A Proofs

Proof of Proposition 1 We know that

$$U(z, x_b) := \arg\max_u \sum_{x_a} p(x_a|x_b, z) r(x_a, x_b, z)$$

and that for all translations $(z, z' = t(r))$

$$D \geq \sum_{x_b} p(x_b|z, z') \mathcal{D}_{\text{KL}}(\beta(z, x_b) \mid\mid \beta(z', x_b)) \ .$$

Applying Pinsker's inequality:

$$\geq 2 \sum_{x_b} p(x_b|z, z') \delta(\beta(z, x_b), \beta(z', x_b))^2$$

and Jensen's inequality:

$$\geq 2 \left(\sum_{x_b} p(x_b|z, z') \delta(\beta(z, x_b), \beta(z', x_b)) \right)^2$$

so

$$\sqrt{D/2} \geq \sum_{x_b} p(x_b|z, z') \delta(\beta(z, x_b), \beta(z', x_b)) \ .$$

The next step relies on the following well-known property of the total variation distance: for distributions p and q and a function f bounded by $[0, 1]$,

$$|\mathbb{E}_p f(x) - \mathbb{E}_q f(x)| \leq \delta(p, q) \ . \qquad (*)$$

For convenience we will write

$$\delta := \delta(\beta(z, x_b), \beta(z', x_b)) \ .$$

A listener using the speaker's language expects a reward of

$$\sum_{x_b} p(x_b) \sum_{x_a} p(x_a|x_b, z) r(x_a, x_b, U(z, x_b))$$
$$\leq \sum_{x_b} p(x_b) \left(\sum_{x_a} p(x_a|x_b, z') r(x_a, x_b, U(z, x_b)) + \delta \right)$$

via (*). From the assumption of player rationality:

$$\leq \sum_{x_b} p(x_b) \left(\sum_{x_a} p(x_a|x_b, z') r(x_a, x_b, U(z', x_b)) + \delta \right)$$

using (*) again:

$$\leq \sum_{x_b} p(x_b) \left(\sum_{x_a} p(x_a|x_b, z) r(x_a, x_b, U(z', x_b)) + 2\delta \right)$$
$$\leq \sum_{x_a, x_b} p(x_a, x_b|z) r(x_a, x_b, U(z', x_b)) + \sqrt{2D} \ .$$

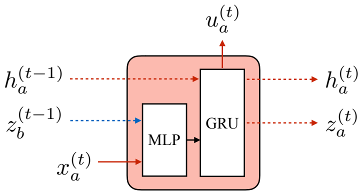
So the true reward achieved by a z'-speaker receiving a translated code is only additively worse than the native z-speaker reward:

$$\left(\sum_{x_a, x_b} p(x_a, x_b|z) r(x_a, x_b, U(z, x_b)) \right) - \sqrt{2D} \qquad \square$$

B Implementation details

B.1 Agents

Learned agents have the following form:



where h is a hidden state, z is a message from the other agent, u is a distribution over actions, and x is an observation of the world. A single hidden layer with 256 units and a $\tanh$ nonlinearity is used for the MLP. The GRU hidden state is also of size 256, and the message vector is of size 64.

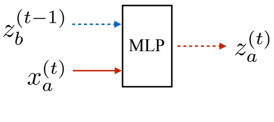
Agents are trained via interaction with the world as in Hausknecht and Stone (2015) using the ADAM optimizer (Kingma and Ba, 2014) and a discount factor of 0.9. The step size was chosen as 0.003 for reference games and 0.0003 for the driving game. An ϵ-greedy exploration strategy is employed, with the exploration parameter for timestep t given by:

$$\epsilon = \max \begin{cases} (1000 - t)/1000 \\ (5000 - t)/50000 \\ 0 \end{cases}$$

As in Foerster et al. (2016), we found it useful to add noise to the communication channel: in this case, isotropic Gaussian noise with mean 0 and standard deviation 0.3. This also helps smooth $p(z|x_a)$ when computing the translation criterion.

B.2 Representational models

As discussed in Section 5, the translation criterion is computed based on the quantity $p(z|x)$. The policy representation above actually defines a distribution $p(z|x, h)$, additionally involving the agent's hidden state h from a previous timestep. While in principle it is possible to eliminate the dependence on h by introducing an additional sampling step into Algorithm 1, we found that it simplified inference to simply learn an additional model of $p(z|x)$ directly. This model is trained alongside the learned agent to imitate its decisions, but does not get to observe the recurrent state, like so:



Here the multilayer perceptron has a single hidden layer with $\tanh$ nonlinearities and size 128. It is also trained with ADAM and a step size of 0.0003.

We use exactly the same model and parameters to implement representations of $p(z|x)$ for human speakers, but in this case the vector z is taken to be a distribution over messages in the natural language inventory, and the model is trained to maximize the likelihood of labeled human traces.

B.3 Tasks

Colors We use the version of the XKCD dataset prepared by McMahan and Stone (2015). Here the input feature vector is simply the LAB representation of each color, and the message inventory taken to be all unigrams that appear at least five times.

Birds We use the dataset of Welinder et al. (2010) with natural language annotations from Reed et al. (2016). The model's input feature representations are a final 256-dimensional hidden feature vector from a compact bilinear pooling model (Gao et al., 2016) pre-trained for classification. The message inventory consists of the 50 most frequent bigrams to appear in natural language descriptions; example human traces are generated by for every frequent (bigram, image) pair in the dataset.

Driving Driving data is collected from pairs of human workers on Mechanical Turk. Workers received the following description of the task:

> Your goal is to drive the red car onto the red square. Be careful! You're driving in a thick fog, and there is another car on the road that you cannot see. However, you can talk to the other driver to make sure you both reach your destinations safely.

Players were restricted to messages of 1–3 words, and required to send at least one message per game. Each player was paid $0.25 per game. 382 games were collected with 5 different road layouts, each represented as an 8x8 grid presented to players as in Figure 8. The action space is discrete: players can move forward, back, turn left, turn right, or wait. These were divided into a 282-game training set and 100-game test set. The message inventory consists of all messages sent more than 3 times. Input features consists of indicators on the agent's current position and orientation, goal position, and map identity. Data is available for download at `http://github.com/jacobandreas/neuralese`.