Detecting Gender Inequality and Language Evolution in Movie Dialogue

Yulin Yu
Master's Thesis
Advised by Prof. Kentaro Toyama
Committee Member: Prof. Paramveer Dhillon
University of Michigan, School of Information
April 16, 2020

Abstract

It is expected that gender inequality in American movies has decreased over the last century, reflecting larger societal trends. But, can this be verified through a computational analysis of movie dialogue? In this study, we apply 13 word-based metrics intended to rate words according to such criteria as gender-ladenness, pleasantness, and age of acquisition, and ask whether any are effective at measuring gender differences in movie dialogue. As a secondary outcome, we also seek to determine whether gender differences have decreased over time and how. We find that metrics that aim to capture word concreteness, imageability, happiness, easiness of understanding, emotional polarization, and gender-ladenness are able to detect gender differences in dialogue. We also find evidence that at least since the 1960s, there has been a steady convergence in female and male dialogue in Hollywood films.

Introduction

Gender inequality in the United States has decreased over the last century (Bureau 2006). Since film are known to reflect social norms(Wedding and Niemiec 2014), Most people also believe in the last decade gender inequality has decrease in movies (Lauzen and Dozier 2005; Neville and Anastasio 2018). In addition, to qualitative analysis, scholars use quantitative analysis in order to understand the gender inequality through the screen time and conversation topics between females. There is one question that arises here: Are these methods reliable in measuring gender inequality?

One informal attempt at quantifying gender inequality in movies is the Bechdel Test. This famous test is designed to measure women's representation in the movies (wik 2020). This test is created by Alison Bechdel in a comic called "Dykes to watch out for" (Bechdel and others 2010). Originally,this is not a theoretical proof. This is an observation based on pop culture. A movie passes the Bechdel Test when it contains a scene that satisfies the following requirements: 1) There are at least two female characters in the scene, 2) these females talk to each other 3) the conversation is about something other than men. This test could be used to measure gender inequality in movies over the years, but only provides a binary metric for any single movie. Also, while intuitively about gender representation in movies, it is not clear exactly what the test measures.

In linguistics and computational social science, many studies try to measure the different uses of language between males and females. Resaerchers approach through lexicon, machine learning, or count-based features (Ramakrishna et al. 2015; Newman et al. 2008; Schofield and Mehr 2016). Some have developed a *lexicon*-based metric to differentiate gender-ladenness of individual words and used the metric to identify gender-related differences in movies across genres (Ramakrishna et al. 2015). Others have tried to use features not necessarily tuned for gender to identify differences in dialogue (Schofield and Mehr 2016). However, none of the existing works have sought specifically to identify which features are most effective in detecting gender differences in dialogue, and none have examined the evolution of gender representation in movies changes over time.

In this paper, we seek to evaluate which of a number of linguistic features is most effective in determining changes in gender inequality in movie dialogues. As a secondary result, we also seek to confirm changes in gender equality in movies over the past several decades.

This work makes three kinds of novel contributions. First, we found dialogues can predict the trend of gender gap change in the movies. Specifically, we found that metrics extracting from the male-female difference that depict the concreteness, imageability, pleasantness, easiness of understanding, and association of gender have significant associations with gender equality trends in the society. Secondly, we also understood how female and male change their language to response to gender inequality change over time: Since the 1960s, Hollywood movies have seen a convergence in language at the word level between female and male characters. Third, we consider what might be the cause of the above results.

The rest of the paper is organized as follows. In Related Work, we introduce the relevant studies in the past. The Empirical Setup section presents our experiment design. Next, we present the results and in the discussion section we discuss the future study and implication.

Related Work

Scholars have conducted significant amounts of research in understanding gender differences in written and spoken language. Social scientists have tried conducting empirical studies to understand the language difference between the two genders. Computer scientists have tried exploring useful features in order to build classifiers to identify the gender of the writer. Meanwhile, there are some researchers specifically studying gender differences in film dialogues.

Gender Differences in Linguistics

Social scientists have studied gender differences in language. Researchers started becoming interested in gender differences in language around 1990. In the early work, researchers tended to use small data-sets to generate patterns of differences within gender in the aspect of the use of words, tone, and emotion. Some studies found females tended to ask more questions (Mulac et al. 1988), to be wordier (Mulac and Lundell 1994), to use more first-person singular pronouns and to refer to emotions more often than males (Newman et al. 2008). While, they found males tended to tell the audience to do something, such as sentence, 'let's go to the park.' Earlier studies also found male offer more opinions than female, use more words and articles, talk more often, and use longer words(Mulac and Lundell 1994; **?**). However, when studies use different samples of data, lots of conflicts resulted. A study of email communication found that both males and females were equally likely to ask questions (Thomson and Murachver 2001). While a study analyzing different genders of managers giving professional criticism in a role play indicates male asks more questions (Mulac, Seibold, and Farris 2000). These conflicting results indicate two things, first, small data set might involve bias when representing language in different genders. More importantly, the gender differences in language might highly correlate to the scenarios where the text data is created.

Thanks to the technology revolution, gender differences in texts can be studied through large scale data analysis. Previous research differentiates gender within text through the style of word, he meaning of word, and embedding. Using Linguistic Inquiry and Word Count(LIWC), Newman (Newman et al. 2008) indicated that the biggest difference between male and females word usage is functional words. Female use more function words, especially the third person pronouns and first-person singular. Another main difference is that females use more social words and sensation words than men; especially vocabulary related to family, friends, feeling, and hearing. male use the following type of words much more than women: words with more than six letters, numbers, articles, swear words, occupation words, and sports-related words (Newman et al. 2008).

These studies provide important insight into how males and females use words differently in spoken and general written language. Our study will contribute to understand how professional scriptwriters consciously or subconsciously reflect patterns of speech in gender over the years. Our study will also focus on studying the historical trend of gender language evolution and to approach difference language evolution in movies for male and female.

