Reasoning About the Things That Go Unsaid: Nonce Features in Reference Games

Evan Kim, Dolapo Adedokun

December 9, 2020

Abstract

When interpreting language, human listeners reason not only about the utterances they hear, but also about the things that go unsaid – that is, the alternative utterances. Some theories of alternatives propose that alternatives are derived from observed utterances through linguistic operations; Other theories propose that alternatives are derived at the level of concepts and not linguistic expressions. In our study, we use reference games to explore rates of scalar implicature when alternatives are conceptually but not linguistically accessible and build upon the Rational Speech Act (RSA) framework to model human reasoning and behaviour.

1 Introduction

Human language is incredible in many regards. Language allows us to express feelings, desires, and to connect with other humans through cognitive communication. Language is also remarkable in its efficiency; the ability for a speaker to convey specific meaning to a listener with only the necessary words and phrases. Specifically, a listener can often infer a speaker's intent by relying on informative assumptions and the context of the language itself – listeners are able to make communicative inferences by considering the intent of the speaker.

Consider, for example, the figure below, taken from Noah Goodman's and Michael C. Frank's "Pragmatic Language Interpretation as Probabilistic Inference" [1, 2]. Goodman's and Frank's work show that if a speaker refers to an object in the figure with the word **"blue"**, a listener would be likely to infer the speaker was referencing the leftmost object, because if they wanted to refer to the middle object, they might more appropriately use the alternative word **"circle"**.

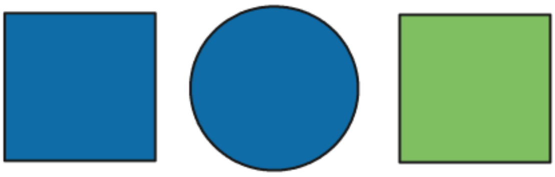

Figure 1: Communicative Inference

In our study, we explore what happens if the blue circle in the figure is replaced with a nonce shape or figure that does not have an easily accessible alternative descriptor. We explore how this change would affect a listeners rate of selecting the leftmost, blue square in comparison to the new nonce object we have introduced.

1.1 Background

Through phrasing, words and sounds, the human language enables communication through the use of discrete materials in order to generate an infinite number of sentences and sequences. As a result, the following interpretation or understanding of a particular phrase or sentence is just as infinite in its variability and further depends on a speaker's context and the previous dialogue or communication [1].

In our study though, we maintain a Gricean framework for pragmatic reasoning where we assume that speakers are cooperative and choose their words and phrases to convey a specific, distinctive, and informative meaning to a listener. Under this framework, from a speakers conveyance, a listener then works backwards to infer the speaker's intent and meaning [1, 2]. In our study, we assume language and communication are forms of rational action, and that the goal of both the listener and speaker are reciprocal in that they both work to maximize their utility; their mutual understanding of one another through efficient use of language.

Furthermore, we specifically assume that a listener presumes that a speaker is approximately rational; that they choose their "utterances in proportion to the utility they expect to gain." For example, a speaker might choose to produce a less meaningful or informative utterance when the alternative is costly or harder to articulate [1, 2].

2 Experiment Design

For this part of the project, Dolapo designed the experiment, experiment stimuli, and hypothesized about potential observations or findings we might discover in our data collection.

2.1 Variable Definitions

Unambiguous feature: A feature is unambiguous if only a single object has that feature within a set of objects
Ambiguous feature: A feature is ambiguous if multiple objects in the same set have this feature.
Nonce feature: A feature that is not easily described in the English lexicon.

2.2 Experiment Assumptions

We assume that the speaker (person asking the question) is rational and is well informed enough to communicate common knowledge. We are interested in looking at the listener's pragmatic understanding when the speaker produces a less informative utterance (because a more informative utterance is costly or more difficult to say).

Participants will be shown a set of objects in various conditions, with at least one object in each set containing an ambiguous feature. In some conditions, some objects within the set may also contain unambiguous or nonce features as well. Participants will then be asked "Which [object] would you choose if you heard [ambiguous feature]."

