LAZY AS AN ANT:

Formalizing The Role of Typicality, Salience, and Alternative Utterances in Metaphor Interpretation

ALEXANDRA MAYN[*]

CONTENTS

ABSTRACT

Although metaphors have been extensively studied in cognitive science, many important questions remain: what factors play into the interpretation of a metaphor? Why are certain metaphors more used than others and why are some easier to interpret than others? We propose a computational model of metaphor understanding within the Rational Speech Act framework that builds on the work of Kao et al. and looks at metaphor interpretation through the lens of typicality and gradient salience. We hypothesize that other features of an entity, as well as possible alternative entities have an impact on how a metaphor is interpreted. We also hypothesize that metaphors can be interpreted inversely when the feature in question is saliently atypical for the entity used as a metaphor. Finally, we assume that when an object is used as a metaphor to convey a certain feature and it has an average typicality for that feature, that will make the metaphor confusing and prone to misinterpretation. We run our model on a small demo dataset and propose a set of behavioral experiments to test the model's predictions.

[*] *Department of Computational Linguistics and Phonetics, Saarland University, Saarbrücken, Germany*

1 INTRODUCTION

Metaphors abound in natural language and are therefore an important phenomenon to capture when modeling processing of discourse. There is a growing body of literature investigating interpretation, processing, and difficulty of metaphors [1]. There have also been attempts to model metaphors probabilistically. Kao et al.[2] were the first to formalize metaphor understanding using the Rational Speech Act Framework[3]. They limit their scope to metaphors of type "X is a Y", where X is a human and Y is a member of an animal category. The crux of their idea is that the speaker wishes to communicate features which are characteristic of this animal and are relevant to humans. For instance, if I tell you that "John is a shark" and you know who I am referring to, you are unlikely to think that John actually has fins and lives in the ocean. Instead, you will probably think that I want to communicate a feature of John's personality by invoking an animal of whom this feature is representative; in this case, probably scariness or meanness.

We would like to propose an alternative model of metaphor understanding which builds on Kao et al.'s work but takes a graded approach and represents features in terms of their salience. We also focus on metaphors of the form "X is a Y", where X is a human male and Y is an animal category. Kao et al. define each category by a vector of 3 binary features which are different depending on the animal category. Our model, on the other hand, defines a member of a category as having all possible features to some degree (a real value between 0 and 1), depending on how salient a given feature is for a certain animal category. This gradient approach captures several important aspects of metaphor interpretation: it takes into account alternative interpretations of an utterance by capturing relative salience of features within an animal ("an ant is more saliently strong than it is fast") as well as alternative utterances for the same intended meaning ("an ox is more saliently strong than an ant"). We derive salience from typicality with the assumption that both very high and very low typicality are salient, whereas average typicality is not and is therefore less likely to be referred to. This presupposes, so to speak, "negative" or inverse use of metaphors which Kao et al. do not consider, e.g. saying "John is a fox" to mean that John is *not* loyal because loyalty is very atypical (and therefore salient) for a fox.

We demonstrate our model's predictions on a small demo dataset and propose a set of behavioral experiments which would allow to show how well this approach to formalizing metaphor interpretation corresponds to human performance. If human data corroborate the model's predictions, that will provide support for the hypothesis that, first, alternative utterances and features play a role in metaphor interpretation, and, second, that metaphors can be interpreted inversely in certain contexts and, lastly, that a metaphor is hard to understand and is prone to interpretation errors if the typicality of the feature in question is not salient for the category.

2 COMPUTATIONAL MODEL

We model metaphor understanding using the Rational Speech Act framework introduced by Frank & Goodman[3] which has been successful at modeling a variety of nonliteral language uses in context including politeness[4], puns[5], and metaphors[2]. At its core is the idea of the listener and the speaker recursively reasoning about each other to arrive at a common interpretation: a first-level pragmatic speaker reasons about a literal listener and acts in such a way as to maximize her utility, which is based on the informativeness of the utterance. In turn, the pragmatic listener reasons about the pragmatic speaker to recover the intended utterance. Kao et al.[2], who first formalized metaphor understanding using RSA, extended the model 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 our methodology and findings will form a natural extension to Kao et al.'s work.