Computational Methods for Gender Difference

Computational methods have been used to determine gender differences. One type of study focuses on building better models to classify text written by different genders (Schofield and Mehr 2016; Mukherjee and Liu 2010; Peersman, Daelemans, and Van Vaerenbergh 2011; Burger et al. 2011; Sboev et al. 2016). Another type of related study uses crowdsourcing methods to generate labels for language features (Ramakrishna et al. 2015). And there is another related field of study which demonstrates methods using existing NLP techniques to analyze the gender bias in movie dialog (Schofield and Mehr 2016).

A large amount of research developed machine learning models with features to classify the gender of the author in the text. (Schofield and Mehr 2016; Mukherjee and Liu 2010; Peersman, Daelemans, and Van Vaerenbergh 2011; Burger et al. 2011; Sboev et al. 2016) One of the important insights in these papers (Schofield and Mehr 2016; Mukherjee and Liu 2010; Peersman, Daelemans, and Van Vaerenbergh 2011; Burger et al. 2011; Sboev et al. 2016) is feature selection. Some proved part of speech, the occurrence of words, emotional words, and types of speech are correlated with gender difference in language.

One study using crowd-sourcing techniques to provide each word a rating and some invlove rating related to gender difference (Ramakrishna et al. 2015). Each word is given a score from -1.0 to 1.0 along each dimension. Then, any text can be evaluated for its average score on some dimension. For example, one of their dimensions is "gender ladenness." Through comparing the gender ladenness in movies from different genres, the authors found different movie genres use language differently in terms of gender ladenness. In this paper, the author generate the gender ladeness features (Malandrakis and Narayanan 2015). Our work uses the features developed by these studies as a basis for determining what can be used as a measure of gendered dialogue and gender inequality in movies.

One type of study specifically use embedding technique to track the social change in gender. Garg (Garg et al. 2018) have proved that pre-trained word embedding model such as pre-trained Google News and Google Books/Corpus of Historical American English embedding can use to quantify the gender stereotype over year. Kozlowski (Kozlowski, Taddy, and Evans 1803) also demonstrate using word embedding to detect difference in culture meaning between gender.

These studies provide techniques to identify or differentiate text produced by different genders. However, these studies have not evaluated how different methods better align with reality and present different aspect of culture change in gender in the lens of movie dialogues, which is the primary focus of this study.

Gender Inequality in US Movies

Studies are trying to interpret and demonstrate gender inequality in movies and the US society. Scholars have conducted these research in movies regarding different aspects

including screen or speaking time, age, stereotype presentation, and sexy attire (Lauzen and Dozier 2005; Neville and Anastasio 2018). On one hand, the screen and speaking time of female characters increased over time. On the other hand, other studies have found the majority of female characters presented as young and physical attractive, while male characters are presented as older (Neville and Anastasio 2018). When considering these aspects, the definition of gender inequality seems complex (Lauzen and Dozier 2005). This also reflects to the trend of gender equality is US society.

The overall gender equality in the US seems to correspond to its trend in the movies. The inequality still existed but also decreased in important aspects such as education, workplace, and political participation (Bureau 2006). However, when considering social life or stereotype, males are still more likely to become the sole source of income in the married couple (Wilkie 1993). Gender inequality in both film and society present complexity in an overall trend. It is likely that gender inequality in movie partially reflects this equality in the society. Our study contributes to understand how dialogues change the response to gender inequality trend in movies.

Empirical Setup

Our main focus is to identify the Ramakrishna features (Ramakrishna et al. 2015) that are most effective at determining gender inequality in movies, based on character dialogues. Then once we have the most effective features, we consider trends in movies over time using those features.

For the first step, we need some proxy for ground truth. For this, we use sociological data (described below) and assume that movie dialogues reflect societal trends. Then we try to determine which of the Ramakrishna features result in the strongest correlations between movie dialogue scores and the social ground truth. We then test the top features with two movie series, Star Wars and James Bond. Those series were rated by human raters to determine their relative rankings with respect to gender inequality.

We will introduce the setup by describing data collection, features extraction, methods, and pipeline process in the following paragraphs.

Data

We use the Cornell movie-dialogues corpus(Danescu-Niculescu-Mizil and Lee 2011) to identify the most effective features for gender inequality. This corpus contains 894,014 movie script lines from 1,068 movie scripts tagged with cast lists, IMDB information, genre, release year, and conversation label. The movies in the corpus were produced from 1930 to 2010. This dataset is automated generated from raw publicly available movie scripts. After dropping all movies which couldn't match with tag mentioned above or had less than 5 IMDB votes, the data have 617 movies left. To the best of our knowledge, this data is the largest data set of movies dialog data with cast and conversation label available.

To evaluate the top metrics, we annotated ten Star Wars movies produced from 1977 to 2017 and six James Bond movies which were randomly picked from each decade from 1950 to 2010. We chose or annotated these movie series because they are likely each belongs to the same genre of movie, and therefore are likely to leave non-gender-related variables roughly constant. In this period, both U.S societies and Hollywood movies have witnessed large changes in gender equality. Thus, there should be observable changes in gender inequality trends in these movie series.

In order to understand the alignment of gender inequality in movies in human judgment and the metrics we extracted, we also collected surveys to understand how people intuitively rank the gender equality for movies within the Star Wars and James Bond series. We collected 93 surveys for James Bond and 129 surveys for Star Wars which can generate one ranking associated with gender equality for each series of movies. These two surveys were created on Google Forms and they asked many questions such as, 'Which James Bond movie has stronger female characters?' or 'Which Star Wars movie has stronger female characters?'. (Both series have strong male characters, so to keep the questions simple, we focused on the strength of female characters.) For example, in the survey we listed the following question: "Which movie presents female characters as being stronger? A New Hope or the Empire Strikes Back?" In the surveys, we listed all the possible combinations of movies. For example, we have 6 movies in the James Bond series, thus, there are 15 total possible pairwise combinations. (We thought this was easier for survey participants to answer than asking them to rank the movies.)

