Colour Me Uncertain: Representing Vagueness with Probabilistic Semantics

Anonymous ACL submission

Abstract

People successfully communicate in everyday situations using vague language. In particular, colour terms have no clear boundaries as to the ranges of colours they describe. We model people's reasoning process in a dyadic reference game using the Rational Speech Acts (RSA) framework and probabilistic semantics, and we find that the implementation of probabilistic semantics requires a modification from pure theory to perform well on real-world data. In addition, we explore approaches to handling target disagreements in reference games, an issue that is rarely discussed in the RSA literature.

1 Introduction

Colour terms are vague. There are no clear boundaries for what red, green, blue, or other colour words denote, causing uncertainty in their interpretations, and yet we are able to effectively communicate using colours in everyday situations.

To explain how we work with uncertainty, proponents of probabilistic semantics (Cooper et al., 2014; Sutton, 2015) consider vagueness to be intrinsic to language, where competent agents make graded judgements as to whether a predicate applies to a situation. This view of semantics allows us to model predicates with conditional probabilities: for example, given a colour patch (e.g.), to what degree would an agent believe that the term *"green"* is appropriate? Aside from probabilistic semantics, other approaches have also been proposed, which we group into two categories: distribution over thresholds and fuzzy truth-values. We discuss their differences in §2.1.

In this paper, we explore the real-world feasibility of modelling vagueness with probabilistic semantics using a colour game dataset in English by Monroe et al. (2017). The game displays three colours and requires the speaker to describe a target colour, which the listener attempts to guess. Monroe et al. apply the Rational Speech Acts (RSA) framework (Frank and Goodman, 2012) with neural listener and speaker models to find that pragmatic inference helps in disambiguating similar colours. We extend their work by replacing their literal listener model, which we argue gives results approximating to fuzzy truth-values, with ones that use probabilistic semantics, and present three main contributions.

First, modelling real-world data with probabilistic semantics requires an additional Gricean assumption that not all world states be false in a given context. Second, the RSA framework is sensitive to the performance of the neural listener and speaker models, with previously observed pragmatic effects diminished after better tuning. Third, we propose various ways to handle target disagreements in dyadic reference games, and find that the removal of disagreements significantly improves model performance on Monroe et al.'s dataset.

2 Background & Related Work

Prior work has employed the RSA framework to combine semantics and pragmatics in an effort to quantify vagueness (Lassiter and Goodman, 2015; Monroe et al., 2017; McDowell and Goodman, 2019). RSA formalises the theory of conversational implicatures (Grice, 1975) by modelling people iteratively reasoning about each other's actions to infer their intentions, and quantifies the interaction by defining explicit objectives for listener and speaker agents. For a survey, see: Degen (2023).

For a given context, we model a literal listener l_0 choosing a state c based on an utterance u's literal interpretation, $\mathcal{L}(u, c)$, and weighted by its prior $P(c)$ (Equation 1). A pragmatic speaker s_1 then chooses an utterance that is most informative by considering the literal listener's choices, subject to a rationality parameter α and utterance cost $\kappa(u)$ (Equation 2). Finally, a pragmatic listener l_2 infers the intended state based on the speaker's choice of utterance (Equation 3).

$$l_0(c \mid u; \mathcal{L}) \propto \mathcal{L}(u, c)P(c) \quad (1)$$

$$s_1(u \mid c, \mathcal{L}) \propto e^{\alpha \log(l_0(c|u;\mathcal{L})) - \kappa(u)} \quad (2)$$

$$l_2(c \mid u, \mathcal{L}) \propto s_1(u \mid c, \mathcal{L})P(c) \quad (3)$$

In Monroe et al.'s game, the states are equally likely so the prior can be discounted. For simplicity, we assume $\kappa = 0$ and $\alpha = 1$.