To enforce a uniform prior before the experiment, participants will be told the following:

"In the following scenarios, you will be given a set of objects and a speaker will ask you about which object is being referred to by a specific word. Select the object that you believe is being referred to based on the word that the speaker has chosen. You can assume all objects have an equal chance of being referred to beforehand."

2.3 Relevant Conditions

In this study, our main area of interest was to explore what happens to the probabilities of selecting different reference objects if we introduce a nonce object. To this end, we designed two main experiment conditions, two supplementary conditions and one filler trial.

2.4 Main Conditions

Typical Condition

In this condition, we have a standard reference game with a scalar implicature setup–one object with no feature, one object with an ambiguous feature, and one object with a unambiguous feature. The speaker's utterance would describe the ambiguous feature. An example of what this condition would look like for a participant is in Figure 2, below. The ambiguous feature would be the hat, and the non ambiguous feature would be the scarf.

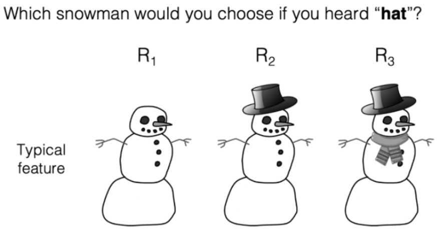

Figure 2: Nonce Condition 1

Nonce Condition 1

In this condition, we maintain the same structure as the standard reference game, but now the unambiguous feature is a nonce object. We have one object with no feature, one object with an ambiguous feature, and one object with a nonce unambiguous feature. The speaker's utterance would describe the ambiguous feature. An example of what this condition would look like for a participant is in Figure 3, below. The ambiguous feature would be the hat, and the non ambiguous feature would be the nonce object.

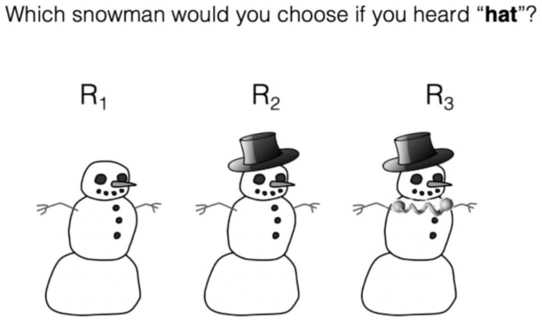

Figure 3: Nonce Condition 1

2.5 Supplementary Conditions

Nonce Condition 2

In this condition, we maintain a similar set up to the previous Nonce Condition, but now another object additionally has an unambiguous feature. We have one object with no feature, one object with an ambiguous feature, and one object with a nonce unambiguous feature. The speaker's utterance would describe the ambiguous feature. An example of what this condition would look like for a participant is in Figure 4 below. The ambiguous feature would be the hat, and the non ambiguous features would be the scarf and nonce object.

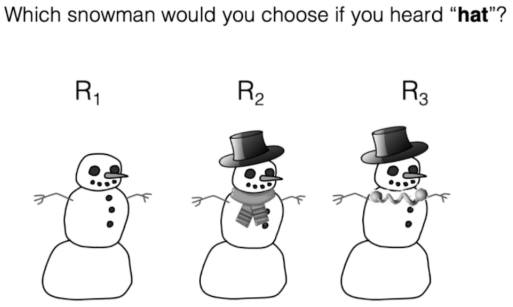

Figure 4: Nonce Condition 2

Nonce Condition 3

In this condition, we maintain a similar setup to Nonce Condition 1, but now another object additionally has a nonce feature. We have two objects with nonce features and one object with an ambiguous feature. The speaker's utterance would describe the ambiguous feature. An example of what this condition would look like for a participant is in figure 5 below. The ambiguous feature would be the hat, and the non ambiguous features would be the two nonce objects.

