rationality hyperparameter.

$$S_1(u|g, deg(f) = d) \propto e^{\alpha U} \tag{4}$$

Finally, the pragmatic listener equation is again similar to that of Kao et al. (2014) with the important difference that the listener interprets *the degree* of an individual feature as opposed to a binary feature vector. The pragmatic listener reasons about the speaker and her possible communicative goals:

$$L_1(c, deg(f) = d|u) \propto P(c) \cdot P(deg(f) = d|c) \\ \cdot \sum_g P(g) \cdot S_1(u|g, deg(f) = d) \tag{5}$$

$P(c)$ is the category prior for the referent, that is, how likely the referent is to be human or the uttered animal. It is a hyperparameter to be fit to the data. $P(deg(f) = d|c)$ represents the probability that a member of category c has a feature f to the degree d – it is the same as the acceptability term in the L_0 and is defined the same way. $P(g)$ is the listener's prior about the feature being communicated given the conversational context (or Question Under Discussion, QUD). A feature might be more likely to be interpreted if it is asked about.

Experiment 3: Metaphor Interpretation Experiment

We now proceed to the main experimental question concerning the role of typicality, salience and QUD on metaphor understanding.

Materials For each non-human animal category we created 4 scenarios (1 for the vague and 3 for the specific communicative goal) of the type used by Kao et al. (2014), in which Bob is talking to his friend about a person he recently met.

In the vague condition, Bob's friend asks the vague question "What is John like?", to which Bob replies by saying "He is a c_a". For the specific goal condition, the friend asks a question mentioning a specific feature, "Is he f_i?". There are three possible scenarios in this condition – an animal category for which the feature in question is extremely typical, neither typical nor atypical, and extremely atypical. Our assumption is that the two extreme cases share high salience, while the averagely typical case lacks it. Table 1 includes examples of each condition.

Table 1: The four conditions in the Metaphor Interpretation Experiment.

G	Typ.	QUD	Utterance
v	–	"What is John like?"	"He is an ox."
s	high	"Is John loyal?"	"He is a dog."
s	avg	"Is John loyal?"	"He is a sloth."
s	low	"Is John loyal?"	"He is a snake."

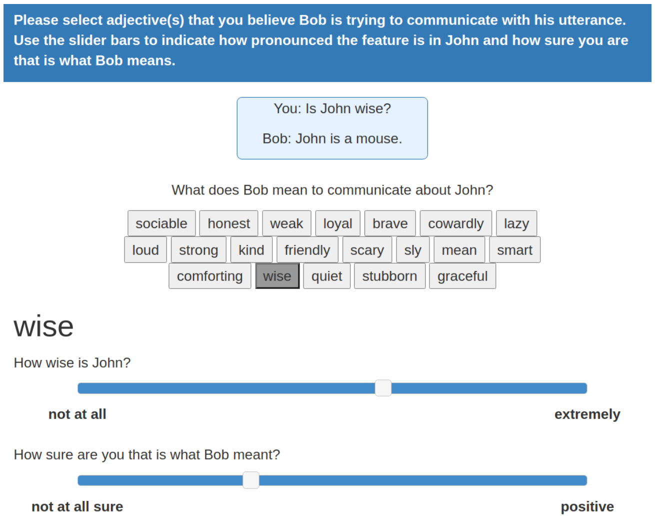
For each of the 20 animals, we took the 2 most typical, 2 least typical, and 1 averagely typical adjective, resulting in 100 items in the specific condition. There were 20 items in the vague condition, one per animal. Out of those 120 items,

[5] The *sd* of the neutral distribution is hypothetically a hyperparameter. We found that changing its value didn't matter very much for the model fit, but fit was slightly higher with *sd*=0.15, so we used that value.

we made 5 balanced lists, such that no adjective appeared more than once and no animal appeared more than twice.

