Team 4: Pamela Casipe, Chris Zepeda, Brianna Garcia

LING 144, Winter 2022

Kelsey Kraus

March 1, 2022

Team Update 3:

Scripted vs. Non Scripted Spontaneous Speech

I. Introduction

Our team will look into the frequency of disfluencies in scripted spontaneous speech in Euhophoria versus natural spontaneous speech. We are interested in examining how they compare and if the disfluencies in scripted speech occur as often as they do in actual spontaneous speech. Our hypothesis is that they will not occur as often in scripted speech as they do in actual speech because scripted spontaneous speech can only mimic actual speech to an extent.

We will take a look at spontaneous speech just for English in our data sets. In terms of the disfluencies we will examine, we have decided to narrow it down to the use of *um* and *uh*. This will allow us to look at the more common disfluencies across the English language and have clearer counts for the disfluencies that occur in both settings. We believe that our topic is particularly interesting because of the way scripted speech examines language patterns to resemble real life scenarios. It is intriguing to see when and why the disfluencies may be used across speech.

One of the articles mentions using one disfluency over the other to indicate a form of politeness in natural spontaneous speech. Looking at these types of patterns may help film writers understand when it is appropriate to use certain disfluencies and insert it correspondingly in scripted speech. This can definitely be very informative when it comes to speech studies and attempting to imitate speech as naturally as possible for not only film writers, but even machine language generators.

Context: A veteran baseball player sees two little children (both about 10 years old) playing baseball at the park.

a. That was, **uh,** quite a crazy catch you got going on there, Kiddos!
b. Thanks sir, that is what you call a, **um,** a home-run!
c. **Um**, do you know him?

Context: A woman is trying to schedule an appointment at the beauty salon.

a. Hello Miss, what type of service were you looking for today?
b. I would like to get my roots covered, layers cut, and **uh**, a fresh set of acrylics please.

II. Background

In his article, Ralph L. Rose actually spoke on the frequency of various disfluencies in natural and scripted spontaneous speech. Similarly, this researcher was interested in the natural occurrence of these disfluencies in everyday speech and how well films and movies were able to replicate the use of them. With the data he gathered, he was able to find that the occurrences for them were actually quite similar across natural spontaneous speech and scripted spontaneous speech.

When it came to differences however, most of them came from the level of structural distribution. This was a reference to the use of disfluencies across boundary and non-boundary positions. Sentence boundaries refer to the capitalization of *um* and *uh,* meanwhile sentence non-boundary refers to disfluencies that are not capitalized.

III. Method

We used two different sets of data for our project: For the scripted spontaneous speech, we drew upon the script of episodes 2, 3, and 4 from the TV show, Euphoria, and we drew upon two transcripts from the UC Santa Barbara natural speech corpus for our natural spontaneous speech. Both collections of data were initially filtered by using Python to remove any unnecessary text such as timestamps, then regular expressions were used to split the data set by newlines in order to break down the data into utterances. Once filtered, both data sets were transferred to a Google Sheets spreadsheet and organized based on text source, utterance, disfluency type, and position of “uh” or “um.”

IV. Analysis & Results

Our team found that the natural spontaneous speech had more instances of the “um” and “uh” disfluencies overall. Based on Figure 1, we can see that there were about three times more instances of the disfluency “uh” found in the Santa Barbara transcripts in comparison to the Euphoria script. There were about two times more instances of the disfluency “um” in the Santa Barabara transcripts as opposed to the Euphoria script as well. In terms of disfluency type, there were more instances of “uh” in both natural and scripted spontaneous speech