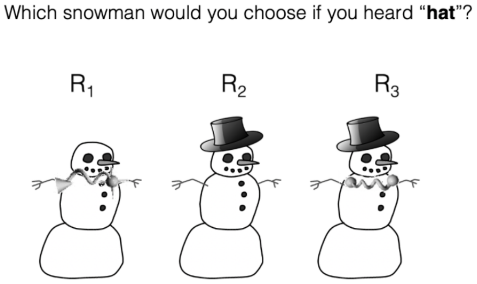

Figure 5: Nonce Condition 3

2.5.1 Filler Trial

We designed a filler trial in addition to the above conditions. The filler trial is meant to be a simple reference game with no ambiguity towards the object being referenced to. This is designed to weed out experiment participants who may not be paying full attention. The filler trial is attached in figure 6 below.

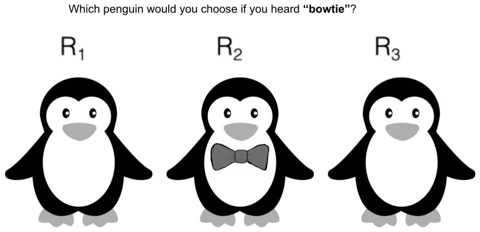

Figure 6: Filler Trial

3 Hypotheses

3.1 Typical Condition

For the typical condition, we expect the probability that a listener chooses the object with the ambiguous feature to be greater than the probability that a speaker chooses the object with the non ambiguous feature. Specifically, in reference to Figure 2, we expect $P(R_2|"hat") > P(R_3|"hat")$. We expect this hypothesis to hold true as this is a replication of what has been performed in previous studies.

3.2 Nonce Conditions

For the Nonce Conditions, we have two main hypotheses. Using Figure 3 as a reference:

3.2.1 H1: Alternative for R3 is more costly or harder to access

In this hypothesis, we predict that a listener will have a higher probability of selecting the object with only the ambiguous feature because the Nonce object is hard to access within the English lexicon. Specifically, we hypothesise $P_{Nonce}(R_2|"hat") - P_{Nonce}(R_3|"hat") < P_{Typical}(R_2|"hat") - P_{Typical}(R_3|"hat")$ for a listener.

3.2.2 H2: Nonce feature has high salience

In this second hypothesis, we hypothesize that the nonce feature on R_3 might actually be so salient such that the listener infers from the speaker that they intend to actually refer to the nonce object. Specifically, we hypothesize that for the listener $P_{Nonce}(R_2|"hat") >> P_{Nonce}(R_3|"hat")$ and that as a result, $P_{Nonce}(R_2|"hat") - P_{Nonce}(R_3|"hat") > P_{Typical}(R_2|"hat") - P_{Typical}(R_3|"hat")$.

4 Model

The Rational Speech Act (RSA) models [1, 2] created by Goodman and Frank have been used to make quantitative predictions about human behavior and explain common phenomena that have not been able to be modeled such as hyperbole and vagueness. When trying to understand how human speakers and listeners will react to features of objects that are conceptually but not linguistically accessible, we immediately turned to the Rational Speech Act to help model our experiment.

The Rational Speech Act framework consists of recursive and hierarchy based reasoning. For this part of the project, Evan implemented the Rational Speech Act model from scratch in Python. The Rational Speech Act is made up of the referents, semantics, prior probabilities, and costs. The referents are the subjects of the game and are the objects with the different features that the speaker will describe. The semantics is a map of messages to referents and is simply binary: it is activated if the message describes the referent and zero if it doesn't. The prior simply represents the likelihood of each referent occurring, and like it was noted above, our prior probability is split even at 50% since we give the disclaimer in the experiment set up. Lastly, the cost represents how “costly” or difficult it is to use a certain message. These features make up the Rational Speech Act model and can be best visualized as a grid with referents over the horizontal axis and messages over the vertical axis as seen in Table 1.

	r1	r2	Cost
”hat”	0	1	0
”glasses”	1	1	0
Prior	0.5	0.5	

Table 1: Sample grid of variables needed for RSA

This model is made of up the literal listener, pragmatic speaker, and pragmatic listener. The probability distribution of the literal listener affects the probability of the pragmatic speaker which in turn affects the pragmatic listener. This is the recursive and hierarchy based reasoning explained previously.

