Introduction: Discourse relation parsing is a research area in Natural Language Processing for analyzing discourse structure by identifying relations. Most previous works are limited to written texts which are well-structured, unlike spontaneous dialogues. I will be contributing a dataset consisting of discourse relations annotated for long dialogues, specifically transcripts of TV show episodes from the Forever Dreaming collection (Chen, 2022).

Methods: Previous research on dialogues includes the STAC dataset (Asher, 2016) which is an annotated multi-party chat log for an online game. After a testing trial of the STAC corpus, I defined discourse relations for the transcripts. I have then annotated an episode using Inception, an online web application. Based on my findings, I revised the annotations and redefined relations. I have noted relevant challenges or new concepts. I also recorded the frequencies of the different types of relations.

Results: I found that certain relations such as “comment” and “continuation” were more prevalent due to the structure of longer dialogues. I observed the use of rhetoric and sarcasm in utterances which were more difficult to define, as expected with creative language. I compiled the data onto a table as well as a bar graph showing the frequencies of 21 different discourse relation types.

Conclusions: There is a major difference between long multi-party dialogues and other written texts. However, to contribute a meaningful dataset, a sample size of at least 1000 episodes is likely required. Relation labels may also still need to be redefined and a variety of TV show genres should be explored.

References:

Acknowledgements:
Yilun (Bobby) Hua
Professor Kathleen McKeown