Method 100 native English speakers who did not take part in any of the previous experiments, average age 38 (*sd*=13), 31 males and 69 females, were recruited on Prolific and received a compensation of £1.75. We aimed at 20 participants per list but for technical reasons participants were assigned in a way which resulted in 18 to 22 ratings per list.

In each trial, participants saw a box with a question-answer pair similar to those in Table 1. An example trial is shown in Figure 1. First, they were asked to select at least one adjective from a list of possible interpretations. In the specific condition, the adjective contained in the question was pre-selected but could be unselected. For each selected adjective, the participants used slider bars to provide a degree rating to indicate to which extent John has the selected property and their certainty of that interpretation.



Please select adjective(s) that you believe Bob is trying to communicate with his utterance. Use the slider bars to indicate how pronounced the feature is in John and how sure you are that is what Bob means.

You: Is John wise?

Bob: John is a mouse.

What does Bob mean to communicate about John?

How wise is John?

How sure are you that is what Bob meant?

Figure 1: Metaphor interpretation experiment setup.

We compared the responses of each participant to the rest to identify any participants who responded at random. We looked at a subset of points where we would expect a certain response (e.g., *quiet-mouse* we expected to be rated highly) and determined which participants' responses diverged from the rest in a seemingly arbitrary way. 3 participants were excluded at this stage.

Results We averaged the obtained degree and certainty ratings. We only considered feature-animal combinations for which we had at least 10 responses.

First, we correlated the typicality priors from Experiment 2 with the obtained degree ratings to test our assumption that people interpret the degree of a relevant feature when hearing a metaphorical utterance. Pearson's r=0.93, p<0.001. This suggests that indeed, when hearing "John is a bear" in response to "How strong is John?", the listener interprets it as John having *the same degree* of strength (presumably compared to other humans) as a typical bear.

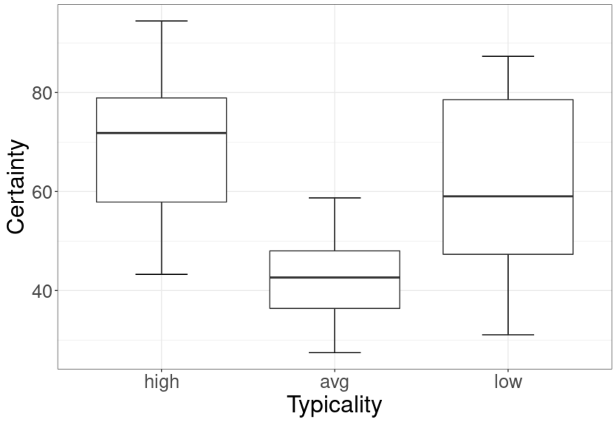


Figure 2: Certainty is significantly lower for the average typicality condition.

The average degree ratings by typicality condition are 77.13 (*sd*=12.75), 41.25 (*sd*=12.51), and 20.88 (*sd*=15.41) for the high, average, and low typicality conditions respectively (Figure 2). The low degree rating in the *low* typicality condition confirms our hypothesis that metaphors can indeed be interpreted inversely (e.g. John is *not loyal* when hearing "John is a fox") when an atypical referring expression is uttered.

Another interesting finding is revealed when analyzing the relationship between typicality and certainty ratings. Typicality of the interpreted trait for the uttered animal is a significant predictor of interpretation certainty (R^2 = 0.29, F(2,98)=19.68, p<0.001). Certainty is highest for the high typicality condition (69.86, *sd*=14.3), followed by low typicality (61.11, *sd*=16.42); certainty is lowest in the average typicality condition (44.62, *sd*=11.77). The difference between means is significant (F(2,98)=19.68, p<0.001 on one-way ANOVA); all means are significantly different (all ps<0.05 on post-hoc Tukey tests). This suggests that high and low typicality are both salient while average typicality is not, and that more salient animal-feature combinations result in higher interpretation certainty.