For the literal listener, this probability distribution takes into account the semantics and prior over referents and normalizes this value.

$$P_{Lit}(r|m) = \frac{[[m]](r)P(r)}{\Sigma_{r' \in R}[[m]](r')P(r')}$$

The pragmatic speaker exponentiates the log of the literal listener and incorporates the cost of that message. After this calculation is complete, normalization is performed.

$$P_S(m|r) = \frac{exp(\alpha(logP_{Lit}(r|m) + C(m)))}{\Sigma_{m' \in M} exp(\alpha(logP_{Lit}(r|m') + C(m')))}$$

Finally, the pragmatic listener again takes into consideration the prior on referents, but unlike the literal listener, uses the distribution it infers about the pragmatic speaker. Again, this value is normalized.

$$P_L(r|m) = \frac{P_S(r|m)P(r)}{\Sigma_{r' \in R} P_S(r'|m)P(r')}$$

To code this Rational Speech Act model in Python (Listing 1) we implemented the numpy package to handle matrix multiplication. Additionally, there was a section to manually pass in values and run the model. The model was verified and checked on multiple examples.

```
import numpy as np

class RSA:
    ”””
    Inputs:
    All returned prob distributions will be in the shape (m x r)
    with the referents as the columns and murmurs as the rows
    ”””
    def __init__(self, referents, murmurs, lexicon, prior, cost,
                 alpha=1):
        self.referents = referents
        self.murmurs = murmurs
        self.lexicon = lexicon
        self.prior = prior
        self.cost = cost
        self.alpha = alpha
        # added parameters

    def literal_listener(self):
        p_lit = np.zeros((self.murmurs.shape[0],
                          self.referents.shape[1]))
        for i in range(self.murmurs.shape[0]):
            margin = np.dot(self.prior, self.lexicon[i])
            p_lit[i] = np.divide(np.multiply(self.lexicon[i],
                                             self.prior), margin)
        return p_lit

    def pragmatic_speaker(self):
        p_lit = self.literal_listener()
        p_s = np.zeros((self.murmurs.shape[0],
                        self.referents.shape[1]))
        for j in range(self.referents.shape[1]):
            with np.errstate(divide='ignore '):
                log_p_lit =  np.log(p_lit[:,[j]])
            exp_col = np.exp(np.multiply(log_p_lit +
                                 self.cost, self.alpha))
            margin = np.sum(exp_col)
            if margin == 0:
                continue
            p_s[:,[j]] = np.divide(exp_col, margin)
        return p_s

    def pragmatic_listener(self):
        p_s = self.pragmatic_speaker()
        p_l = np.zeros((self.murmurs.shape[0],
                        self.referents.shape[1]))
        for i in range(self.murmurs.shape[0]):
            margin = np.dot(p_s[i].reshape(1, self.referents.shape[1]),
                                            self.prior.T)
            p_l[i] = np.divide(np.multiply(p_s[i], self.prior), margin)
        return p_l
```

Listing 1: Self-implemented Rational Speech Act Model in Python

	r1	r2	r3	Cost
”hat”	0	1	1	0
”glasses”	0	0	1	0
Prior	0.33	0.33	0.33	

Table 2: Grid of inputs for the typical condition

	r1	r2	r3
”hat”	0	0.75	0.75
”scarf”	0	0	1

Table 3: Grid of pragmatic listener for the typical condition

We will now observe our model in the different conditions of our experiment that we plan to run.

Typical Condition

In the typical condition, we are observing the standard reference game with a scalar implicature setup. This set up includes no nonce objects and simply one referent with no features, one with an unambiguous feature, and another with the shared feature. Consider the simple snowman example in Figure 4. This example has the inputs in Table 2.

After passing these values into the model we were prompted with the following results in Table 3.