2.1 Linguistic Approaches to Vagueness

Many approaches to modelling vagueness have been proposed (for a recent survey, see: Burnett and Sutton, 2020). Of particular interest are fuzzy and probabilistic approaches, because of their compatibility with neural network models.

In **fuzzy** logic, truth is not binary, but instead any real value from 0 to 1, which allows a direct account of vagueness (Zadeh, 1965). Logical operations such as AND and OR have fuzzy versions which are truth-functional, meaning that they are defined as functions taking fuzzy truth-values as input. The simplicity of a truth-functional approach means that fuzzy logic is unable to express correlations between truth-values (Fine, 1975). For example, considering a borderline red/orange shade, where "red" and "orange" are both 0.5 true, fuzzy logic treats "red or orange" the same as "red or not red". This does not match empirical facts about the use of vague terms (Sauerland, 2011).

In **probabilistic** logic, truth is binary but uncertain, and this can also be used to account for vagueness (Edgington, 1992, 1997). In contrast to fuzzy logic, there can be correlations between truth-values, which avoids the problems with the fuzzy account. However, this requires us to define a joint distribution over all truth-values.

To build up to a joint distribution, we first consider marginal probabilities. For a predicate u, we can define a probabilistic truth-conditional function that gives the probability of the truth-value T_c being true, for state c, as in Equation 4. This function gives the marginal probability for one truth-value, ignoring all other truth-values (for other states c').

$$t_u(c) = \mathbb{P}(T_c = \top; u) \quad (4)$$

A simple approach to define a joint distribution is to define a global threshold for truth, uniformly sampled from $[0, 1]$, against which marginal probabilities of truth are compared. Combining this with the RSA framework can capture various aspects of how vague terms are used (Lassiter, 2011; Lassiter and Goodman, 2015).

However, using a global threshold is restrictive. Emerson (2023) shows how we can see such a model as one instance in a broader class of probabilistic models. The most general model class would consider all possible joint distributions, but this is intractable. Tractability can be maintained by restricting to models that only require: the marginal probability for each truth-value, and the correlation between each pair of truth-values. A global threshold corresponds to maximising all correlations.

3 Methodology

We adopt the model architectures in Monroe et al., with a few refinements, to train an RSA system on the colour game dataset. As in Andreas and Klein (2016), neural models enable listener and speaker agents to be trained on real-world language use. The literal listener uses an LSTM to process utterances and based on its final state it outputs parameters for a score function. The literal speaker generates utterances by encoding the colour context as input to a second LSTM.

We refine Monroe et al.'s model by switching the speaker's decoding process from sampling to beam search, as well as making the colour encoder permutation invariant to the order of inputs (Zaheer et al., 2017), so as to improve performance.

The literal listener's score function is given in Equation 5, where f is the Fourier-transformed vector representation of a colour (a deterministic transformation, following Monroe et al., 2016), and μ and Σ are output by the LSTM.

$$\text{score}(f) = -(f - \mu)^T \Sigma (f - \mu) \quad (5)$$

If Σ is positive definite, which Monroe et al. note is the case for over 95% of their inputs, the score is the logarithm of a probability density function (a multivariate Gaussian).

3.1 Base Literal Listener Model

Our baseline model follows Monroe et al. (2017), normalising the scores with an exponential softmax to give the listener's beliefs about the intended colour. Viewing this under the approaches in §2.1, it can be seen as implementing fuzzy logic, since the exponential of the score is a fuzzy truth-value and normalising fuzzy truth-values is a truth-functional operation (for details, see Appendix A).

As this interpretation only holds if Σ is positive definite, we include a model in our experiments where scores are clamped to be non-positive so that it can be clearly contrasted with other approaches.

3.2 Probabilistic Literal Listener Model L_0^{prob}

Instead of normalising the scores directly, our L_0^{prob} probabilistic literal listener model interprets them as log-probabilities of truth. We clamp the scores to be non-positive and take their exponentials to get marginal probabilities $t_u(c)$ for each colour c.

