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

Professor Kraus

LING 144

6 March 2022

Team Project Update 3

The question we are investigating for this project is how the filler utterances “uhm” and “like” are positionally distributed in film scripted speech and how the overall rate of occurrences contrast from one another. More specifically, how these utterances are distributed in a variety of syntactic positions. Also, the rate of occurrence we are seeking to capture is on a comparative basis between the different filler utterances, along the entire corpus.

The hypotheses that we had for this investigation includes the postulation that the “uhm” and “like” filler utterances will occur most frequently at a clause initial positions. This is based on the assumption that pauses come most naturally intra-sententially. In regards to the rate of total occurrences, regardless of positioning, we believe that the filler utterance “uhm” will provide the higher rate in comparison to the other filler utterance “like” investigated. Quantitatively, we hypothesize that the rate of occurrence for the filler word “uhm” will approximate to about one for every five-eight sentences.

The language variety that we are targeting is Standard American English (SAE) with a focus on California English. 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 influence it has on our modern conception of speech. By targeting film scripts we are able to discover the reflection that modern speech has in popular culture. We also hoped to reveal the syntactic distribution of the different filler utterances, and provide a foundation for discussion about their linguistic function.

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., which stand in contrast to what we will call *quasi-lexical* fillers like *uh* or *um.* After all, fillers are not simply errors or meaningless noise, but broadly implementable into the structure of the syntax and discourse (Fox Tree 2007). Speakers can tell the difference between different fillers, indicating that there ought to be more to their distribution than just sheer chance. Additionally, a distinction between these categories (albeit with the names *discourse markers* and *filled pauses*) is well supported in the literature historically (Laserna et al. 2014). Additionally, the broader effect on social perception linked to filler usage is also relatively well studied, as well as many of the factors in real-time discourse that might lead to filler words being needed and correlate with their usage, such as nervousness or extemporaneousness (Duvall et al. 2014).

Although lexicals function basically the same as quasi-lexical fillers in many contexts, they give a feeling of fluency which could be traced to the fact that they don’t stick out as non-words. However, 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. These factors make them highly comparable but still distinct

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 40 years (*Clueless, Fast Times at Ridgemont High*, *Mean Girls*, and *Lady Bird*), whose dialect of Coastal Californian English 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 list of sentences with a quasi-lexical fillers. The filler-*like* will provide the most difficult task. Every instance of the other two fillers will be gathered based on our assumption that all uses of them function as fillers. However there are non-filler uses that are common to the word *like.* 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.).

Results

As a result, we found 20 more of the traditional filler "uhm" than the "like" filler(Figure 1). This meets our expectations since we have hypothesized that we are likely to find the traditional filler than the "like" filler. Through the process of annotation, we were able to find different places where two different filler words appear. As for "like", within its 24 occurrence we gathered, there were 13 occurrence that "like" was uttered after the copula, there were 3 occurrences where "like" occurred as clause starter, and there were 8 occurrence where "like" occurs elsewhere, which we couldn't find syntactic similarity since it's relied on pragmatics (Figure 2). For filler "uhm", out of 44 occurrences we captured, 31 of the occurrences were placed as sentence initial and 13 occurrences were elsewhere where we couldn't syntactically group (figure 3).

-RESULTS IN GRAPHS:

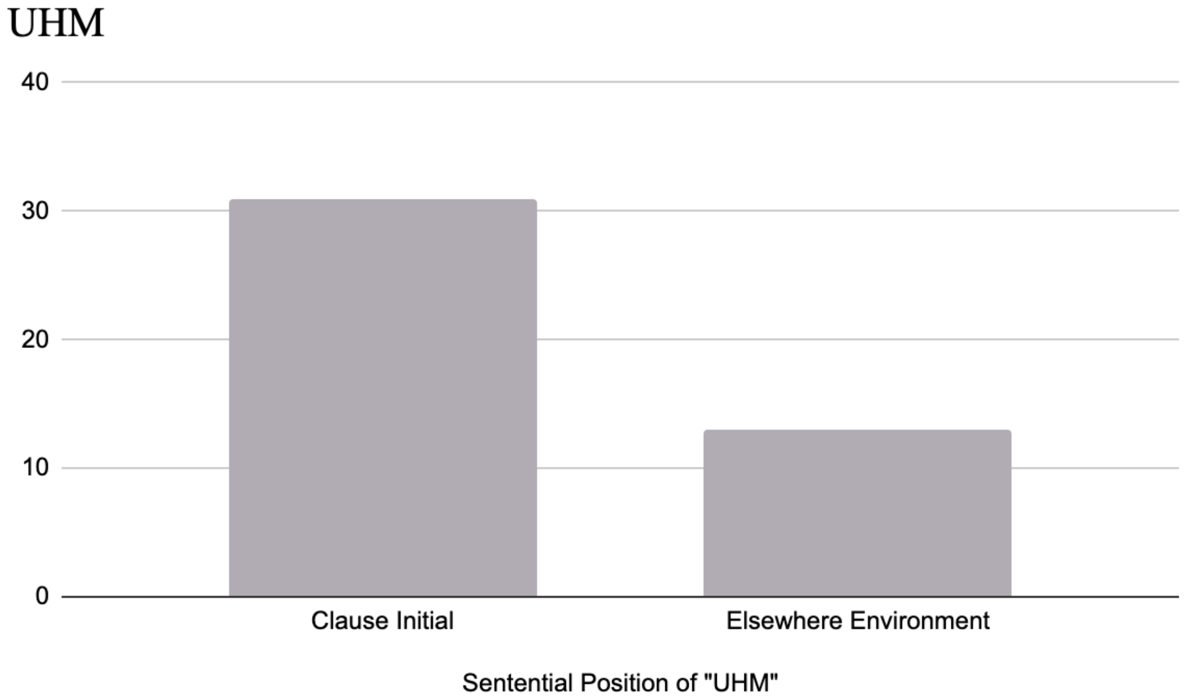

Figure 1

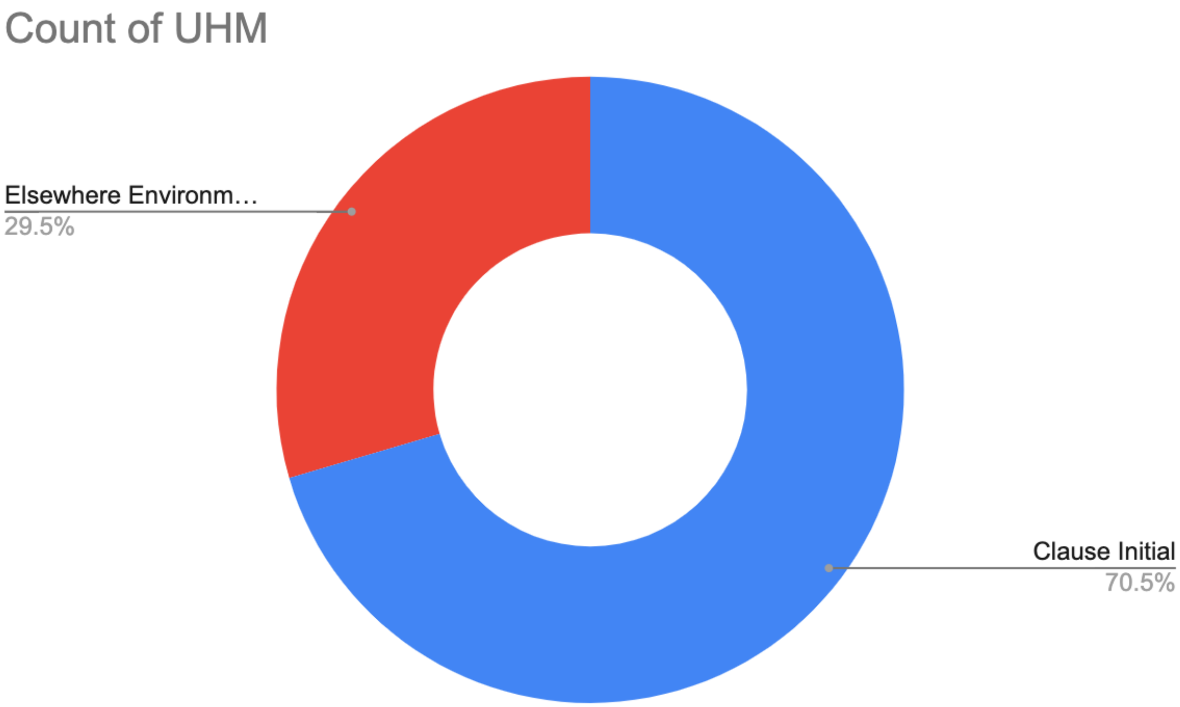

Figure 2.1

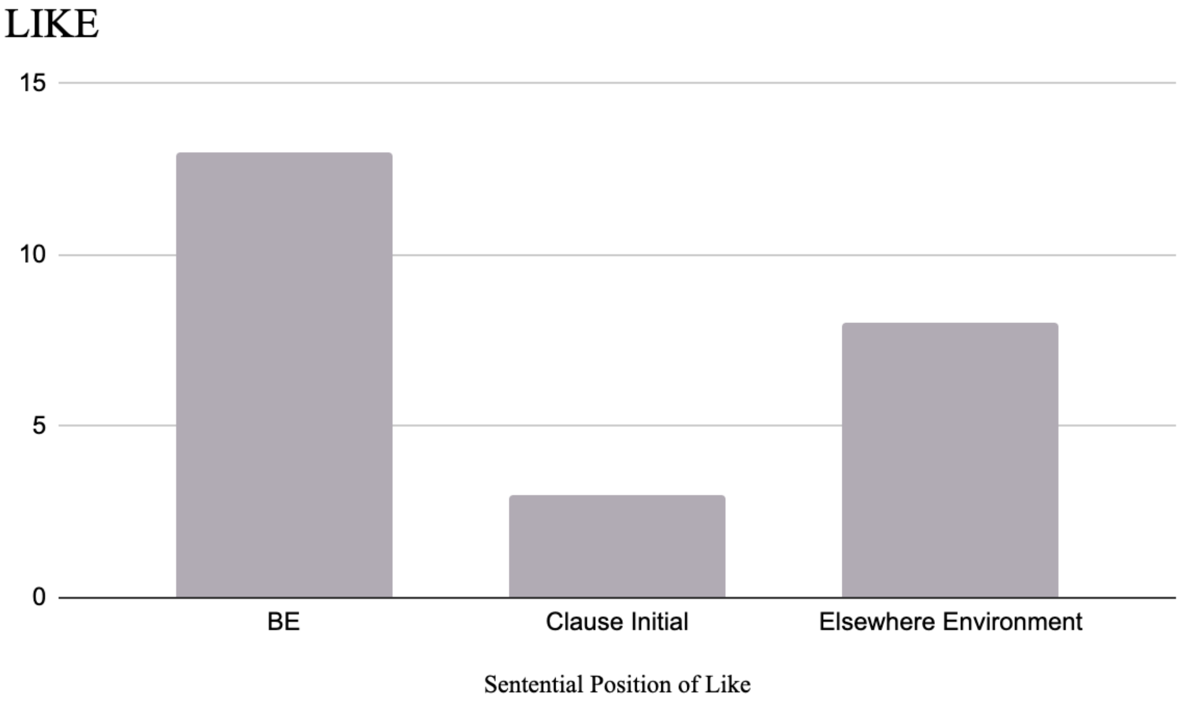

Figure 2.2

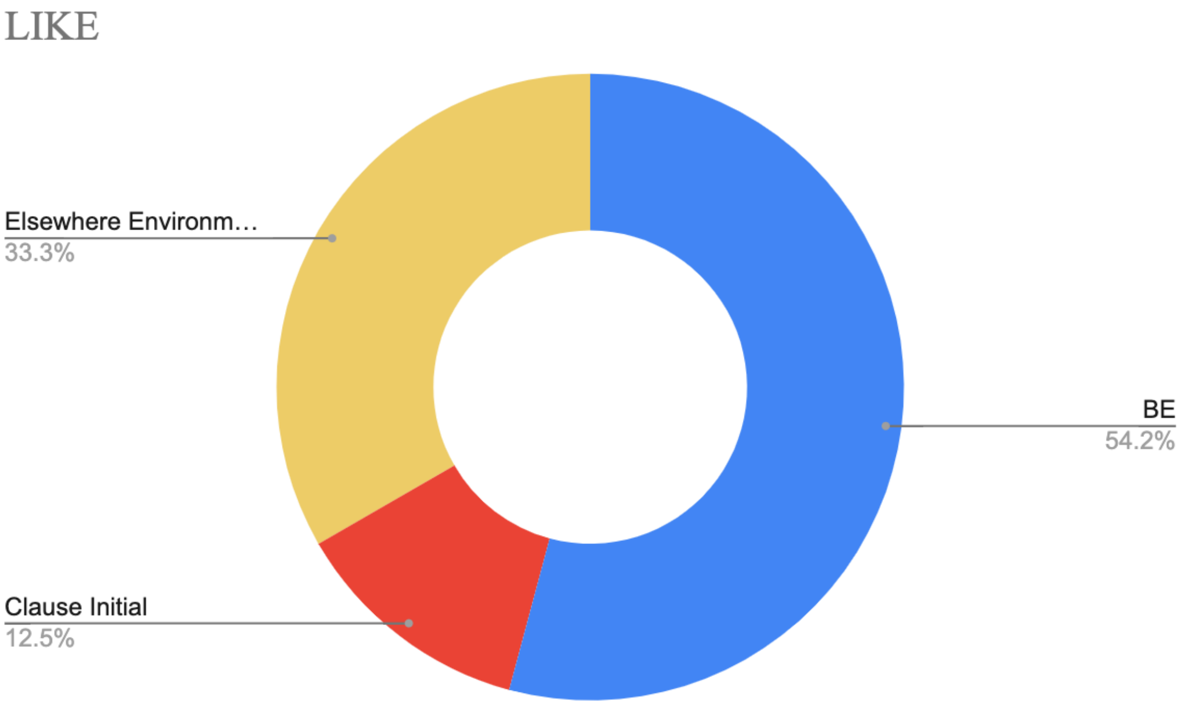

Figure 3.1

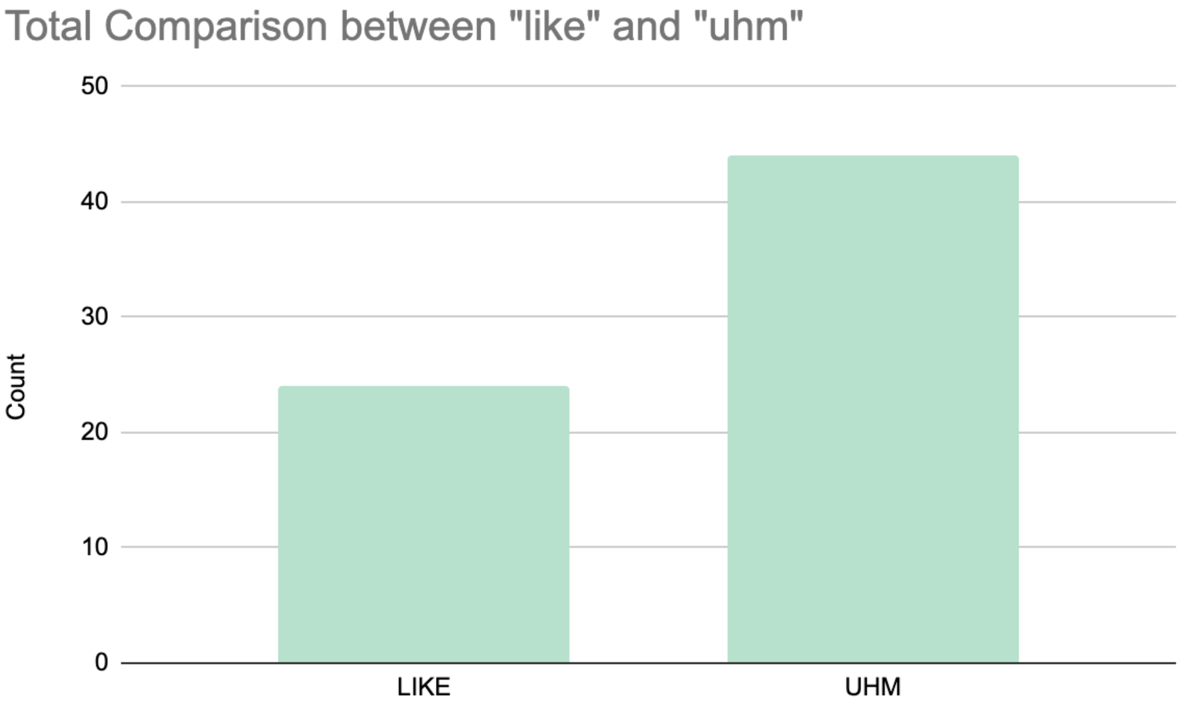

Figure 3.2

Discussion

At the start of this investigation, we attempted to approximate the distribution of the filler word utterances “like” and “uhm”. We stated that both filler utterances ought to occur most frequently at clause initial position. Our reasoning was based on what we deemed filler words to function as, namely, pauses with sound. In reference to the syntactic distribution of the filler utterance “uhm”, we found in our data was that there was a significant majority of clause initial occurrences (70.5% of “uhm” occurrences were located at a clause initial position). However, in relation to the evidential distribution of occurrences for the filler word “like”, there was a stark contrast. Although there were a few instances of “like” being used to introduce clauses (12.5% of “like” occurrences were located at clause initial positions), the distribution pattern lended itself to a demonstration of “like” being tied to the copula (54.2% of “like” occurrences were preceded by the copula). This leads us to tie systematic depictions of “like” and “uhm” to separate mechanisms.