The result from the Rational Speech Act model was consistent with our hypothesis: that would explain the second referent without the unambiguous feature with a higher weight. This is because if the speaker was trying to reference the third referent, they would use the message “scarf” as opposed to “hat”. Additionally, we see that “scarf” is the only message for the speaker to use to describe the third referent according to the model. We were expecting this because the Rational Speech Act is modeled to handle this specific case, and there has been much research done in this area.

Nonce Condition

The Rational Speech Act model has performed particularly well in many applications however, with the current Rational Speech Act model, we cannot capture the projected behavior of human listeners when presented with a nonce object. We first will introduce the nonce condition example again in Figure 3.

We can note this when looking at the semantics for this example, it is identical to the typical case, except the message is now a nonce message as opposed to “scarf” as seen in Table 4.

Since we are passing in the same values to our model, we will therefore get the same results. However, this is not consistent with any of our hypotheses we raised.

As we have discussed previously, our first hypothesis (H1) considered that the alternative for the

	r1	r2	r3	Cost
”hat”	0	1	1	0
nonce	0	0	1	0
Prior	0.33	0.33	0.33	

Table 4: Grid of inputs for the typical condition

	r1, null	r2, hat	r3, hat	r3, nonce
"hat"	0	1	1	0
nonce	0	0	0	1

Table 5: Feature based grid for the inputs of nonce condition

	r1, null	r2, hat	r3, hat	r3, nonce
"hat"	0	0.5	0.5	0
nonce	0	0	0	1

Table 6: Results from feature based inputs for the nonce condition

third referent is harder to access, so $P_{L\_Nonce}(R2|"hat") - P_{L\_Nonce}(R3|"hat") < P_{L\_Typical}(R2|"hat") - P_{L\_Typical}(R3|"hat)$. The second hypothesis we created (H2) looked at the opposite and noted that the nonce feature may be so salient such that $P_{S\_Nonce}("hat"|R3) << P_{S\_Nonce}(nonce|R3)$ which would mean the opposite of H1: $P_{L\_Nonce}(R2|"hat") - P_{L\_Nonce}(R3|"hat") > P_{L\_Typical}(R2|"hat") - P_{L\_Typical}(R3|"hat)$. Our current model does not capture either of these hypotheses. Therefore we must adjust the model to be consistent with these hypotheses.

When considering this challenge, we noted that this problem was very similar to a few-shot/zero-shot learning problem. After doing additional research, we could not find anything of significance that would assist us in adjusting our model.

One of the first changes we made was not to the model, but to the inputs we were passing into the model. We experimented with a new semantic – one that was feature-based (Table 5) as opposed to one that is referent-based.

Unfortunately, we did not get promising results as seen in Table 6 and decided to go down another approach.

Our consensus was to augment the pragmatic speaker since this is where the predicament of choosing a message first appears. By editing the pragmatic speaker, we will in turn modify the pragmatic listener since the pragmatic listener depends on the pragmatic speaker.

The first consideration was to include a linear variable in the pragmatic speaker which would represent how "hard" a message is to access. This plays into the idea of a nonce feature since the speaker would not know what word to use to describe this object. Doing so would give our model more flexibility to weigh different messages that the speaker was considering. This makes conceptual sense for H1: if a speaker is struggling to find a word to use for the nonce object, there is a higher chance that the speaker will just give up and use the ambiguous word. This would decrease the difference between the two referents in the nonce condition compared to the typical condition, satisfying H1.

For our model we updated the pragmatic speaker to the following where $H(m)$ is the hardness term which represents how hard it is to pick a certain word.:

$$P_S(r|m) = \frac{exp(\alpha(logP_{Lit}(r|m) + C(m) - H(m)))}{\Sigma_{m' \in M} exp(\alpha(logP_{Lit}(r|m') + C(m')))}$$

And we got the following results from the updated Rational Speech Act model as seen in Table 7 when we set $H(nonce) = 6$. These results are consistent with H1, as $P_{L\_Nonce}(R2|"hat") - P_{L\_Nonce}(R3|"hat") < P_{L\_Typical}(R2|"hat") - P_{L\_Typical}(R3|"hat)$ and our outputs show $0.00247 < 0.5$.

