Lign 167 Human-based metric(s)

1. Model generates text (sentence) 3 different ways, using the character based model, the phoneme based model, and the word based model. This is repeated 100 times per model type for 300 total sentences generated.

   1a. For each repetition the surveyees are presented with 3 different sentences from each model type; along with a control sentence randomly chosen from the corpus. Behind the scenes, when making the survey, each model type, along with the control group, is randomly assigned a letter for a label.

   1b. The surveyees are then asked to choose, from the 4 sentences provided at each question, which of the 4 sentences “makes the most sense,” or “Is the most clear”

2. At the end of the survey, the 4 different letter labels will have a score associated with them, based on the surveyor totaling up the scores for each label for each question. At most, 1 letter will have a score of 100, and all other letters will have a score of 0.

3. We will then compare the scores for each of the models. We would expect the control group to always have the highest score. Therefore, our second-highest score is the best version (character-based, phoneme-based, word-based) of our model, according to this metric.

There would be two versions of the survey, one with labeling for each sentence origin type, (for scoring at the end) and one without labeling. For the surveyee, each question would have 4 randomly arranged sentences, and they simply need to circle/put a checkmark next to the one they choose to be the most clear.