The average typicality for the vague condition is 83.36 (*sd*=8.84), and the certainty is is quite high, 71.45 (*sd*=17.29), meaning that when provided with no context, people interpret the metaphor as a highly typical feature and are fairly certain of that interpretation.

Model Evaluation

We used the typicality priors collected in Experiment 2 to compute the predictions of our model and compare them to human judgements.

We performed a hyperparameter search and fit α=1, $P(g)$= 0.05, and $P(c_{human})$=0.99.

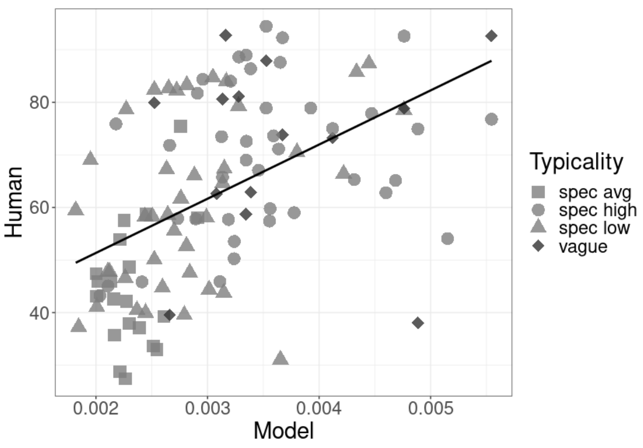


Figure 3: Human certainty ratings vs. model probabilities for the 120 metaphors. The model achieves a fit of r=0.5.

We correlated the probability ratings of the L_1 from the model with the certainty ratings from Experiment 3 (Figure 3). The certainty ratings represent the participants' perceived likelihood that the speaker meant to communicate the interpreted feature to the interpreted degree.[6] The Pearson correlation coefficient r was 0.5 (p<0.001), suggesting that such a model captures important principles at play in metaphor understanding, namely relative salience of degrees within and between metaphors, while there still being considerable individual variation.

Unexpectedly, model fit is best when the goal prior $P(g)$ is uniform, seemingly negating the role of conversational context in metaphor interpretation. The reason for that appears to be our dataset: there are only two data points where the interpreted feature was not the one in the question, both of which had high certainty, *weak-sloth* and *kind-snake*. Namely, "He is a sloth" in response to "Is John weak?" was interpreted as the speaker wanting to communicate *laziness*, the reasoning presumably being as follows: if the speaker had wanted to communicate weakness, she would have chosen an animal for whom weakness is salient. However, she chose a sloth, for whom the most salient feature is laziness, so that must be what she means." The other data point is the pair *kind-snake* being interpreted as the speaker wanting to communicate *slyness*. There were more such mismatched responses which we did not include in our dataset because fewer than 10 participants agreed on them. However, in a post-hoc analysis, we looked at all the mismatched responses people gave, and found that their certainty was 65.6 (sd=22.5), comparable to the overall certainty of the highly typical responses (69.86, sd=14.3). In other words, participants seemed to only give a mismatched response if they were fairly sure it was correct, and as a result, we have no mismatched data points with low certainty, which would have resulted in a higher QUD prior.

Comparison to Kao et al. (2014)

We also ran Kao et al. (2014)'s model on our data. This required some adjustments. Because in their model, an animal is defined as a vector of binary 3 most typical features,[7] for each item in our dataset that did not include one of those, we swapped out the third most typical feature for the one in question to calculate a prediction. Also, in order to apply their model to our data, we needed to obtain prior probabilities for the binary feature vectors. When the value in the feature vector was 0, signifying the absence of a feature, we computed the probability as *1-typicality* of that feature for that animal.