If the data we collect ends up satisfying H1, we can infer that human behavior treats nonce objects with a negative cost and believes that they are much harder to explain, therefore making

	r1	r2	r3
”hat”	0	0.501	0.498
nonce	0	0	1

Table 7: Output from the updated RSA model the includes ”hardness” of a word term

	r1	r2	r3
”hat”	0	1	1
nonce	0	0	1
Saliency	1	1	10
Prior	0.33	0.33	0.33

Table 8: Input to the updated RSA model the includes ”saliency” of a referent

the difference between the two referents with the ambiguous features less clear.

As noted, H2 demonstrates the opposite logic as H1. To handle this we took some inspiration from paper from Henry [3] and ultimately altered what the pragmatic speaker was conditioning on. Instead of just conditioning on the referent, we introduced the saliency of a given referent.

To do so, we introduced a new term which the probability distributions would condition on: saliency. Saliency would be like the hardness term but would represent somethings contextually different. Saliency would take into account how novel a certain feature is when describing a referent. For example, our nonce objects would have a very high saliency score, since the speaker is likely not familiar with the object. As noted in H2, the $P_{S\_Nonce}(”hat”|R3) << P_{S\_Nonce}(nonce|R3)$ and since the listener will infer that since the nonce object is so salient, it pushes us in a direction such that “hat” has a very high probability to describe referent 2 in the nonce condition.

To include this idea of saliency, we added a salient term ($S(r)$) that would be included in the literal listener and pragmatic speaker terms. However, since the Rational Speech Act is recursive, its effects will trickle down to the pragmatic listener. A large salience value will mean that the referent is fairly salient. One thing to note, is in our formula we will take the inverse of the saliency term and multiply it to the probability distribution, since salience will negatively impact the literal listener when reasoning about the speaker.

$$P_{Lit}(s, r|m) = \frac{[[m]](r)P(r)S(r)}{\Sigma_{r' \in R}[[m]](r')P(r')}$$

$$P_S(m|s, r) = \frac{exp(\alpha(logP_{Lit}(r|m) + C(m)))}{\Sigma_{m' \in M} exp(\alpha(logP_{Lit}(r|m') + C(m')))}$$

Now our input grid would look like the grid from Table 8.

The results that we received were promising (Table 9) as H2 was satisfied. $P_{L\_Nonce}(R2|”hat”) - P_{L\_Nonce}(R3|”hat”) > P_{L\_Typical}(R2|”hat”) - P_{L\_Typical}(R3|”hat)$ still holds as $0.983 > 0.5$.

	r1	r2	r3
”hat”	0	0.991	0.008
nonce	0	0	1

Table 9: Output from the updated RSA model the includes ”saliency” of a referent

These results were now consistent with our predictions for H2. This can make logical sense since the nonce feature will be so salient that it is the clear choice to describe the referent with that nonce feature. Therefore, using the ambiguous feature to describe the referent will highly suggest that it represents the referent without the nonce object.

If the human data collected satisfies the results from H2, our model notes that human behavior may consider the idea of how salient a feature is, and if it is so salient, we will polarize each of the possible messages we can use to describe referents with ambiguous features.

5 Discussion

5.1 Future Work

There is much future research and work that needs to be completed to gain valuable insight, however, our team was able to lay the framework so an experiment can be run and analyzed quickly.

The next step that will take the highest priority is the implementation of the experiment. After placing the model in the Psiturk framework, this experiment can be hosted on our TA's lab's VPS and thrown on Amazon Mechanical Turk to collect data. Amazon Mechanical Turk will allow us to get 9 participants per 24 hours and we will be able to run this experiment as many times as we see fit.

Once data has been obtained, the next step would be to analyze the data. It is critical that this data is filtered before analysis and some features that will need to be filtered out are participants that failed a certain number of filler tasks and those that have a different native language than English. This will help remove potential biases and inaccurate data. Once data has been filtered, the statistical analysis must be done on the data to uncover potential insights. As described above, various different utterances will be displayed for the different participants, so comparing the results between different trials will be particularly useful and interesting.