Kao et al. define possible features of X as a vector of features of length 3, in which each of the features are binary. The limitations of that approach are that, first, it does not represent relative salience of features both between and within animals (i.e., of types "an ant is more saliently strong than it is fast" and "an ox is more saliently strong than an ant"), which means that it does not take into account alternative utterances and alternative interpretations of a given utterance, which, we suppose, both play a crucial role in metaphor interpretation. Therefore, we define an animal as a vector of size [total number of features], where each of the features is its normalized salience for a given animal. This approach also allows for "negative" use of metaphors which is also worth investigating – do people in some contexts interpret "He is a fox" as "He is not loyal", for instance, since foxes are associated with being sly and cunning and therefore have a low typicality (and, therefore, high salience) for loyalty? In this paper we only consider the listener interpreting the metaphor as one feature but we hope to extend it to make it possible to interpret multiple features in future work. Based on the above considerations, we define the literal listener Lo as:

$$L_0(c, f|u) = \begin{cases} \text{salience(f,u)} & \text{if } c = u \\ 0 & \text{otherwise} \end{cases}$$

where *f* is the feature in question, *c* is the animal category of the metaphor, and *u* is the utterance. The literal listener will hear the utterance "John is a shark" and interpret it as John literally belonging to the category "shark" and having the feature in question to the extent to which it is salient for a shark. We are not the first work to define the literal listener gradiently. Graf et al.[6] define their literal listener as the degree of acceptability of an object for a given category.

The pragmatic speaker acts in accordance with her utility, which is defined as the log of the literal listener. This is similar to the literal listener equation in Kao et al. but without a projection of the feature vector along one feature dimension since only one feature is being considered:

$$U(u|f) = log(L_0)$$

The choice of the utterance is governed by a softmax choice rule based on utility of the utterance given the feature that the speaker wants to communicate:

$$S_1(u|f) \propto e^{\lambda U(u|f)} = U(u|f)^{\lambda}$$

where λ is an optimality parameter that is fit to the data. As Kao et al. point out, since the speaker's goal is to communicate the strength of the feature (the scariness of a shark), S_1's goal is satisfied because she knows that the literal listener L_0 will interpret the utterance as person being literally a shark and therefore having the feature of being scary. Finally, we define the pragmatic listener as:

$$L_1(c, f, i|u) \propto P(c) \cdot P(f|c) \cdot P(f|i) \cdot S_1(u|f)$$

The definition of the pragmatic listener consists of the following components:

- $P(c)$ is the prior that the entity in question is either a human or a non-human animal. Following Kao et al., we set $P(c = c_a) = 0.01$ for every animal and $P(c = c_h) = 0.99$. We assume that the only two interpretations the listener is considering are a person and that animal.
- $P(f|c)$ is prior probability that a feature will be referred to given the category. We define it as a the feature's salience for a given category. Our method for deriving salience from typicality is described in detail in 2.1; it is important

to note that both low and high typicality have high salience (e.g., it is salient that a fox is *not* loyal, i.e. that loyalty is not a typical attribute of a fox).

- $P(f|i)$ is the probability of talking about a feature given the intention - that is, the speaker's communicative goal - of the speaker as understood by the listener. The two communicative goals we consider are *specific*, when the listener assumes that the speaker seeks to communicate a specific feature, and *vague*, when the feature the speaker is trying to communicate is unclear to the listener. For the demo dataset, we set the $P(f|i = specific)$ to 0.5 but in general it is a parameter that can be fit to the actual data but is expected to be at least 0.5; $P(f|i = vague)$ is uniform for all features.

2.1 Deriving Salience from Typicality

Frank and Goodman[3] introduce the idea of salience in their model as the prior probability that an object will be referred to. We make the assumption here that when the presence of a feature is a spectrum, both ends of that spectrum are salient, and therefore an animal is likely to be referred to also if the feature in question is very *atypical* of it. We define the salience of a feature for an object as follows:

$$salience(f, o) = |typicality(f, o) - \mu)|^{\kappa}$$

where (normalized) typicality is a real number between 0 and 1 and μ and κ are hyperparameters to be fit to the data. μ shifts the typicality values to be centered around 1-μ, the absolute value ensures symmetry of typicalities around the center; κ allows for a better fit of the function to the data. The salience matrix is then normalized by column (animal), so that salience of a feature can be interpreted as a probability that it will be referred to.