These marginals are then used to calculate the joint distribution. Given three colours in the context, there are $2^3 = 8$ possible joint outcomes for truth-values. The joint distribution is not fully determined by the marginals, but also depends on correlations between the truth-values. We assume correlations are fixed (see Emerson, 2023 for more options), and explore two possibilities: 1. truth-values are independent (*Prob Indep*), and 2. truth-values are maximally correlated (*Prob Max*).

Finally, the joint distribution over truth-values determines the distribution over listener actions. If ties are randomly broken (u is true for more than one colour, or false for all colours), then the chance of picking the target colour is given in Equation 6, where $p_{...}$ is the joint probability of truth ($\top$) or falsehood ($\bot$) for each colour.

$$L_0^{\text{prob}}(c_0 \mid u, C; \theta) = p_{\top\bot\bot} + \frac{1}{2}p_{\top\top\bot} + \frac{1}{2}p_{\top\bot\top} + \frac{1}{3}p_{\top\top\top} + \frac{1}{3}p_{\bot\bot\bot} \quad (6)$$

However, we notice a problem with training a model to maximise the "pure" probabilistic objective in Equation 6. Suppose an utterance is definitely false for some colour. In the case where all truth-values are false, the "definitely false" colour is chosen with a one-third chance. The only way for the model to avoid this outcome is to set the marginal probability of another colour to 1, but by doing so it cannot convey uncertainty.

To avoid this problem, we introduce an "applied" version of the model, where the all-false outcome is excluded. In other words, if the speaker makes an utterance, it must be true of something, which is grounded on Grice's maxim of quality.

3.3 Target Disagreements

In supervised learning, it is assumed there is an objectively correct output for each input. This assumption does not hold for our language reference game. While there is a correct answer in the context of the game (i.e. the target colour), the listener and the speaker's choices cannot be wrong given our objective of modelling linguistic behaviour. From the speaker's perspective, the utterance they uttered applies to the target colour; from the listener's perspective, the colour they chose best matches the utterance they received. As such, we propose and investigate three alternative strategies for modelling data with target disagreements:

Listener-Speaker (L-S): Train on the listener's choice but evaluate on the speaker's target. The aim is for the literal listener to emulate a human listener's literal interpretation function, and for the pragmatic listener to apply pragmatic reasoning to select the intended target.

Listener-Listener (L-L): Both train and evaluate on the listener's choice. This changes the objective to emulating listener behaviour rather than selecting the "correct" target.

No Disagreements (ND): Remove training data with disagreements between speaker and listener, but evaluate on the unaltered test set. The aim is to understand if disagreements add noise to training.

Model	L_0 Accuracy	L_2 Accuracy
Monroe et al. (2017)	85.08	86.98[1]
Base	**87.65 ± 0.05**	88.03 ± 0.04
Base Clamped	87.51 ± 0.05	87.94 ± 0.04
Pure Prob Indep	76.06 ± 0.07	76.98 ± 0.12
Pure Prob Max	75.84 ± 0.08	76.85 ± 0.11
Applied Prob Indep	87.65 ± 0.03	87.96 ± 0.05
Applied Prob Max	87.58 ± 0.04	**88.05 ± 0.06**

Table 1: Mean accuracies for the main models evaluated on the test set, shown with standard errors of the means. Highest accuracy for each category in bold.

3.4 Experiment Setup

Hyperparameters were determined with grid search on the validation set, using the original data split. Details of grid search and chosen hyperparameters are given in Appendix B. Every model type was trained 10 times to reduce the effect of random initialisation (Reimers and Gurevych, 2017). Since an RSA model contains two neural nets (listener and speaker), they were arbitrarily paired up and the same dyads used for all evaluations.

4 Results & Discussion

The accuracies of the main model types are summarised in Table 1. Two-tailed p-values were above 0.1 between all pairs of the Base and Applied Prob

[1]This is for Monroe et al.'s best performing blended model, L_e, as they did not report L_2 accuracy on the test set.