In order to answer these questions, we need participants who are very familiar with these movies. Thus, we found Star Wars and James Bond enthusiasts on social media platforms such as Facebook group and Reddit communities to participate in the survey.

To set up a ground truth indicating the social change of gender difference or gender bias through years, we use occupation census data to calculate the average percentage of occupation between gender for each decade (Garg et al. 2018). We extracted historical US census data in occupation to represent the change of gender inequality over time. Census data include a basic demographic survey for every person living in the US and it conducted the survey for each five years. We calculated the average occupation percentages difference between female and male in occupation for each decade from 1920 to 2010 (Ruggles et al. 2015). This data will play a role as the ground truth for the social dynamic of gender bias change over time.

- **`AOD:`** Average occupation difference between male and female per decade

  p_{women} - p_{man}:
  p_{women} = percentage of occupation that is women
  p_{men} = percentage of occupation that is men

Experiments

In order to understand the evolution of female and male's language cognitively over year, we use psycholinguistic normatives metrics to capture the underlying emotional state of

Table 1: Annotated movies from the Star Wars and James Bond series, arranged by decade of release.

Star Wars	James Bond
	Dr. No(1962)
A New Hope(1977)	Diamonds Are Forever(1971)
The Empire Strikes Back(1980)	For Your Eyes Only(1981)
Return of the Jedi(1983)	
The Phantom Menace(1999)	GoldenEye(1995)
Attack of the Clones(2002)	Casino Royale(2006)
Revenge of the Sith(2005)	
The Force Awakens(2015)	Spectre(2015)
Rogue One(2016)	
The Last Jedi(2017)	
Solo(2018)	

the speaker. These are the Ramakrishna features (Ramakrishna et al. 2015).

Psycholinguistic normatives are a series of metrics that rate individual words according to different dimensions such as "concreteness" or "pronounceability." The metrics provide a single numeric rating between -1 and 1 for every word in a given vocabulary list. This list of nonnatives have been used to measure mental status mostly in the psychotherapy field (Ramakrishna et al. 2015).

We generate four scores for each movie. First, we generate two average psycholinguistic normatives scores for each movie, one using female dialogues and one using male dialogues. We map the rating associated with the words in the dialogues for each movie and calculate the average. Thus, for each movie, there are a female dialogues rating and female dialogues rating associated with each metric. From here, we only able to calculate the rating excited in the psycholinguistic normatives dataset which contains ratings for 274596 common English words. Finally, we also calculate the actual and absolute difference between female and male dialogue ratings.

Detail of psycholinguistic normatives metrics are listed below and all the metrics are on the scale from -1 to 1:

Arousal, valence, and dominance metrics come from Affective Norms for English Words (ANEW) (Bradley and Lang 1999). These three dimensions is widely used to model the emotional status.

- **`Arousal:`** Describe the intensity of emotion o with -1 indicating the lease intensity of emotion and +1 indicating the most intensity of emotion in the text
- **`Valence:`**Describe the polarity of emotion with -1 indicating the most negative of emotion and +1 indicating the most positive of emotion in the text
- **`Dominance:`** Describe the feel of control with -1 indicating the lease feeling of control and +1 indicating the most feeling of control in the text

Concreteness, imagability, age of acquisition, and familiarity metrics originate from the MRC Psycholinguistic database (Wilson 1988) and it use to describe the cognitive status of speakers including memory and comprehension.

- **`Concreteness:`** Describe if the text can be perceived by using five senses with -1 indicating very abstract and +1 indicating very concrete
- **`Imagability:`** The degree of creating images of the word's subject with -1 indicating a word is the least likely to create mental images and +1 indicating a word is very likely to create images.
- **`Age of acquisition:`** Given a word what is the degree expected age with -1 indicating a word is closest to age 0 and +1 indicating a word is reach to the max ages.
- **`Familiarity:`** Knowledge exposure to the word with -1 indicating a word have very limited knowledge expose to and +1 indicating a word exposure to the max amount of the knowledge.

Pleasantness, pronounceability, context availability, and gender ladenness matrics originate from the Paivio, Yuille and Madigan norms (Clark and Paivio 2004). These words built on the pass cognitive metrics to reveal more associative and referential processes in the human brain.

- **`Pleasantness:`** Describe the Degree of pleasant feelings with -1 indicating the lease pleasantness and +1 indicating the most pleasantness in the text
- **`Pronounceable:`** Describe the easiness of a word is to pronounce with -1 indicating a word is the most difficult to pronounce and +1 indicating a word is the easiest to pronounce.
- **`Context Availability:`** Describe how easy to think of a specific circumstance and context when the word appears -1 indicating very hard and +1 indicating very easy.
- **`Gender Ladenness:`** Describe the degree of the feminine and masculine association of a word with -1 indicating a word mostly associated with masculine and +1 indicating a word is mostly associated with feminine.
- **`Colorado:`** Describe the number of associated word with -1 indicating a word have limited amount of association with other words and +1 indicating a word can associate with other words
- **`Paivio:`**Describe the number of associated word with -1 indicating a word have limited amount of association with other words and +1 indicating a word can associate with other words

Pearson Correlation and Spearman's rank correlation

In order to estimate how well the metrics speak the gender stereotype change. We use Pearson correlation to understand the relationship between the average occupation change between gender and the difference between gender in all features. Computing coefficient of Pearson correlation(Benesty et al. 2009) helps us understand how will the trend of the occupation change aligns with the trend of gender difference in each feature.

