Team 6: Jackson Confer, Sarah Lee, Emilio Gonzalez, Gaosong Liu

Professor Kraus

LING 144

17 February 2022

Team Project Update 2

Introduction – set up the question you are investigating, and state your hypothesis. This includes talking about the language variety you are using (even if it's English!), why the project you're embarking on is of interest, and a short description of what you are trying to achieve.

The question we are investigating for this project is how the filler utterances "um", "uh", and "like" are positionally distributed in spontaneous speech and how the overall rate of occurrences contrast from one another. More specifically, are the positions of these utterances held in sentence-initial, sentence-middle, or sentence-final positions. Also, the rate of occurrence we are seeking to capture is on a per sentence basis, unless we find a high enough rate to merit further demarcation.

The hypothesis that we have for this investigation is that the "uh", "um", and "like" filler utterances will occur most frequently at a sentence-initial position. This is based on the assumption that pauses come most naturally intra-sententially. In regards to the rate of occurrences, we believe that the filler utterance "uh" will provide the highest rate in comparison to the other filler utterances investigated. Quantitatively, we hypothesize that the rate of occurrence for the filler word "uh" will approximate to about one for every three sentences.

The language variety that we are targeting is Standard American English (SAE). We are in the process of approaching this issue, so naturally we find the most immediate language to have the most facility for a basic investigation. This endeavor is taken with the knowledge that there almost are certainly other filler utterances in different languages and dialects which express themselves quite differently.

The project we are embarking on is of interest because filler utterances play a significant role in modern discourse. This fact has become marred by the triviality often attributed to the linguistic tendency in media (film, television, internet media, etc.). We hope that this investigation will provide more information and analysis on the phenomenon which may work to elucidate the amount of work it does in modern speech.

a. Background – what have others said? What other work has been done on this topic?

The way we use language in natural spoken dialogue is instructive. Language can reveal many aspects of the diversity of a person's social identity. For example, the dialect and spoken language in the language can be used to infer the person's upbringing environment, and the different speaking patterns can also infer the attitude of others towards you. This paper studies frequently-occurring words in English daily oral communication, such as: like, um, uh. The research text on filler words is English movie dialogue.

As a kind of spoken American English, English film dialogue is an excellent resource for English learners and linguistic scholars to study. Although it is a modified language, in order to fully reproduce real life, there are a large number of oral phenomena such as non-fluent filler words in the film dialogue. The advent of the Internet age allows us to easily obtain a large number of original sound movies, and how to make full use of movie resources to obtain the data we need is very important.

Many researchers believe that in order to ensure uninterrupted speech flow, speakers should consider using lexical fillers, such as I mean, you know, like, listen, etc. After all, lexical fillers are meaningful words. Although they function basically the same as quasi-lexical fillers, they give a feeling of fluency. But in fact, quasi-lexical fillers are more labor-saving in pronunciation than lexical fillers, and because they have no semantics and do not require brain thinking resources, they may be more conducive to planning and monitoring utterances. The large number of such words in spoken American English also proves the necessity of their existence.

b. Methods – How did you/will you collect your data? What language variety and context of use does this represent? What kind of data is/will be present in your dataset? How did you narrow it down (i.e. how did you filter after you found the initial data)? How is it/will it be annotated?

We will primarily be drawing our data from corpora of movie transcripts. There are many well-known movies that focus on characters who use filler-*like* somewhat regularly in their speech. In particular, we will be drawing from movies centered on young, white, Californians within the past 30 years such as *Clueless, Fast Times at Ridgemont High* and *Lady Bird*, whose dialect is a well-cited example of *like*-filling. In particular, we looked for transcriptions that transcribed both *like* and more traditional filler words. We aim to create a program that filters out the dialogue, leaving us only with instances where the actors are using filler words, where we can then easily compare occurrences of *like* with *uh* or *um.*

We will filter our data with a Python program, and generate a list of sentences with *like* and lists of sentences with a traditional filler. Filler-*like* more often than not has a comma immediately following it orthographically, making it easy to find the specific instances of the word with regex. Ideally we want to devise a way to gather instances transcribed without commas, which will have a distinct syntax, although how easy the occurrences are to target is another question. We're not sure of the exact parameters of how to narrow it down yet, but the idea behind the main filter is there. We're sorting the data by the filler word used, which will more easily allow us to compare the distributions of the words, before annotating based on whatever linguistic features we notice (position in syntax, etc.).