Model	Far	Split	Close
Pure Prob Indep	93.00	75.04	62.76
Applied Prob Indep	96.25	87.76	79.78
Δ (Applied - Pure)	3.25	12.72	17.02

Table 2: Comparison of the mean accuracies between the pure and applied probabilistic (Independent) models across different context types. Similar results were obtained using the Max Correlation models.

Model	$t_u(c) < 0.01$	$t_u(c) > 0.99$
Base Clamped	94.05%	3.98%
Pure Prob Indep	5.61%	89.22%
Applied Prob Indep	56.93%	7.91%

Table 3: Percentage of target colour samples that were assigned extreme marginal probabilities $t_u(c)$.

Train-Test Target	L_0 Accuracy	L_2 Accuracy
S-S	87.65 ± 0.03	87.96 ± 0.05
L-S	86.32 ± 0.04	86.70 ± 0.05
L-L	85.02 ± 0.04	85.14 ± 0.04
S-S ND	**87.85 ± 0.04**	**88.18 ± 0.06**

Table 4: Mean accuracies for the probabilistic (independent) models, using the specified target disagreement strategy, shown with standard errors of the means. Highest accuracy for each category in bold.

models,[2] so there is no evidence to suggest a performance difference between these four model types.

Although the Base listener uses Monroe et al.'s architecture, its accuracy is much higher, highlighting the impact of model tuning and hyperparameter selection. The best optimisation algorithm found in grid search, AdamW, was not available at the time their work was published. Also, they did not state if their models were regularised, but we found a dropout rate of 0.5 provided the best performance. The narrower gap between our L_0 and L_2 accuracies suggests that some of the improvements from pragmatic reasoning that Monroe et al. observed could be attributed to an under-tuned model.

In addition, we find that the Base model produces positive scores for over 36% of the test set, compared to less than 5% noted by Monroe et al.. For the Base Clamped model, this drops to 3.1% for the raw scores before clamping, demonstrating that training dynamics affect the interpretation of the model as producing fuzzy truth-values.

4.1 Pure vs Applied Probabilistic Models

The performance differences between corresponding Pure and Applied models are significant at p<0.00001. The limitation of the Pure models is apparent when comparing different difficulty contexts in Table 2. For the Pure models, the especially poor results in contexts with two or more similar colours (*split* and *close*) can be attributed to the high marginal probabilities generated, as shown in Table 3 (for full distributions, see Appendix C). If two or more colours in a given context have high marginal probability, the literal listener's output distribution will be skewed towards having equal probabilities for those colours, drowning out any signal from the utterance. In contrast, the Applied models produce less extreme marginal probabilities and achieve better performance in all context types.

4.2 Target Disagreements

The results of our proposed strategies to deal with target disagreements are shown in Table 4. The models trained on listener choices performed poorer not only in predicting speaker targets, but also in predicting listener choices. However, the removal of target disagreements from training resulted in significantly better performance than the S-S models trained on the full dataset.[3] This suggests that the data samples with target disagreements added noise during the training process, leading to poorer performance.

5 Conclusion

We demonstrated that a probabilistic semantic model benefits from an assumption to exclude an all-false outcome. While our results do not conclusively decide between probabilistic or fuzzy approaches to vagueness, this paper adds to a growing body of work that people exhibit pragmatic behaviours as posited by the RSA framework. However, careful tuning of the literal listener model reduces the effect size of pragmatic reasoning compared to previous work. Finally, we explored the previously undiscussed issue of target disagreements. For the 'Colors in Context' dataset, we found that disagreements may be best seen as noise.

[2] Bootstrap tests using 100,000 rounds of resampling were performed over the six pairs of these four model types.

[3] Two-tailed p-value of 0.0296 for the Prob Indep models in Table 4. Results for other model types are similar; for details see Appendix D.