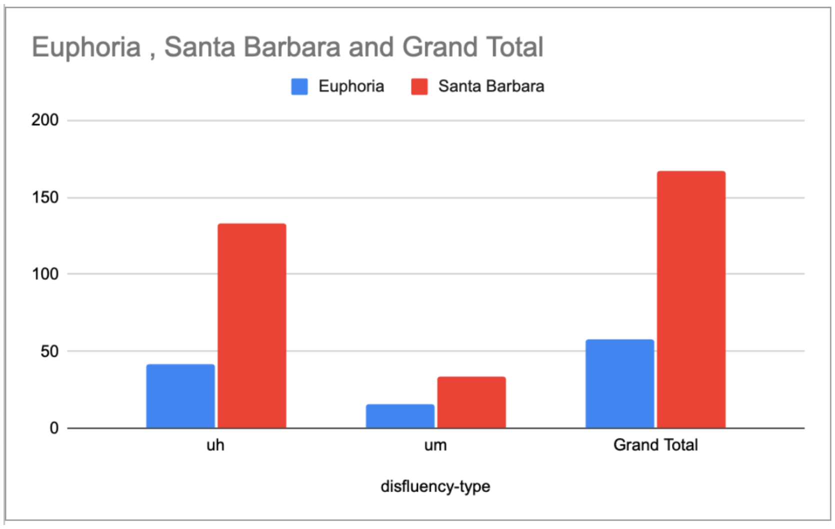


Figure 1

Figures 2 and 3 depict the amounts of the different positions of the disfluencies. Our team found that the initial position was the most prevalent in the data, with initial "uh" being the most prevalent, and the middle position was the least prevalent.

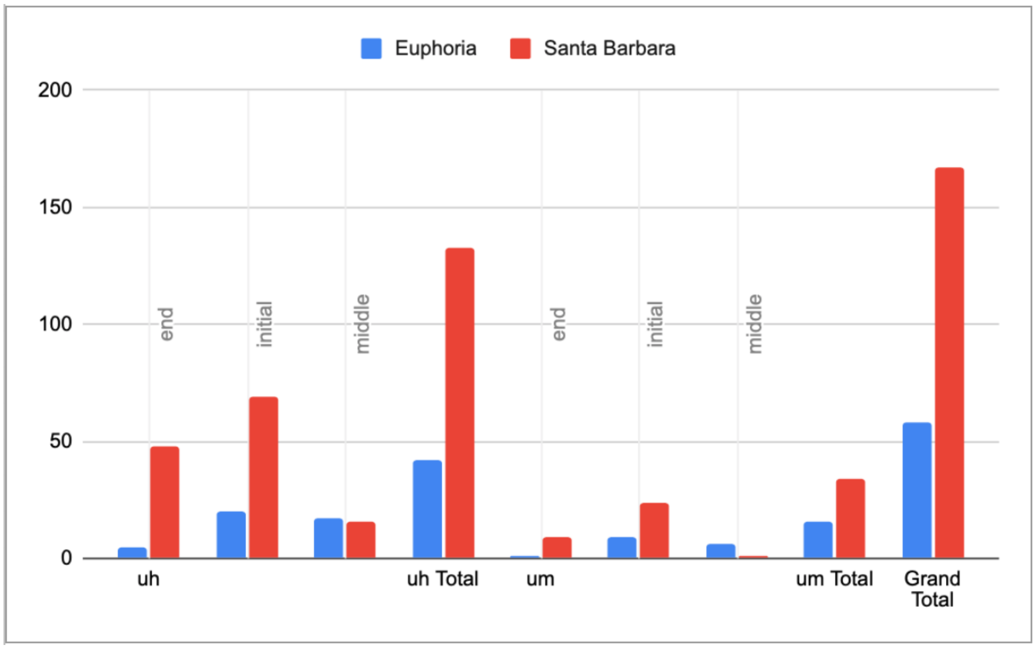
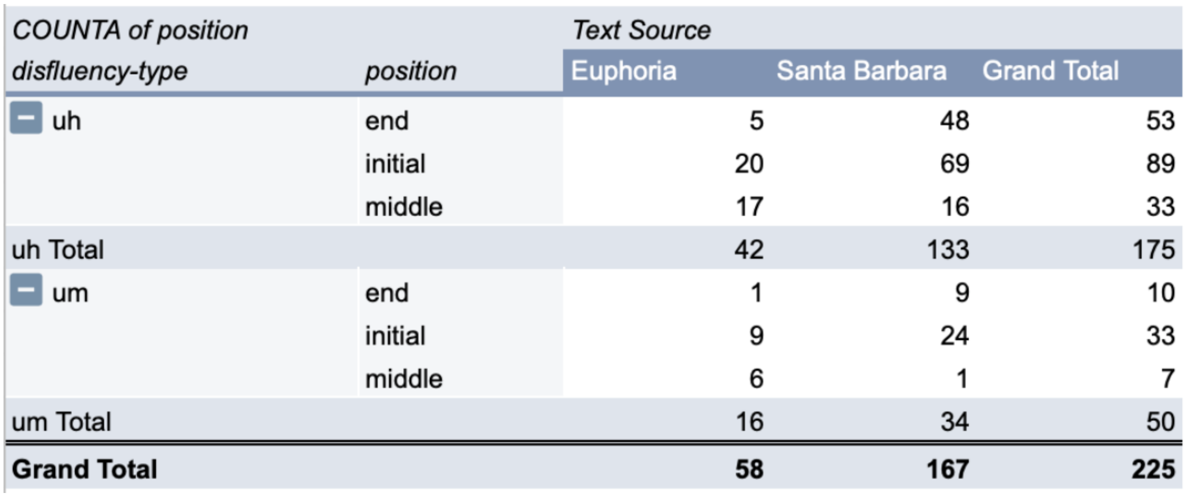