To study the how the result of metrics align with human judgment in gender stereotype change in the Star Wars and James Bond series, we use Superman's rank correlation(Zar

1972) to measure their relationship. Spearman's rank correlation(Zar 1972) between two variables are based on Pearson correlation between two list of rank values. In this specific project, we calculate the rank relationship between level of strength of female characters ranked by human judgment and ranked by different scores. The 'd' is the difference between the two inputted ranks in each observation and the 'n' is the number of observations (Zar 1972).

$$R_s = 1 - \left(\frac{6\Sigma d^2}{n^3 - n}\right)$$

Research pipeline

We first want to understand can the psychological metrics capture the difference between female and man's dialogues. We extract the metrics from male, female, and actual and absolute differences between male and female's dialogues. Thus, each metrics such as `Valence` is associated with four specific metrics. Then, we plotted the trend of these metrics with the average occupation difference between males and females in each decade showing in figure 1.

To test the hypothesis that some psychological metrics are significantly associated with the gender stereotype change in movies, we compute the Pearson correlation between the decade change in each metrics including male, female, and the actual and absolute difference between male and female's dialogues. (We assumed that the social stereotype change in the society is directly reflected in the movies.)

Besides looking at how the metrics associated with the big picture change in gender stereotype in the society. We want to test the power of the metrics in the new data set. In this set of data, we are also interested in understanding how the automatic prediction aligns with human judgment. We use the metrics described above to generate a rank and compute the rank correlation with the rank describing strongness of female characters generated by humans.

Results

In this section, we present the empirical results of the study. Our analysis are divided into three parts. Firstly, we use scatter plots to illustrate the trend of gender language evolution. Secondly, we estimate the correlation of word based rating metrics and the gender stereotype trend in the society. Thirdly, we study how these metrics align with human judgement using Star Wars and James Bond movie series.

Gender language evolution in movies

We found most Ramakrishna features can differentiate male and female dialogue well, though there are some exceptions. The definition of 'well' is both the trend of female's and male's metrics are not overlapping with each other in most of the period. For example, plot (a) in figure 1 describes the score of valence change over time. This plot indicates a clear difference between male and female's metrics where the 95 confidence level of these do not overlap before 1990. However, some metrics did not capture this difference between gender. For example, `PAIVIO` (plot (d) in figure 1) shows both the metrics of male and female overlapping between 1950 to 1980.

Our metrics suggest that the difference between men's and women's dialogue was small in 1930, and gradually increased to its largest point in 1960. The difference then decreased to its smallest point in 2000. For example, in the `Valence` metric the difference between male and women's dialogue started from 0.01 in 1930 and reached to 0.02 in 1960, and gradually decrease to 0.005 in 2000. Nine out of 12 metrics follow this trend. Some metrics, such as `Paivio` reach their peak in 1970, not 1960.

Significant metrics that tell the gender stereotype change

A few metrics capture the decrease of difference between female and male's dialogues. Metrics such as `Valence`, `Pleasantness`, `Age of Acquisition`, and `Gender Ladenness` show the female dialogues gradually change to the characteristic of male dialogues and become more similar to male's. We believe these metrics can potentially be used to measure gender inequality as well.

Meanwhile, These metrics also indicate both dialogues of males and females developed in a more 'masculine' trend over time. For example, in `Gender Ladenness` metrics, both female's and male's dialogues become more negative over time, indicating that both women and men are using more masculine words over time. Similarly, with `Valence`, both women and men appear to use more negative vocabulary over time.

Ultimately, since we are interested in gender *inequality*, however, it is not enough to find a correlation with `AOD`, we also want to find a *convergence* between female and male dialogue. Along these lines, we find that neither of the difference-based metrics (i.e., difference between female and male dialogue, relative or absolute value) exhibit a strong correlation with AOD from 1930–2000. All of these metrics have a p-values larger than 0.05 when we computing the Pearson correlation.

However, visual inspection of the graphs reveals an unexpected thing. Women's dialogue frequently displays a peak (or a valley) in 1960 or 1970. (Such a trend is much weaker or absent for men's dialogue.) It is possible that movie dialogue for female evolved in one direction until 1960 or so, and then changed course. One explanation for this course change could be that women's dialogue became more traditionally feminine from 1930 through the 1960s and only then became less feminized.

In fact, many of the male-female difference features have significant correlations with AOD from 1960 to 2000. We found that differences between female and male dialogue in the following metrics showed significant decreases from 1960–2000: `Concreteness`, `Imageability`, `Context Availability`, `Gender Ladenness`, `Familiarity`, `Age of Acquisition`,

Pronounceability, and Pleasantness. During this time, the metrics for women and men also move in the same direction, generally in a direction toward more 'masculine' direction. In addition, Colorado, which measures the meaningfulness of the word, also shows some convergence. In the next section, we will test if these significant metrics align with human judgement.

Alignment with human judgement

Since all of the Star Wars and James Bond movies were produced after 1960, it is meaningful to test them with the metrics extracted from 1960 onward. When looking at performance of metrics in the James Bond series, we found almost all of the convergent metrics mentioned above have high rank correlation (corr $>$ 0.6), and among them, Imageability, Age of Acquisition, Valence, and Concreteness, have the highest rank correlation (corr $>$ 0.7). Unexpectedly, one exception is Gender Ladenness (corr = 0.429), the metric that was deliberately designed to capture gender.

With the Star Wars series, however, very few metrics capture human assessments of the gender inequality in the movies. Most metrics have low correlation (corr $<$ 0.5). One exception is Concreteness (corr = -0.782). Again and unexpectedly, Gender Ladenness shows low rank correlation (corr = -0.370).

