One of the main challenges of creating a model of this size is the time required to train it. Even using the cloud computing resources it took us several hours to train for a few hundred epochs. Because of this we wanted to make sure we got as much training done in the time provided as possible and so we chose to use a GRU implementation for or model. The accuracy of GRU models is generally similar to the performance of LSTM models which was the type that seemed most obvious initially. However, GRU models are computationally more efficient that LSTM models and thus we could get more training done in the same period of time.

We constructed our training dataset by downloading Wikipedia articles in 15 different languages. We chose languages based on number of speakers in the present day, and also on which countries currently have space programs or astronauts on the International Space Station.

We chose to include English, Chinese and Arabic because they are spoken by so many people that they must be included no matter what. Japanese, French, Danish, German, Italian, Norwegian, and Swedish are all spoken by nations currently using the ISS. And we included Russian and Hindi because Russia and India both have large space programs even though they aren’t on the ISS.

Our model itself was constructed on the implementation of GRU from PyTorch’s library nn. We trained the model for 250 epochs and for each epoch we randomly selected a 400 character chunk of one article from each language as our training data. At first we tried using a chunk of just one article, but the model wasn’t consistently decreasing in loss each epoch because it was skewing towards the language of the article selected. Using one chunk of every language solved this problem by keeping the training data balanced across all languages every epoch.