Figure 2



COUNTA of position		*Text Source*		
disfluency-type	*position*	Euphoria	Santa Barbara	Grand Total
⊟ uh	end	5	48	53
	initial	20	69	89
	middle	17	16	33
uh Total		42	133	175
⊟ um	end	1	9	10
	initial	9	24	33
	middle	6	1	7
um Total		16	34	50
Grand Total		**58**	**167**	**225**

Figure 3

V. Discussion

Our findings matched with what we had originally hypothesized. They show that the disfluency “uh” and the initial position were the most prevalent in both types of data. The most popular position being the initial position was quite surprising, but understandable.

The data could have definitely been different if different data sets were used. Euphoria episodes 2, 3, and 4 were specifically chosen because the first episode didn’t have a lot of instances of any disfluency, which shows that the Euphoria data was extremely dependent on which episodes were examined. Although we weren’t able to do so, we think that our data set would have been most reflective if we used the actual transcript of each episode, but the transcripts were not anywhere online and we did not have the time to create our own transcript.

These findings show that initial “uh” is the type of disfluency that natural spontaneous speech gravitates towards, which is information that can be used for creating scripted spontaneous speech. The ratio of initial “uh”/”um” and end “uh”/”um” can be utilized to create more natural scripted spontaneous speech in various forms of entertainment such as books, plays, TV shows, and films.

VI. Conclusion

Our team found that natural spontaneous speech had more instances of disfluencies, specifically “uh” was most prevalent in its initial position for both data sets. Some things that we could have done differently were to make the utterances from the UC Santa Barbara transcript longer. Because we split the data by newlines, the Santa Barbara dataset had more utterances in total than the Euphoria data set. This could have skewed our results and not give a completely accurate representation of the phenomenon.

Our research could be taken further by examining the semantic role of each type of disfluency based on their position. The positions can be studied to see if there is a difference in semantic role depending on if the disfluency occurs in the initial, middle, or ending positions. Additionally, there could be research done on why initial "uh" was most prevalent as opposed to end "uh." We could also examine other types of disfluencies such as "like" and "oh" and compare those with "uh" and "um."

VII. References

1. "Santa Barbara Corpus of Spoken American English." *Department of Linguistics - UC Santa Barbara*, https://www.linguistics.ucsb.edu/research/santa-barbara-corpus.
2. Zayats, Vicky, et al. "Disfluencies and Human Speech Transcription Errors - Arxiv." *Disfluencies and Human Speech Transcription Errors*, https://arxiv.org/pdf/1904.04398.pdf.
3. "EUPHORIA" SCRIPTS & TRANSCRIPTS FROM EVERY EPISODE - https://8flix.com/transcripts/euphoria/season-1-8772296/tt8772296s1-dialogue-transcripts/
4. Rose, Ralph L. "A Comparison of Filled Pauses in Scripted and Non-scripted Spontaneous Speech." http://www.roselab.sci.waseda.ac.jp/resources/file/LPSS2019_Rose_scripted_filled_pauses_paper.pdf
5. Shriberg, Elizabeth. "To 'errrr' is human: ecology and acoustics of speech disfluencies." https://www.jstor.org/stable/44645157
6. Corley, Martin and Stewart, Oliver W. "Hesitation Disfluencies in Spontaneous Speech: The Meaning of um." https://compass-onlinelibrary-wiley-com.oca.ucsc.edu/doi/full/10.1111/j.1749-818X.2008.00068.x