Discussion

We evaluated which of a number of Ramakrishna linguistic features(Ramakrishna et al. 2015) is most effective in determining changes in gender inequality in movie dialogues. Through evaluating how well these linguistic measure associate with gender inequality indicators (AOD) in the Cornell movie-dialogues corpus(Danescu-Niculescu-Mizil and Lee 2011), we found that Concreteness, Imageability, Context Availability, Gender Ladenness, Familiarity, Age of Acquisition, Pronounceability, and Pleasantness are significant in measuring gender inequality in movie dialogues. Then, we tested these metrics against the James Bond and Star Wars movies data sets. We found that most of the metrics successfully measured gender inequality in the James Bond movies, but not in the Star Wars movies.

Some of our results challenge previous findings. For example, previous studies have proven that Gender Ladenness metrics in particular are sensitive in predicting gender difference in language (Narayanan et al. 2019). While we find that Gender Ladenness shows convergence in movies over time, it does not do especially well in capturing human assessments of gender differences in movies. It is likely that gender dynamics in movies that goes beyond single-word choices is a factor.

In addition, while others have found that word-based metrics not specifically focused on gender also reveal gender differences (Schofield and Mehr 2016), we find our results differ somewhat from that work. Past work, for example, suggests that Arousal will differentiate between female and male dialogue (Schofield and Mehr 2016). It is also a stereotype that females are always associated with stronger emotions and use more emotional language (Newman et al. 2008). However, our results show that both the dialogues of males and females have similar emotional intensity over time. One interpretation is that the difference of emotional intensity between genders is smaller in movies because both genders need to present relative high emotional intensity to fit the needs of drama. Ultimately, the reason why non-gender-specific metrics can also measure gender differences is still unclear.

Male-female difference features including Concreteness, Imageability, Context Availability, Gender ladenness, Familiarity, Age of Acquisition, Pronounceability, and Pleasantness show significant correlation with AOD after 1960. Through looking into how female and male dialogues change in these metrics, we found that the majority of metrics become 'more male' in both dialogues. Both male and female use more sensory, imageable, masculine, and less sophisticated words in movies over time. In some metrics, female dialogues changed in the male direction, while male dialogues changed in the female direction. Examples of this are Context Availability and Pronounceability. From 1960, females' used words that are more difficult to pronounce and to think of in specific circumstances, while males used words that are easier to pronounce and think of context. In Pleasantness metrics, male dialogues do not change much, while female dialogues come to resemble male dialogues. In other words, female use less pleasant words over time. These changes may reflect changes in the language use of the larger society, or merely changes in how movie dialogue is written. For example, it is possible that scriptwriting overall has converged toward shorter, Anglo-Saxon word choice over time.

We also found a lot of results suggesting that the 1960–1969 period seems to present a peak of most female metrics and a turning point in how closely female metrics resemble male metrics. This also caused most our metrics to align with gender inequality trends after 1960 but not before 1960. These results seem to respond to the movie industry's turning points in gender equality due to the second-wave feminism movement. Film study scholars have discussed movie production adapted to the second wave feminism movement that happened around 1960 (Benshoff and Griffin 2011). The '70s was marked as a turning pointing in the Hollywood movie industry since Hollywood began to employ female as directors of movies (Benshoff and Griffin 2011). We also found male and female metrics to have some overlap before 1960. However, this overlap is not found after 1960. The above reasons might also explain this phenomena, but in-depth causal inference studies or literature studies is needed to answer this question.

There are some limitations to our study. We need to explain how the trend of language metrics changes with gender inequality changes. Currently, we use model-free evidence from data visualization. This method is not as robust since we did not determine the meaningfulness of the increase and decrease. For example, female Valence metrics decreased from around 0.360 to 0.345 and the scale

of metrics spans from -1 to 1. One question raised here is if the decrease from 0.360 to 0.345 in `Valence` the metrics is meaningful. Furthermore, we primarily apply a linear model to measure the correlation between language metrics and gender inequality indicators (AOD). The relationship between language change and the gender inequality indicators might be non-linear. However, the primary focus of our study is to determine the metrics that can measure or detect gender inequality well. Linear models are sufficient to show the association between language metrics and gender inequality indicators (AOD) or language metrics and human judgment indicators consistently. Another drawback of this study is the size of the datasets. We only have around 500 availables movies from 1930 to 2000, which is considered a relatively small dataset. However, due to the randomized nature of the collection within this dataset, and the fact that this dataset is the largest available movie dataset with gender and intereaction labels, we believe this data can provide significant language training data. The gender inequality indicators we used, which is the occupational difference in change over time between males and females, can successfully reflect gender inequality norms in society. Many successful studies use this approach to represent gender bias or stereotype change over time (Garg et al. 2018). However, this might not an entirely accurate appraisal of the movie industry.For instance, movies might delay presenting social norm. However, since our gender inequality indicator is mostly decreases gradually over time, the lag effect won't change correlations between the trends significantly. Finally, another limitation of this study is that we only evaluate lexical features which Ramakrishna produced (Ramakrishna et al. 2015). Other important linguistic features are also worthy of evaluating, for example, emotional status lexicon, word embedding, and count-based metrics. We will also measure these features in a future study.

Conclusion

We evaluate which of a number of Ramakrishna linguistic features(Ramakrishna et al. 2015) is most effective in determining changes in gender inequality in movie dialogues. Through evaluating how well these linguistic measure associate with gender inequality indicator (AOD) in the Cornell movie-dialogues corpus(Danescu-Niculescu-Mizil and Lee 2011), we found `Concreteness`, `Imageability`, `Context Availability`, `Gender Ladenness`, `Familiarity`, `Age of Acquisition`, `Pronounceability`, and `Pleasantness` are significant in measuring gender inequality in movies dialogues. These metrics can use to check gender inequality trend automatically in a simple manner, while provide applicable case study to understand language evolution using lexicons. This approach can also use to understand how other linguistic features, such as embedding, word count, and other lexicon features, associate with gender inequality change.

References

Bechdel, A., et al. 2010. Dykes to watch out for. *Artlink* 30(2):70.