Limitations

As our work focuses on one dataset, we are not able to generalise about the effectiveness of our proposed strategies to handle target disagreements on other dyadic reference games. We have given a theoretical justification and empirical analysis of our results, and so we would expect our conclusions to generalise, but further work would be needed to confirm this on other datasets. In addition, we applied fixed global correlations between truth-values when exploring the probabilistic approach, and leave for future work to investigate the impact of varying correlations locally.

References

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Heather Burnett and Peter R. Sutton. 2020. Vagueness and natural language semantics. In D. Gutzmann, L. Matthewson, C. Meier, H. Rullmann, and T. Zimmermann, editors, *The Wiley Blackwell Companion to Semantics*. Wiley.

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2014. A probabilistic rich type theory for semantic interpretation. In *Proceedings of the EACL 2014 Workshop on Type Theory and Natural Language Semantics (TTNLS)*, pages 72–79, Gothenburg, Sweden. Association for Computational Linguistics.

Judith Degen. 2023. The Rational Speech Act framework. *Annual Review of Linguistics*, 9:519–540.

Timothy Dozat. 2016. Incorporating Nesterov momentum into Adam. In *Proceedings of the 4th International Conference on Learning Representations*.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159.

Dorothy Edgington. 1992. Validity, uncertainty and vagueness. *Analysis*, 52(4):193–204.

Dorothy Edgington. 1997. Vagueness by degrees. In Rosanna Keefe and Peter Smith, editors, *Vagueness: A Reader*. MIT Press.

Guy Emerson. 2023. Probabilistic lexical semantics: From Gaussian embeddings to Bernoulli Fields. In Jean-Philippe Bernardy, Rasmus Blanck, Stergios Chatzikyriakidis, Shalom Lappin, and Aleksandre Maskharashvili, editors, *Probabilistic Approaches to Linguistic Theory*, chapter 3, pages 65–121. University of Chicago Press.

Kit Fine. 1975. Vagueness, truth and logic. *Synthese*, 30(3/4):265–300.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

H. Paul Grice. 1975. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Daniel Lassiter. 2011. Vagueness as probabilistic linguistic knowledge. In *Vagueness in Communication*, pages 127–150, Berlin, Heidelberg. Springer Berlin Heidelberg.

Daniel Lassiter and Noah Goodman. 2015. Adjectival vagueness in a bayesian model of interpretation. *Synthese*, 194:3801–3836.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*.

Bill McDowell and Noah Goodman. 2019. Learning from omission. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 619–628, Florence, Italy. Association for Computational Linguistics.

Will Monroe, Noah D. Goodman, and Christopher Potts. 2016. Learning to generate compositional color descriptions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2243–2248, Austin, Texas. Association for Computational Linguistics.

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.

Uli Sauerland. 2011. Vagueness in language: The case against fuzzy logic revisited. In Petr Cintula, Christian G. Fermüller, Lluís Godo, and Petr Hájek, editors, *Understanding Vagueness: Logical, Philosophical and Linguistic Perspectives*, pages 185–198. College Publications.

Peter R. Sutton. 2015. Towards a probabilistic semantics for vague adjectives. In Henk Zeevat and Hans-Christian Schmitz, editors, *Bayesian Natural Language Semantics and Pragmatics*, pages 221–246. Springer International Publishing, Cham.

L.A. Zadeh. 1965. Fuzzy sets. *Information and Control*, 8(3):338–353.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R. Salakhutdinov, and Alexander J. Smola. 2017. Deep sets. *Advances in Neural Information Processing Systems*, 30.

Matthew D. Zeiler. 2012. Adadelta: An adaptive learning rate method. ArXiv preprint 1212.5701.