The other hypothesis we stated during the onset of this investigation was that the filler utterance “uhm” will provide the higher rate of occurrence in comparison to “like”. That hypothesis was borne out in our data set with the total amount of “uhm” occurrences (44 occurrences) reaching a higher total than the total amount of “like” occurrences (24 occurrences).

Something particularly surprising about our results was that the filler utterance “like” was found to have a remarkable patterning with the copula “be”. We did not foresee this result, which is striking considering its clear distribution demonstrated in our data. The surprise was bolstered by the fact that the filler utterance “uhm” had not one instance of following the copula. This suggests in itself that there was an unforeseen connection between the copula and the systematic use of filler words.

Another surprising result was the variation in which these filler utterances were used. This was so evident that we created an elsewhere category to group these occurrences. A crucial note is that the elsewhere environments, although they may have not been sentence initial, commonly were used with conjunctions which suggests further evidence for filler utterances introducing clauses. However there were clear examples of the contrary, such as the filler utterance “like” being used to separate a preposition with its sister DP (“Could you give us some privacy for like one second?(Mean Girls)).

We recognize the fact that our data could have resulted in a different analysis using a different data set. However, the search for filler utterance patterns in California English movie scripts, and perhaps California English more generally, can begin with a data set like this. Most of the movie scripts we used were based in California, and are widely recognized as expressions of California English in cinema. Additionally, even though Mean Girls is based in Evanston, Illinois, it is a glaring example of California’s influence on midwestern culture. These points direct us toward the belief that our data set provided an appropriate target for this investigation.

Conclusion

With the result we see from our research, we can conclude that there are different syntactic locations and pragmatic functions between “like” and “uhm”. We can take further steps and take a look at the natural language data since our data only consists of the movie scripts and observe how these two different filler words act differently in natural language. Some different approach we could have taken is looking at the pragmatic significance behind the filler words and being able to sort out the ‘elsewhere’ case of two filler words. Within the process, it would have been much more convenient during the annotation to have a more sophisticated regex that could remove the exceptions that get caught up in the filtered data. For example: “(Class breaks