Benesty, J.; Chen, J.; Huang, Y.; and Cohen, I. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*. Springer. 1–4.

Benshoff, H. M., and Griffin, S. 2011. *America on film: Representing race, class, gender, and sexuality at the movies*. John Wiley & Sons.

Bradley, M. M., and Lang, P. J. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology ....

Burger, J. D.; Henderson, J.; Kim, G.; and Zarrella, G. 2011. Discriminating gender on twitter. In *Proceedings of the conference on empirical methods in natural language processing*, 1301–1309. Association for Computational Linguistics.

Clark, J. M., and Paivio, A. 2004. Extensions of the paivio, yuille, and madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers* 36(3):371–383.

Danescu-Niculescu-Mizil, C., and Lee, L. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115(16):E3635–E3644.

Kozlowski, A.; Taddy, M.; and Evans, J. 1803. The geometry of culture: analyzing meaning through word embeddings. 2018. *arXiv preprint arXiv:1803.09288*.

Lauzen, M. M., and Dozier, D. M. 2005. Maintaining the double standard: Portrayals of age and gender in popular films. *Sex Roles* 52(7):437–446.

Malandrakis, N., and Narayanan, S. S. 2015. Therapy language analysis using automatically generated psycholinguistic norms. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Mukherjee, A., and Liu, B. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*, 207–217. Association for Computational Linguistics.

Mulac, A., and Lundell, T. L. 1994. Effects of gender-linked language differences in adults' written discourse: Multivariate tests of language effects. *Language & Communication* 14(3):299–309.

Mulac, A.; Wiemann, J. M.; Widenmann, S. J.; and Gibson, T. W. 1988. Male/female language differences and effects in same-sex and mixed-sex dyads: The gender-linked language effect. *Communications Monographs* 55(4):315–335.

Mulac, A.; Seibold, D. R.; and Farris, J. L. 2000. Female and male managers' and professionals' criticism giving: Differences in language use and effects. *Journal of Language and Social Psychology* 19(4):389–415.

Narayanan, S.; Palacios, V. M.; Ramakrishna, A.; Somandepalli, K.; Malandrakis, N.; and Singla, K. 2019. Linguistic analysis of differences in portrayal of movie characters. US Patent App. 16/137,091.

Neville, C., and Anastasio, P. 2018. Fewer, younger, but increasingly powerful: How portrayals of women, age, and power have changed from 2002 to 2016 in the 50 top-grossing u.s. films. *Sex Roles* 80.

Newman, M. L.; Groom, C. J.; Handelman, L. D.; and Pennebaker, J. W. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes* 45(3):211–236.

Peersman, C.; Daelemans, W.; and Van Vaerenbergh, L. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, 37–44.

Ramakrishna, A.; Malandrakis, N.; Staruk, E.; and Narayanan, S. 2015. A quantitative analysis of gender differences in movies using psycholinguistic normatives. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1996–2001.

Ruggles, S.; Genadek, K.; Goeken, R.; Grover, J.; and Sobek, M. 2015. Integrated public use microdata series: Version 6.0 [dataset]. *Minneapolis: University of Minnesota* 23:56.

Sboev, A.; Litvinova, T.; Gudovskikh, D.; Rybka, R.; and Moloshnikov, I. 2016. Machine learning models of text categorization by author gender using topic-independent features. *Procedia Computer Science* 101:135–142.

Schofield, A., and Mehr, L. 2016. Gender-distinguishing features in film dialogue. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, 32–39.

Thomson, R., and Murachver, T. 2001. Predicting gender from electronic discourse. *British Journal of Social Psychology* 40(2):193–208.

Wedding, D., and Niemiec, R. M. 2014. *Movies and mental illness: Using films to understand psychopathology*. Hogrefe Publishing.

2020. Bechdel test.

Wilkie, J. R. 1993. Changes in us men's attitudes toward the family provider role, 1972-1989. *Gender & Society* 7(2):261–279.

Wilson, M. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers* 20(1):6–10.

Zar, J. H. 1972. Significance testing of the spearman rank correlation coefficient. *Journal of the American Statistical Association* 67(339):578–580.

wordtype	corr	p
AROUSAL	-0.799	0.017
AGE OF ACQUISITION	-0.772	0.025
CONTEXT AVAILABILITY	-0.662	0.073
PRONOUNCEABILITY	-0.210	0.618
FAMILIARITY	0.155	0.715
MEANINGFULNESS (COLORADO)	0.219	0.603
CONCRETENESS	0.261	0.532
IMAGABILITY	0.325	0.432
DOMINANCE	0.451	0.262
VALENCE	0.629	0.095
PLEASANTNESS	0.644	0.085
MEANINGFULNESS (PAIVIO)	0.735	0.038
GENDER LADENNESS	0.753	0.031

Table 2: Pearson correlation in linguistic metrics scores over time in **female's dialogues**. Corr column contains the coefficient and p column contains the p-value.

wordtype	corr	p
AGE OF ACQUISITION	-0.827	0.011
AROUSAL	-0.816	0.014
CONTEXT AVAILABILITY	-0.342	0.407
MEANINGFULNESS (COLORADO)	-0.018	0.965
PRONOUNCEABILITY	0.034	0.937
CONCRETENESS	0.067	0.874
IMAGABILITY	0.096	0.821
FAMILIARITY	0.338	0.412
DOMINANCE	0.468	0.242
MEANINGFULNESS (PAIVIO)	0.637	0.090
GENDER LADENNESS	0.710	0.048
PLEASANTNESS	0.788	0.020
VALENCE	0.827	0.011

Table 3: Pearson correlation between linguistic metrics score and the average percentage of difference from 1930 to 2000 in **male's dialogues**. Corr column contains the coefficient and p column contain the p-value.