A Proof of Fuzzy Logic in the Base Literal Listener Model

The score function in Equation 5 is repeated below as Equation 7. For a given utterance u, the base literal listener determines μ and Σ, then applies this score function to each colour representation f. The scores are passed through an exponential softmax to give a probability distribution over the colours.

$$\text{score}(f) = -(f-\mu)^T \Sigma (f-\mu) \tag{7}$$

Given representations f_i for a set of colours $C = \{c_0, \ldots, c_n\}$, the probability of choosing each colour is therefore given by:

$$L_0^{\text{base}}(c_i|u, C; (L)) = \frac{\exp(\text{score}(f_i))}{\sum_j \exp(\text{score}(f_j))} \tag{8}$$

To define a Gaussian distribution, as suggested by Monroe et al., the exp-scores must be rescaled so that they integrate to 1. However, multiplying all exp-scores by a constant leaves the distribution in Equation 8 unchanged, and so does not change any predictions of the model.

If Σ is positive definite, the score function achieves its maximum value of 0 when $f = \mu$. The exp-scores are therefore guaranteed to lie in the range $[0, 1]$, and so can be interpreted as fuzzy truth-values for the utterance u. The distribution in Equation 8 is therefore a normalisation of these fuzzy truth-values. The normalisation only depends on the truth-values (with no further dependence on u or f_i), and so it is a truth-functional operation.

B Grid Search and Hyperparameters

We performed grid search to identify the most performant optimisation algorithms, learning rates, and dropout values for training the neural listener and speaker models. Five optimisation algorithms were explored in the grid search process: Adam (Kingma and Ba, 2015), AdamW (Loshchilov and Hutter, 2019), NAdam (Dozat, 2016), Adadelta (Zeiler, 2012), and Adagrad (Duchi et al., 2011). The Adam and Adadelta algorithms were chosen because they were used in Monroe et al. (2017), while the other three were selected as alternative adaptive optimisation algorithms. For the learning rates, values ranging from 1 to 10^{-4} were selected at regular logarithmic intervals, and dropout rates ranging from 0 to 0.5 were selected at intervals of 0.1.

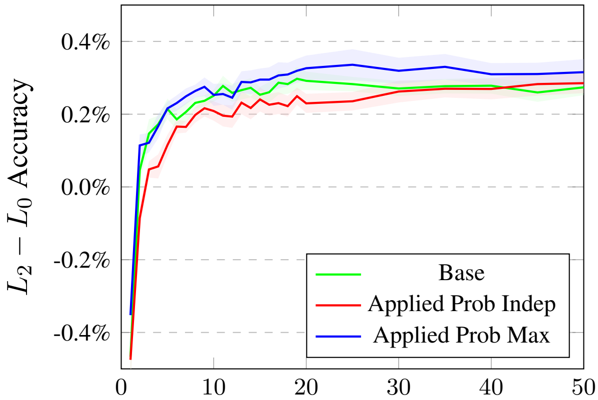

Figure 1: Mean deltas between L_2 accuracy and L_0 accuracy on the validation set, with varying numbers of alternative utterances per colour. Shaded regions mark the standard errors of the means. Number of utterances were incremented by 1 between 1 and 20 utterances, and incremented by 5 between 20 and 50 utterances.

Based on the results from grid search, we trained the listener models with AdamW using a learning rate of 0.001 and 0.0004 for the base and probabilistic models respectively, and the speaker model with Adam using a learning rate of 0.001. Dropout of 0.5 was applied to listener models, but not to the speaker models as their performance degraded significantly with any dropout. The neural models used the same embedding and hidden dimension sizes as in Monroe et al. (2017), which was 100.

