translation), we focus on developing automatic measures of system performance. We use the available training data to develop simulated models of human decisions; by first showing that these models track well with human judgments, we can be confident that their use in evaluations will correlate with human understanding. We employ the following two metrics:

Belief evaluation This evaluation focuses on the denotational perspective in semantics that motivated the initial development of our model. We have successfully understood the semantics of a message z_r if, after translating $z_r \mapsto z_h$, a human listener can form a correct belief about the state in which z_r was produced. We construct a simple state-guessing game where the listener is presented with a translated message and two state observations, and must guess which state the speaker was in when the message was emitted.

When translating from natural language to neuralese, we use the learned agent model to directly guess the hidden state. For neuralese to natural language we must first construct a "model human listener" to map from strings back to state representations; we do this by using the training data to fit a simple regression model that scores (state, sentence) pairs using a bag-of-words sentence representation. We find that our "model human" matches the judgments of real humans 83% of the time on the colors task, 77% of the time on the birds task, and 77% of the time on the driving task. This gives us confidence that the model human gives a reasonably accurate proxy for human interpretation.

Behavior evaluation This evaluation focuses on the cooperative aspects of interpretability: we measure the extent to which learned models are able to interoperate with each other by way of a translation layer. In the case of reference games, the goal of this semantic evaluation is identical to the goal of the game itself (to identify the hidden state of the speaker), so we perform this additional pragmatic evaluation only for the driving game. We found that the most data-efficient and reliable way to make use of human game traces was to construct a "deaf" model human. The evaluation selects a full game trace from a human player, and replays both the human's actions and messages exactly (disregarding any incoming messages); the evaluation measures the quality of the natural-language-to-neuralese translator, and the extent to which the learned agent model can accommodate a (real) human given translations of the human's messages.

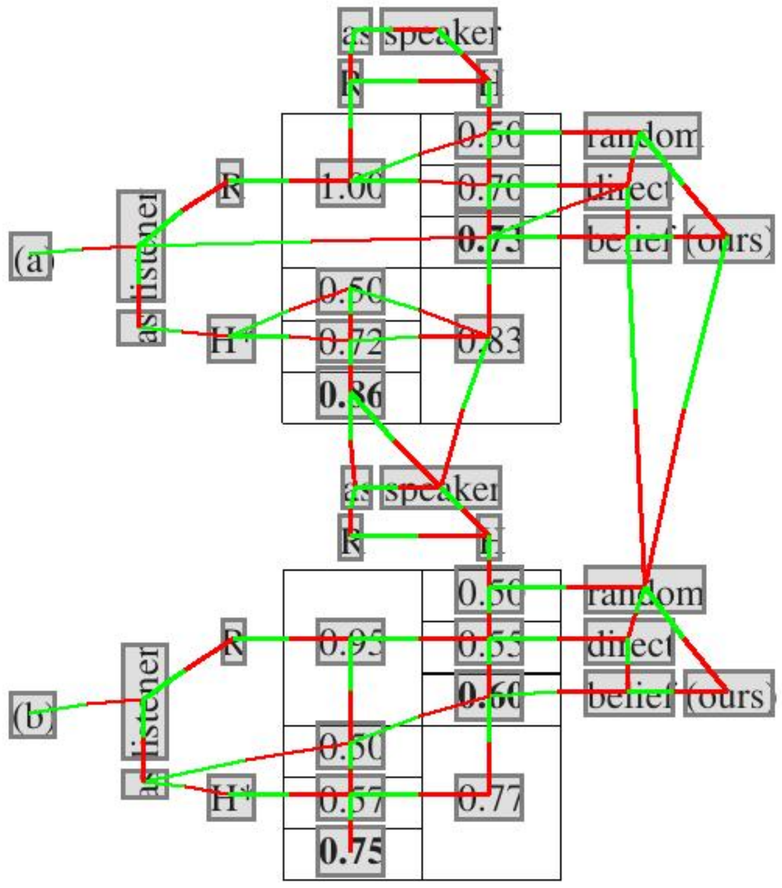

Table 1: Evaluation results for reference games. (a) The colors task. (b) The birds task. Whether the model human is in a listener or speaker role, translation based on belief matching outperforms both random and machine translation baselines.

Baselines We compare our approach to two baselines: a *random* baseline that chooses a translation of each input uniformly from messages observed during training, and a *direct* baseline that directly maximizes $p(z'|z)$ (by analogy to a conventional machine translation system). This is accomplished by sampling from a DCP speaker in training states labeled with natural language strings.

8 Results

In all below, "R" indicates a DCP agent, "H" indicates a real human, and "H*" indicates a model human player.

Reference games Results for the two reference games are shown in Table 1. The end-to-end trained model achieves nearly perfect accuracy in both

Figure 7: Best-scoring translations generated for color task.