Since we only had typicalities for the individual features, we needed to combine them in some way. To do that, we compare averaging and taking the minimum and the maximum of the three individual probabilities; of those, slightly better results are achieved when taking the minimum. We then computed the model predictions as described in Kao et al. (2014) and compared them to the human certainty ratings, like for our model. Since in our model, the degree of a feature is real-valued, and in Kao et al. (2014)'s model it is binary, we had to binarize the predictions: whenever the interpreted degree was greater than some threshold, the listener inferred 1 (e.g. that John is loyal), otherwise 0 (i.e. John is disloyal). We tried out different thresholds. Figure 4 displays Kao et al. (2014)'s model fit in terms of the binarization threshold and the three ways of combining individual probabilities. The highest fit is achieved with the threshold=0.3 and taking the minimum of the three probabilities. That is quite a low threshold; when it is raised to e.g. 0.5, the model performs poorly. This suggests that people are fairly conservative in their inverse interpretations. For instance, if the feature under discussion is meanness, then the animal has to be quite low in meanness (below 0.3 out of 1) for the inverse interpretation *not mean* to occur. Anything above that people still tend to interpret positively, in line with the polarity of the QUD.

The correlation between Kao et al. (2014)'s model's predictions and participant judgements collected in section was at most r=0.26 (p=0.01), which is significantly worse than for our model (t=2.09, p=0.02 on a two-tailed paired correlation test). While it is true that an average or min of three individual feature priors might not correspond perfectly to their joint prior, it is unlikely to lead to a dramatic difference in predictions. Therefore, we conclude that our model is better able to account for the data beyond incorporating degrees, providing further evidence that relative salience and alterna-

[6] Perhaps a more direct way of phrasing the question to get at likelihood might have been "How *likely* does it seem to you that this is what John meant?" as opposed to "how *certain*". However, we think that for our experiment, these two phrasings tap into the same concept. A more direct phrasing would be more appropriate if the feature interpreted was a nonhuman feature, e.g. that John has fins when hearing "John is a shark". In that case, the listener is not uncertain that the speaker didn't mean fins but rather she is certain that it's very unlikely.

[7] It is hypothetically possible to define animals as vectors of more than 3 features in this model but that quickly gets computationally unrealistic. Including all 20 features would involve enumerating 2^{20} vectors for each item.

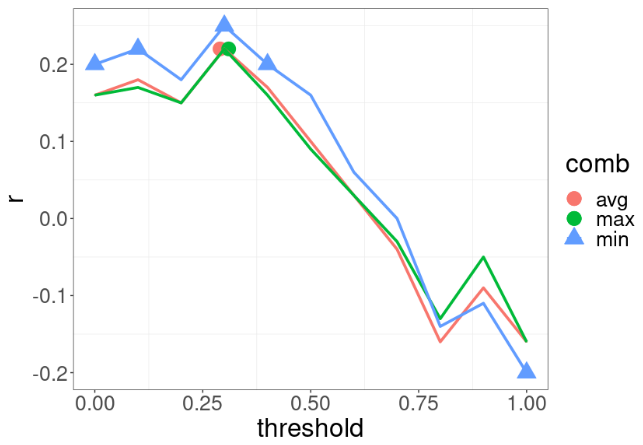


Figure 4: Kao et al.'s model fit to our data as a function of the binarization threshold. Significant correlations are displayed as points on the line. The best fit of r=0.26 is achieved when taking the minimum of the individual probabilities and setting the threshold to 0.3.

tive metaphorical utterances indeed play a role in metaphor interpretation.

Table 2: Model comparison

model	fit (Pearson's r)
our model	0.5
Kao et al.	0.26

Discussion

We proposed an RSA-based model of metaphor comprehension which builds on Kao et al. (2014)'s model but takes a gradient approach. We hypothesized that when hearing a metaphor, the listener interprets it as a degree of a relevant feature. We also assumed that salience affects metaphor interpretation, and that both high and low typicality are salient while average typicality is not. Both assumptions were supported by the experimental results.

