Investigating the Prevalence of Stereotypical Language in Actual and Movie Conversations

Student: Monika Ciecka
Advisor: Rivka Levitan
CUNY Brooklyn College

Goals and purpose of project

The goals of this year's research were to gain an exposure to natural language processing techniques by applying these techniques to analyze gender representation in movies. Robin Lakoff's paper *Language and Woman's Place* introduced linguistic markers that are typically characteristic of "women's language", which include: hedges, profanity, formality, and tag questions. The quality of a character's representation in a movie would be reflected in their language, therefore making this fit for a computational language analysis.

We analyzed these features in the contexts of actual and movies conversations to see if there was a large difference between stereotypical language between them. The hypotheses that led our analysis included: stereotypical female language will be more prevalent in movies than in actual conversation, and the stark contrast between male and female speech will be greater in movie conversations.

Process used in completing the research

For our research, two sets of data were used: movie data and Switchboard data, which provided the actual conversation portion. The movie data was obtained from IMSDB prior to the start of CREU research. With the data now available, a literature review was conducted. This helped spur some ideas for features that would classify hedges and other features.

Tag questions were detected by a regular expression of the pattern <comma>, <tag phrase>, <question mark> where the tag phrases were found within the data set. Hedges were detected by checking for the presence of certain hedge words from a bag of words. Profanity were detected by using a regular expression to match variations from a list of profanity words that was put together. Formality markers were found using a calculation including various part of speech tags.

Collectively, these features were used to create a stereotype index that serves to give a rating to a movie in terms of how well or poorly it performed in terms of representation.

It was necessary to ensure accuracy of the data, so the genders of the top speaking male and female characters were manually checked against the IMDB database.

Conclusions and results achieved

This project has applied Lakoff's hypotheses, along with our own, to two different corpuses. We have found that in the actual conversations, females are no less likely to use profanity and no more likely to use hedges, and their language is less formal than males. This contradicts our hypotheses. However, the results for the movies does mostly support our hypotheses. There was a significant difference for hedges, profanity, and formality. Searching for these features aids to add another layer of analysis to the relatively straightforward Bechdel test by analyzing the content of female speech as opposed to only its existence in a movie script. In the future, more features could be added in order to add more to the analysis and create a more advanced stereotype index.