We varied the beam size in the literal speaker's decoding process to analyse the impact on the pragmatic listener's performance. Since the literal speaker produces alternative utterances as a proxy for the set of all possible utterances that theoretical pragmatic agents would consider, we conjectured that generating a larger number of utterances should improve pragmatic performance. As seen in Figure 1, the pragmatic effect increases until around 15 to 20 utterances per colour before

Model	L_0 Accuracy	L_2 Accuracy
Base: Speaker-Speaker	87.65 ± 0.05	88.03 ± 0.04
Base: Listener-Speaker	86.29 ± 0.04	86.74 ± 0.05
Base: Listener-Listener	84.97 ± 0.04	85.16 ± 0.04
Base: Speaker-Speaker, No Disagreements	**87.98 ± 0.04**	**88.27 ± 0.04**
Applied Prob Independent: Speaker-Speaker	87.65 ± 0.03	87.96 ± 0.05
Applied Prob Independent: Listener-Speaker	86.32 ± 0.04	86.70 ± 0.05
Applied Prob Independent: Listener-Listener	85.02 ± 0.04	85.14 ± 0.04
Applied Prob Independent: Speaker-Speaker, No Disagreements	**87.85 ± 0.04**	**88.18 ± 0.06**
Applied Prob Max Correlation: Speaker-Speaker	87.58 ± 0.04	88.05 ± 0.06
Applied Prob Max Correlation: Listener-Speaker	86.15 ± 0.06	86.66 ± 0.07
Applied Prob Max Correlation: Listener-Listener	84.90 ± 0.05	85.11 ± 0.05
Applied Prob Max Correlation: Speaker-Speaker, No Disagreements	**87.88 ± 0.04**	**88.20 ± 0.05**

Table 5: Mean accuracies for the base and applied probabilistic models, using the specified target disagreement strategy, shown with standard errors of the means. Highest accuracy for each category in bold.

plateauing, so we chose a beam size of 15 to maintain the trade-off between computation time and performance.

For the grid search process, analysis of alternative utterances, and model checkpointing, accuracy was evaluated using the validation set based on the train/validation/test data split that Monroe et al. created.

C Full Distribution of Marginal Probabilities

Illustrations of the full distributions of marginal probabilities produced by the literal listener models are shown in Figure 2, as opposed to the summary statistics given in Table 3.

D Target Disagreements – Full Results

Table 5 lists the full results of various target disagreement strategies for each model type. Compared against Table 4, we see the same trends where the No Disagreements strategy performed the best, followed by Speaker-Speaker, Listener-Speaker, and lastly the Listener-Listener strategy.

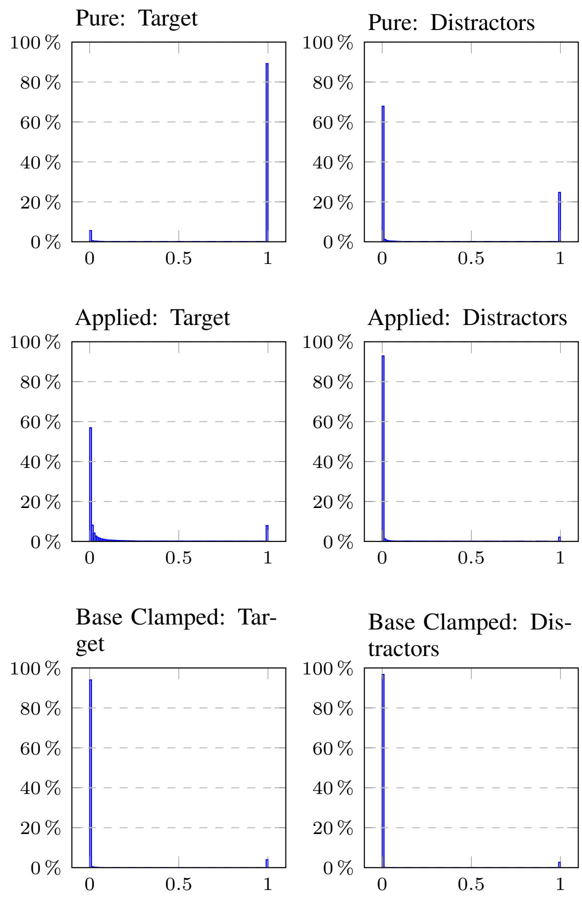

Figure 2: Distribution of marginal probabilities produced by literal listener models for the target and distractor colours in the test set.