While our model captures several important facets of the process at hand, achieving a fairly good fit (r=0.5), there is a lot of unexplained variance. We saw that there are 3 data points in particular (Figure 3) where the model assigns a much higher probability to the metaphor-feature pairs than our participants. Those are "sly mule" in the specific low condition, to which participants assign a near-zero probability, "loud parrot" in the vague condition, and "lazy cat" in the specific high condition. When those three data points are excluded, model fit increases to r=0.59 ($p < 0.001$). The average typicality of slyness for a mule is 27.6 so the model presumes it to be salient and therefore assigns it high probability; participants, however, give it a low rating. It seems likely that one of the reasons for that is that there's a very salient association of mules and stubbornness ("stubborn as a mule" is a conventionalized phrase), so participants might assume that the speaker is unlikely to use the mule for anything else. Interestingly, the mean typicality rating for stubbornness of a mule is not as high as we might expect (73.37), so the model which does not know anything about this pre-existing association strength treats the pairings "stubborn mule" and "sly mule" as approximately equally probable since they are equally far from the mean. This points, for one, to the importance of distinguishing between novel and conventionalized metaphors in modeling, which could be explored in future work. This also suggests that there might be an asymmetry in salience of two the ends of the typicality scale: perhaps only more extreme low typicality values are salient. This is supported by our experimental results (Figure 2).

A factor which presumably plays a large role is individual variability. We took averages when computing the typicality priors, and the span for some responses is quite large. For example, individual ratings for the *dog-kind* combination range from 37 to 100. We all have slightly different associations with animals, which can also vary depending on the context. One could give a rating of 37 for kindness to a guard dog and one of 100 to a puppy, for instance. A sentence-level model will not be able to capture these context effects. Additionally, the way metaphors are used and interpreted is influenced by the cultural context: for example, *lucky* might be a salient property of a cat in the context of Japanese culture, but arguably not in the context of European culture.

There are other questions we hope to explore in future work. First, our model assumes that the speaker is only trying to communicate the degree of one feature but it is possible that the speaker's intention and part of the reason for choosing to use a metaphor in the first place is to talk about multiple features (Kao et al., 2014). Second, currently we only consider metaphorical utterances but there is, naturally, always the possibility of saying the literal version of the message. For instance, instead of using a low-salience combination like "John is a giraffe" to say that John is averagely kind, the speaker could say "John is pretty kind".

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. This project is supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 948878).

References

Frank, M., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*, *336(6084)*.

Franke, M. (2012). On scales, salience and referential language use. In *Logic, language and meaning* (pp. 311–320). Springer.

Glucksberg, M., & Manfredi, D. (1997). Property attribution in metaphor comprehension. *Journal of memory and language*, *36*(1), 50–67.

Graf, C., Degen, J., Hawkins, R., & Goodman, N. (2016). Animal, dog, or dalmatian? levels of abstraction in nominal referring expressions. *CogSci*.

Kao, J., Bergen, L., & Goodman, N. D. (2014). Formalizing the pragmatics of metaphor understanding. *Proceedings of the annual meeting of the Cognitive Science Society*, *36*.

Kao, J., Levy, R., & Goodman, N. D. (2013). The funny thing about incongruity: A computational model of humor in puns. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Katz, A. N. (1982). Metaphoric relationships: The role of feature saliency. *Journal of Psycholinguistic Research*, *11*(4), 283–296.

Lakoff, G. (1993). The contemporary theory of metaphor.

Oka, R., & Kusumi, T. (2020). Distinctive features influence perceived metaphor aptness and preference for metaphor use. *Metaphor and Symbol*, *35*(1), 12–22.

Tendahl, M., & Gibbs Jr, R. W. (2008). Complementary perspectives on metaphor: Cognitive linguistics and relevance theory. *Journal of pragmatics*, *40*(11), 1823–1864.

Wilson, D., & Carston, R. (2006). Metaphor, relevance and the 'emergent property'issue. *Mind & Language*, *21*(3), 404–433.

Wilson, D., & Sperber, D. (2002). *Relevance theory*. Blackwell.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2016). Talking with tact: Polite language as a balance between kindness and informativity. *Proceedings of the 38th annual conference of the cognitive science society*.