into applause)CHER Thank you very much.MR HALL UHM, Amber (Clueless).” Also having regex code which can remove the character names within the program that can provide more clean data would be better. Such as “REGINABut if you like him…” to “But if you like him…”

References

Duvall, E., Robbins, A., Graham, T., & Divett, S. (2014) "Exploring filler words and their impact." *Schwa*. Language & Linguistics 11.

Fox Tree, J. E. (2007). “Folk notions of *uh, um, you know,* and *like**.” *Text & Talk* 27(3). De Gruyter Mouton.

Gerwig, G. (Director). (2017). *Lady Bird* [Film]. IAC Pictures, Scott Rudin Productions, Management 360. (transcript accessed from https://www.dailyscript.com/scripts/LADY_BIRD_shooting_script.pdf)

Heckerling, A. (Director). (1982). *Fast Times at Ridgemont High.* [Film]. Refugee Films. (transcript accessed from https://imsdb.com/scripts/Fast-Times-at-Ridgemont-High.html)

Heckerling, A. (Director). (1996). *Clueless*. [Film]. Paramount Pictures. (transcript accessed from https://imsdb.com/scripts/Clueless.html)

Laserna, C. M., Seih, Y. T., & Pennebaker, J. W. (2014). “Um... Who Like Says You Know: Filler Word Use as a function of Age, Gender, and Personality." *Journal of Language and Psychology,* 33(3) (328-338).

Waters, M. (Director). (2004). *Mean Girls* [Film]. Paramount Pictures. (transcript accessed from https://www.scriptslug.com/assets/scripts/mean-girls-2004.pdf)