wordtype	diff_corr	p_val
AROUSAL	-0.234	0.577
CONCRETENESS	-0.231	0.582
IMAGABILITY	-0.205	0.626
CONTEXT AVAILABILITY	-0.189	0.654
PRONOUNCEABILITY	-0.178	0.674
FAMILIARITY	-0.132	0.755
PLEASANTNESS	0.110	0.796
VALENCE	0.179	0.671
GENDER LADENNESS	0.212	0.614
AGE OF ACQUISITION	0.219	0.602
DOMINANCE	0.239	0.568
MEANINGFULNESS (PAIVIO)	0.503	0.204
MEANINGFULNESS (COLORADO)	0.530	0.176

Table 4: Pearson correlation between linguistic metrics score and the average percentage of difference in **actual difference between female and male's score** from 1930 to 2000. Corr column contain the coefficient and p column contains the p-value.

wordtype	fm_corr	fm_p_val
AGE OF ACQUISITION	-0.219	0.602
CONTEXT AVAILABILITY	-0.189	0.654
PRONOUNCEABILITY	-0.178	0.674
FAMILIARITY	-0.132	0.755
AROUSAL	-0.128	0.762
PLEASANTNESS	0.110	0.796
VALENCE	0.179	0.671
GENDER LADENNESS	0.212	0.614
DOMINANCE	0.239	0.568
CONCRETENESS	0.281	0.499
MEANINGFULNESS (PAIVIO)	0.285	0.494
MEANINGFULNESS (COLORADO)	0.351	0.394
IMAGABILITY	0.356	0.386

Table 5: Pearson correlation between linguistic metrics score in **absolute difference between female and male's score** and the average percentage of difference from 1930 to 2000. Corr column contains the coefficient and p column contain the p-value.

wordtype	corr(actual)	p(actual)	corr(absolute)	p(absolute)
CONCRETENESS	-0.966	0.007	0.966	0.007
IMAGABILITY	-0.956	0.011	0.956	0.011
AGE OF ACQUISITION	-0.894	0.041	0.894	0.041
MEANINGFULNESS (PAIVIO)	-0.640	0.244	0.640	0.244
MEANINGFULNESS (COLORADO)	-0.279	0.649	0.910	0.032
AROUSAL	0.577	0.308	0.519	0.370
DOMINANCE	0.813	0.095	0.813	0.095
VALENCE	0.827	0.084	0.827	0.084
PLEASANTNESS	0.881	0.048	0.881	0.048
PRONOUNCEABILITY	0.884	0.047	0.884	0.047
FAMILIARITY	0.902	0.036	0.902	0.036
GENDER LADENNESS	0.927	0.024	0.927	0.024
CONTEXT AVAILABILITY	0.940	0.017	0.940	0.017

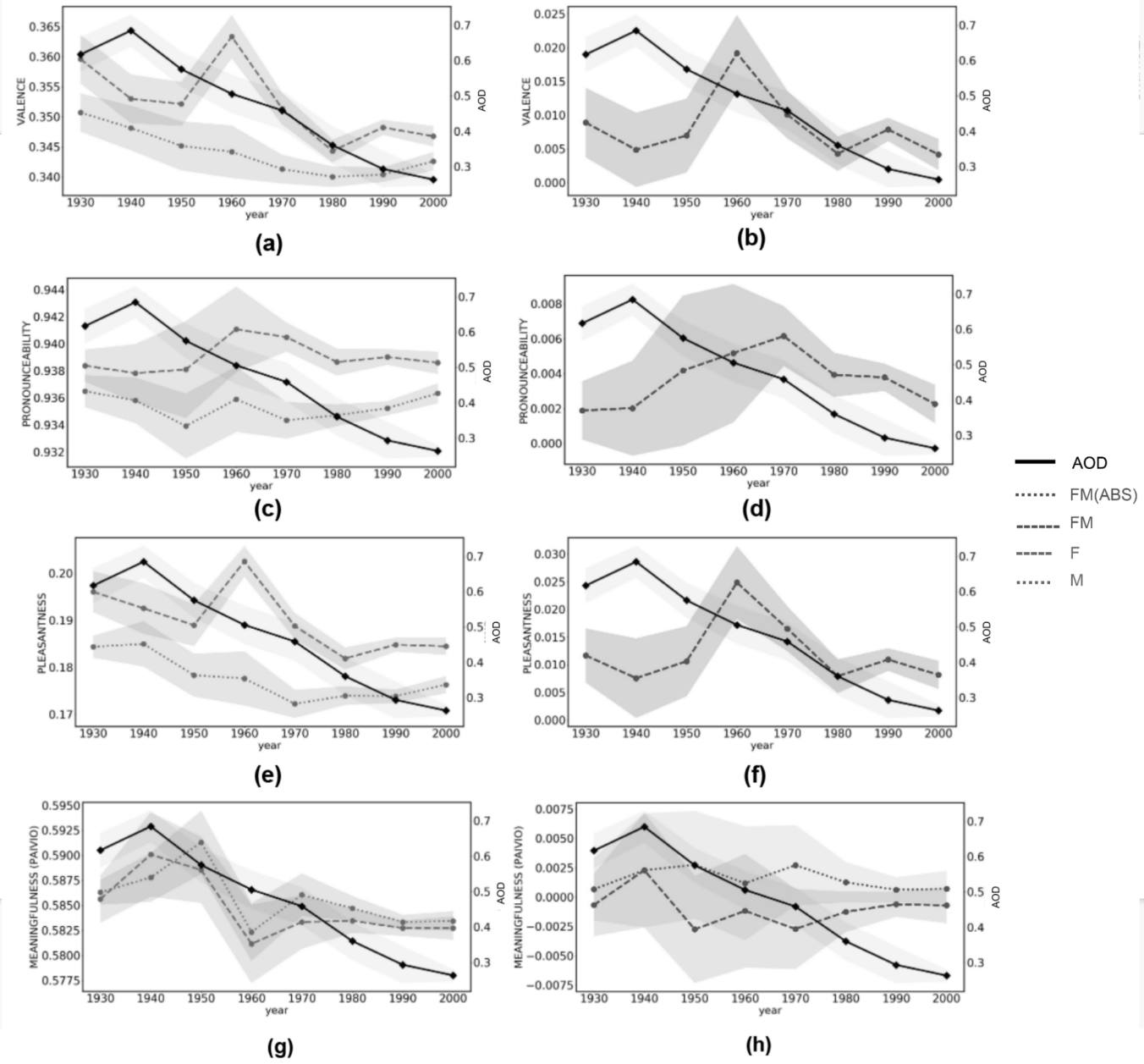
Table 6: Pearson correlation between linguistic metrics score from 1960 to 2000 in **absolute and the actual difference between female and male's score** and the average percentage of difference. Corr and p column contain the coefficient and the p-value. (actual) indicate the correlation is associated with the actual difference between female and male's score and (absolute) is associated with the absolute difference between female and male's score.