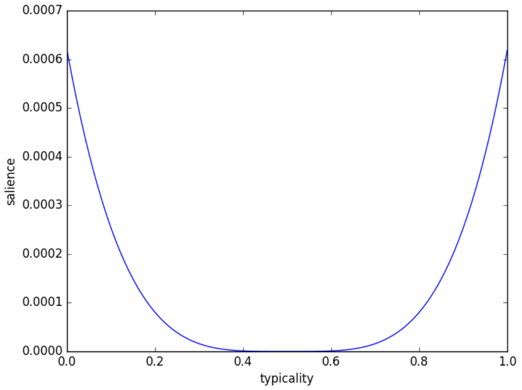

Figure 1: Salience for a given typicality value with $\mu = 0.5$ and $\kappa = 4$

For the demo dataset, we set $\mu = 0.5$ and $\kappa = 4$. It is noteworthy that here we are making a tentative assumption that salience is symmetric for the two ends of the typicality spectrum. However, it does not have to be the case. In a different context, Frank[7] found that people did not violate literal semantics to refer to an object when doing so was the only way to uniquely identify it; that is, they did not utter X to mean "the object that does not have an X". Therefore, it is possible that salience is asymmetric and that high atypicality of a feature is less salient than typicality that is at an equal distance from the mean. This will be investigated in the behavioral experiments and then μ can be fit to the data accordingly.

3 BEHAVIORAL EXPERIMENTS

We propose a series of behavioral experiments to elicit people's interpretation of metaphors in vague and specific goal conditions with varying typicality.

3.1 Feature Elicitation

MATERIALS 20 common non-human animal categories which share some characteristics (e.g., ants and oxes are both strong) and a preliminary list of characteristics typical of at least one of the animals (the purpose of this experiment is to make this list closer to human judgment).

METHOD 25 native English speakers will read animal categories along with the full preliminary list of characteristics. The will be asked to put a check next to characteristics they associate with that animal. Also, to ensure that the list of features captures all important features describing these animals, for each animal along with the prepared list there will be a box to have the participants type in features that are in their view characteristic of that animal if they are not on the list prepared by the author.

RESULTS Based on the participant responses, the list will be edited in the following way: if a feature from the preliminary list was not checked (or was only checked once) as being characteristic of one or more of the animals, it will be removed from the list. Then each of the adjectives participants inputted in the free-response box will be examined. The author will use her judgment to add some of those adjectives to the final list based on whether synonyms already exist on the list and whether multiple participants inputted the same feature.

3.2 Typicality Elicitation

MATERIALS Using the feature list obtained in the experiment above, we will now elicit typicality ratings from a separate group of 25 native English speakers.

METHOD Each participant will see every animal category along with the full list of features. They will then use a 7-point Likert scale to answer the question: "How typical is this feature for this animal", 0 being "not at all typical" and 6 being "highly typical". Typicality of each of these features for human males will also be elicited.

RESULTS The typicality ratings for each feature for each animal category (e.g., friendliness for a dolphin) will then be averaged and normalized (divided by 6, the maximum typicality value), resulting in a typicality rating for each animal category, including human.

3.3 Metaphor Interpretation

MATERIALS Based on the animal categories and typicality ratings obtained in the first two experiments, we now proceed to the main experimental question. Since we are interested in the role of typicality and salience, as well as of communicative goal, on metaphor understanding, for each of the 20 non-human animal categories we we will create 4 scenarios (1 for vague and 3 for the specific communicative goal) of type used by Kao et al.[2], in which Bob is talking to his friend about a person he recently met.

In the case of a vague communicative goal, Bob's friend asks a vague question "What is he 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, moderately typical, and extremely atypical. What the two extreme cases share and what the averagely typical case lacks is high salience. Table 1 includes examples of each condition.

Table 1: Utterance conditions

Goal	Typicality	Example question	Example utterance
vague	–	"What is John like?"	"He is an ox."
specific	high	"Is John loyal?"	"He is a dog."
specific	average	"Is John loyal?"	"He is a dolphin."
specific	low	"Is John loyal?"	"He is a fox."

METHOD 24 native English speakers will participate in this experiment. Every participant will see all data points in the vague condition and one data point for each animal in the specific condition (20 x 2 = 40 trials in total). This will result in 24 data points per animal in the vague condition and 8 data points for each combination of animal and feature with a given typicality level (e.g., dog-loyal).

For each trial, the participants will see a question-answer pair and will be asked to interpret the metaphor. First, they will select whether they think John is an animal or a person. Second, they will be asked: "What did Bob mean to say about John?" and they will have to select from a drop-down list of features and for each feature they will use a 7-point Likert scale to indicate the degree to which the feature is pronounced in John (e.g. '0 - John is not at all loyal' to '6 - John is very loyal'.) Finally, for each data point they will answer the question: "How certain are you that that is what Bob meant?" by reporting a number on a 7-point Likert scale.