Lastly, the alpha parameter in the Rational Speech Act is often fit on the human data it is trying to model. With the data from the experiment readily accessible, the alpha parameter should be fit on this new data to represent this task at hand. In order to refute or support the hypothesizes that we generated, one will need to compare the human behavior with the results generated by our model. In the case of Nonce Condition 1 where we had two differing hypotheses, one will be able to see if human behavior behaves like the first augmented Rational Speech Act or second Rational Speech Act we proposed. If it behaves consistent with the first augmented Rational Speech Act, it can be explained that human listeners infer that the speakers see the nonce object as costly to explain. Perhaps it is because they believe the speaker cannot find a good word that would explain the nonce feature accurately. However, if the opposite is true, it can be seen that the listener believes the speaker is conditioning on the state of the world and the nonce object. This would be consistent with the nonce feature being so salient that a speaker would almost only refer to a nonce description when describing that particular object.

Comparing the collected data to our model, one will be able to uncover what aspects of our model are able to emulate human behavior in the nonce condition. On the other hand, one can observe and note the pitfalls of the model and areas of improvement. This will allow researchers to better understand our updated Rational Speech Act model and aspects that represent nonce objects.

In the future when data is collected and analyzed, a recursive process can be applied to update the augmented Rational Speech Act model we proposed. After collecting the human data from

Amazon Mechanical Turk and analyzing the collected data, one will be able to adjust the Rational Speech Act given an alpha value that is fitted to the collected data. Recursively observing the downfalls of the model compared to the collected data and making adjustments will allow the augmented model to behave much better and much more like human behavior.

Another necessary extension is to understand how "salient" a feature is and what an appropriate "saliency" value is to pass into the model for different features. Currently, we passed in arbitrary values for these inputs.

Once there is more understanding of the current task, it will be interesting to include an experiment that would explore the interpolation of nonce objects. If this is done, the data can be looked at another level, as we will have additional data for human behaviors on a spectrum of how “nonce” an object is. This could be an interesting extension that our new model should be able to handle.

5.2 Conclusion

Throughout this project, we were able to design an experiment, create stimuli, implement the Rational Speech Act by hand, and provide the framework such that the experiment could be run. To our knowledge, understanding how nonce features impact the Rational Speech Act has not yet been explored yet and our contributions will hopefully encourage additional work to be done to model human behavior in speech. We learned about different ways to update the Bayesian model by introducing linear terms and additional variables to condition on.

6 Acknowledgments

We would like to thank Jennifer Hu, our TA sponsor, for introducing us to this topic and project. Without her help, we would not have the necessary resources or technical knowledge to work on this project. We would also like to thank Professor Tenenbaum and the 9.66 course staff for their continued support and encouragement throughout this project and the course.

7 Code

All code used for this project can be accessed at https://github.com/evankim20/9.66-Final-Project.

References

[1] Frank, Michael C. “Rational Speech Act Models of Pragmatic Reasoning in Reference Games.” 2016, doi:10.31234/osf.io/f9y6b.

[2] Goodman, Noah D., and Michael C. Frank. “Pragmatic Language Interpretation as Probabilistic Inference.” Trends in Cognitive Sciences, Elsevier Current Trends, 28 Sept. 2016, www.sciencedirect.com/science/article/pii/S136466131630122X.

[3] Tessler, Michael Henry, et al. “Informational Goals, Sentence Structure, and Comparison Class Inference.” doi:10.31234/osf.io/n8eyj.

[4] Kao, J. T., et al. “Nonliteral Understanding of Number Words.” Proceedings of the National Academy of Sciences, vol. 111, no. 33, 2014, pp. 12002–12007., doi:10.1073/pnas.1407479111.

[5] One-Shot Learning with a Hierarchical Nonparametric Bayesian Model. (2012, June). Ruslan Salakhutdinov, Josh Tenenbaum, Antonio Torralba. http://proceedings.mlr.press/v27/salakhutdinov12a/salakhutdinov12a.pdf