For our semester long project for CS 6740, we want to extend from the above paper in which various factors in phrasing affects their memorability. Ji Hun has already spoken to Professor Lillian Lee, who said she is willing to advise us throughout the project. (Heejung will speak with her soon as well, but because she hasn’t spoken to Professor Lee, we are going with the assumption that we will be working together).

For our extension, we are going to address one feature of phrasing that the paper failed to include: pronunciation. That is, do certain stress patterns in phrases serve as an indicator to whether the phrase is memorable? We plan to work off the publicly available dataset from the paper (http://www.cs.cornell.edu/~cristian/memorability_files/README_v1.0.txt) which includes a comprehensive set of movie quotes from IMDB, classified as either memorable or nonmemorable.

Currently, we have downloaded the dataset and started parsing through it. If we feel the need for it, we will also add onto the current dataset with phrases from more recent movies. We plan to add onto the already available dataset by tagging each phrase with its stress patterns using the following resource:

http://www.speech.cs.cmu.edu/cgi-bin/cmudict?in=syllable&stress=s

Afterwards, we will apply similar techniques that the paper presented in determining whether there is any relationship between the pronunciation of the phrases and their memorability. In addition, we wish to apply concepts that were taught in CS 4740, such as calculating for perplexity of certain quotes, where the language model we would be basing the measure off of would be the pronunciation of words of memorable quotes and unmemorable quotes. Although our plan of execution is a bit general at the moment, we plan to get advice from Professor Lee on a weekly-to-bimonthly basis to push our project into the right direction. One end goal we see, however, is creating a binary model that employs machine learning techniques to predict whether a quote is memorable or not based on their stress patterns.

In terms of an evaluation plan, we do not foresee too many problems because we are working off of a tagged dataset. One thing we could definitely do to validate our model is to perform cross-validation when we formally generate our model. Like we mentioned prior, we could also scrape more datasets from IMDB if we need to validate our results further.