RESULTS For the vague goal condition and for each typicality subcondition of the specific goal we obtain the following characteristics (by averaging):

1. The percent of responses that John is a person.
2. For the vague condition, the features interpreted and the degree to which they are present in John.
3. For each typicality subcondition of the specific goal condition, proportion of items in which the intended feature was indeed the one interpreted.
4. For each typicality subcondition of the specific goal condition, the degree to which a feature is present in John (only for correctly interpreted features; out of 6, later normalized).
5. Average certainty ratings (out of 6, later normalized).

We then perform the following statistical comparisons to test our predictions:

1. **Certainty that John is a human:** Proportion of responses that John is a human. We expect that to be very high and not significantly different between the 4 subconditions. That would illustrate that people were able to deduce that the utterance was being used metaphorically.
2. **Correct inference of feature:** For each subcondition of the specific goal condition, pairwise comparisons of proportion of correctly interpreted features. We expect that, in general, the average typicality sub-condition will have significantly more misinterpretations than the other two because the feature is not (positively or negatively) salient for the animal. Prediction about the direction of difference between the high and low typicality conditions are harder to make: it seems probable that there will be more misinterpretations in the low typicality case, which would suggest that people use metaphors "positively" more often; but it could also be the case that there would be no difference in the proportion of correct interpretations for the two categories.
3. **Informativeness:** The way in which the utterance informed us about John is represented by the difference between the priors for humans and the ratings of the degree to which the feature is present in John. We expect, first of all, the averages for the vague condition and for the specific high typicality subcondition to be positive (the listener learned that the feature is present in John) and the average for the low typicality condition to be negative (the listener learned that the feature is absent in John, i.e. that John is **not** loyal).

It is again harder to make predictions about the average typicality case; it is probable that the difference is positive based on our tentative hypothesis that people use metaphors "positively" more often. In terms of the absolute value of the change, we expect it to be significantly smaller for the average typicality condition (it is more confusing and less informative) than for either of the other two specific subconditions and for the vague condition. Following Kao et al., we expect the change to be greater in the high typicality subcondition than in the vague condition.

4. **Certainty:** We expect certainty ratings to be the lowest for the average typicality subcondition and the highest for the high typicality condition. The low typicality condition and the vague condition may or may not be signficantly different from the high typicality condition.
5. **Features inferred for the vague condition:** Average salience of the features inferred for the vague condition. We expect that number to be high (the metaphor conveys salient features of an animal.)

4 MODEL EVALUATION

We used a handwritten demo dataset with 10 features and 6 nonhuman animal categories (see 7.1) to run the model. To be able to compare the model predictions to the experimental data, we would need to use the typicality priors elicited in the Experiment described in 3.2. When deriving salience, we fit the hyperparameters to the data, which resulted in $\mu = 0.5$ and $\kappa = 4$. We set $P(f|i) = 0.5$ when the goal is to communicate the feature in question and distributed the rest of the probability mass uniformly across the rest of the features. In other words, the speaker is pretty sure that in response to a question about a person's scariness she will get an answer about their scariness but there is a small probability that the speaker will be communicating some other feature, e.g. the person's loyalty. Following Kao et al., we set $P(c_a) = 0.01$, indicating that there is a very small but non-zero probability that the entity in question is a member of a nonhuman animal category. We also fit $\lambda = 1$ for the speaker's utility.

In all cases, regardless of the intention and typicality of a given feature, our model inferred that the entity described was a person and not an animal ($P(c_p|u) = 1.00$).

In the uniform goal condition, the feature interpreted is the one associated with the highest salience (typicality of 0 or 1.)

To look at the three typicality subconditions, we consider 0 and 1 as low, 2-4 average, and 5 and 6 high typicality. We get the following results:

Table 2: Correct predictions and typicality

Typicality	Proportion of correct predictions
high	0.46
avg	0.18
low	0.27

These numbers are preliminary as they are based on very few data points but the pattern suggests that the model captures the hypothesis stated in 3.3: it is easiest to interpret a metaphor if it is about a typical feature, followed by an atypical feature. Interpreting metaphor with average typicality for a feature is very difficult.

Finally, average probabilities per typicality category (Table 3) are the model equivalent of certainty ratings. We can see that the numbers are the highest for high typicality, closely followed by low typicality, and noticeably lower for the vague condition and the average typicality feature-animal combinations.

Table 3: Prediction probability and typicality

Condition	Average probability
high typ	0.0086
avg typ	0.003
low typ	0.0084
vague	0.0049

5 DISCUSSION

We have proposed a computational model of metaphor interpretation in the Rational Speech Act framework which combines basic principles of communication and communicative goals with gradient salience of features to arrive at an interpretation. We propose behavioral experiments to obtain feature priors and test predictions of the model.

