Dataset:

We are planning to use dialog-based language learning datasets. Since the system is used for sending natural language messages for astronauts, it would be better to use dialogue related data instead of books and articles and train models with vocabularies about astronauts' working environments. There is plenty of previous research done on similar topics, and we are able to find many available resources online, such as Kaggle, Meta research and Google research. Additionally, since the astronauts are guaranteed to speak one random language, perhaps we can also find datasets in different languages.

Method:

We implemented the program using a n-gram approach which looks at the frequency of our candidates appearing together. Our vocab size would be the size of all ASCII characters. Then we try to predict the top 3 characters using MLE with training data that we gathered about astronauts conversation. We used Python, NP and part of A3’s implementation to write this program.