The model predictions suggest several conclusions about the nature of metaphor interpretation: most importantly, multiple factors play a role. In a vague context where the communicative goal of the speaker is not clear to the listener, the model predicts that the listener will interpret the metaphor as its most salient (which is to say, most or least typical) feature. When it is known to the listener that the speaker is trying to communicate a specific feature, the probability of that feature being interpreted increases. However, in our model, the listener is always considering both the relative salience of features within the metaphor and comparing it to the alternatives the speaker could have uttered. For example, if the listener hears "John is a fox" in response to her question "Is John scary?", the model predicts that the listener will think that the speaker is trying to communicate John's gracefulness because if she had wanted to say that John *is* or *is not* scary, she would have chosen an animal that has a high or low typicality, respectively, for scariness. However, the fact that she chose to say "fox" suggests that she wants to communicate something about John that is salient for a fox, namely gracefulness.

Klintsch et al.[1] hypothesized that a metaphor is difficult to interpret if there are very few lexical items that are strongly related to both the metaphor and the intended referent. Our model makes another prediction about what can make a metaphor hard to understand. It suggests that if the referring object has an average typicality rating (i.e., low salience) for the feature of interest, that makes the metaphor confusing and leads to misinterpretation.

We hope that the current work will make a contribution to the understanding of pragmatic metaphor interpretation. While our model captures several important features of the process at hand, there are several limitations we hope to address in future work. First, our model currently assumes that the speaker is only trying to communicate 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 communicate multiple features[2]. Secondly, currently we only consider alternative metaphorical utterances but there is, naturally, always the possibility of saying the literal version of the message which needs to be taken into account.

6 ACKNOWLEDGEMENTS

The author would like to thank **Martin van Harmelen** for assistance with implementation and helpful insights about deriving salience from typicality.

7 APPENDICES

7.1 Demo dataset

We used the following small manually created dataset to test the model's performance and predictions. It contained typicality ratings (out of 6) for 10 features for humans and 6 non-human animal categories.

Table 4: Typicality ratings (demo dataset)

	dog	dolphin	ant	cat	ox	fox	human
strong	4	4	5	1	6	2	3
happy	6	5	3	4	2	2	3
loyal	6	4	4	2	3	1	2
sly	0	0	0	5	0	6	4
smart	4	6	4	5	2	5	4
slow	0	0	1	2	5	0	2
scary	2	0	1	2	3	3	3
lazy	1	1	0	5	0	4	5
ferocious	2	0	0	1	3	4	2
graceful	1	4	0	6	1	6	3

7.2 Example model output

Here are a couple of examples of the model's predictions.

In the example below, 'lazy' has 0 typicality for ants (see 7.1.), therefore it has high salience. The example below is therefore in the specific, low salience condition and corresponds to the experimental data point "Is John lazy? John is an ant." Probability displayed is the maximum probability of an interpretation for this intention-utterance pair. The model predicts that the person intends to say that John is a human who is not at all lazy.

The person wanted to communicate the feature: LAZY
To do that, she uttered: ANT
The model's inference was that she meant HUMAN with the feature LAZY
Probability of the prediction: 0.0124
Typicality of the feature LAZY for ANT : 0.0

The example below is from the vague condition, so it is not clear to the listener what feature the speaker wanted to communicate. This corresponds to the experimental datapoint "What is John like? He is a dog." The model predicts the interpretation that John is happy since happy is the most salient (and typical) feature for a dog.

The person wanted to communicate the feature: UNIFORM
To do that, she uttered: DOG
The model's inference was that she meant HUMAN with the feature HAPPY
Probability of the prediction: 0.0094
Typicality of the feature HAPPY for DOG : 1.0

REFERENCES

[1] W. Klintsch and A.R. Bowles. Metaphor comprehension: What makes a metaphor difficult to understand? *Metaphor and symbol*, 17(4), 2002.

[2] J. Kao et al. Formalizing the pragmatics of metaphor understanding. *Proceedings of the annual meeting of the Cognitive Science Society*, 36, 2014.

[3] M.C. Frank and N. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084), 2012.

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

[5] Levy R. Goodman N. D Kao, J. T. The funny thing about incongruity: A computational model of humor in puns.

[6] C. et al. Graf. Animal, dog, or dalmatian? levels of abstraction in nominal referring expressions. *CogSci*, 2016.

[7] M.C. Frank. Rational speech act models of pragmatic reasoning in reference games. 2016.