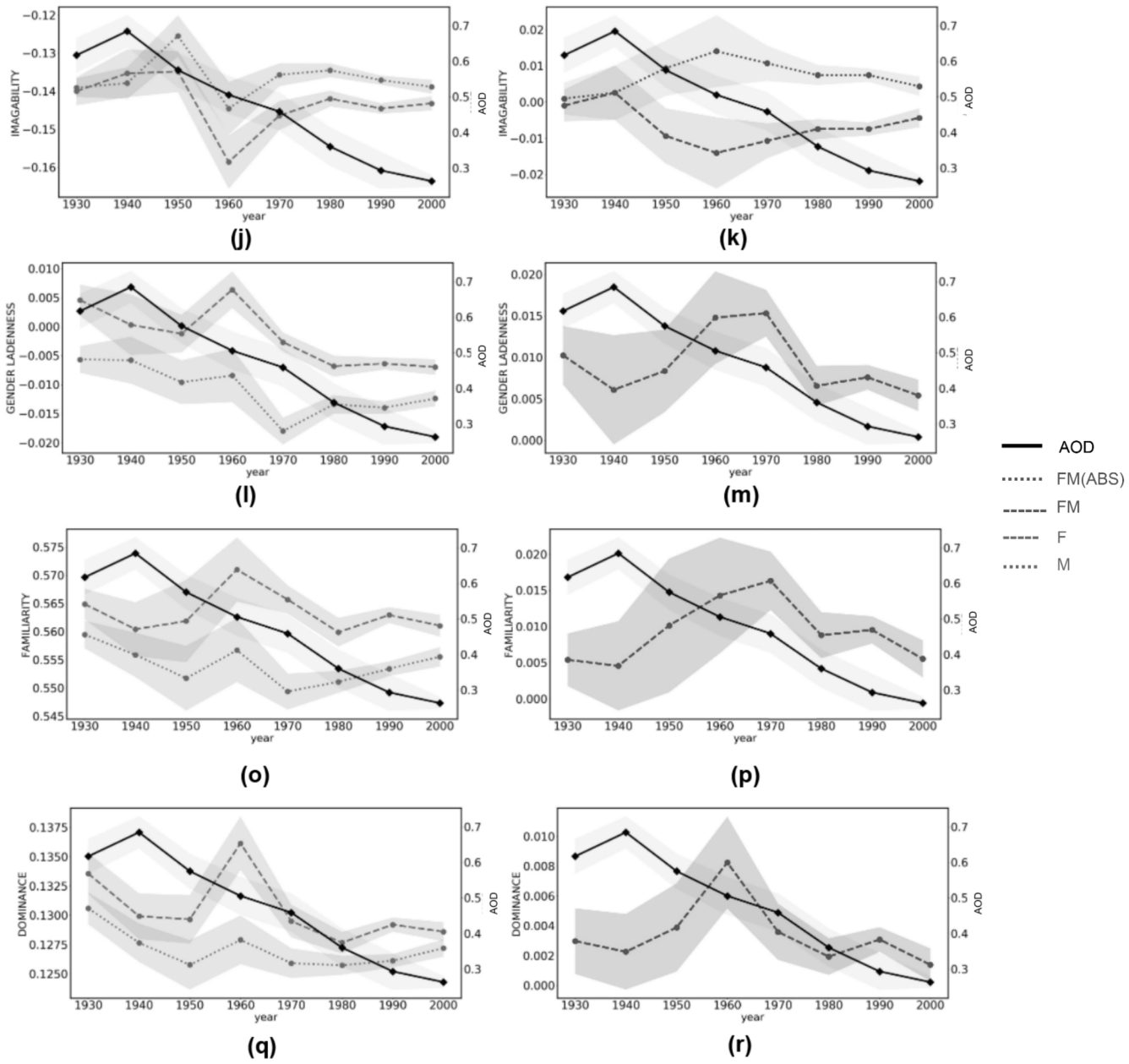
	diff(abs)	female	male
IMAGABILITY	0.829	-0.486	0.086
AGE OF ACQUISITION	-0.829	0.771	0.029
VALENCE	-0.771	-0.257	0.143
CONCRETENESS	0.714	-0.257	0.029
CONTEXT AVAILABILITY	-0.600	-0.771	-0.143
FAMILIARITY	-0.600	-0.543	-0.143
PLEASANTNESS	-0.600	-0.086	0.143
PRONOUNCEABILITY	-0.600	-0.543	-0.029
DOMINANCE	-0.543	0.029	0.143
GENDER LADENNESS	-0.429	-0.314	-0.257
MEANINGFULNESS (COLORADO)	-0.429	0.029	-0.257
AROUSAL	0.143	0.486	0.943
MEANINGFULNESS (PAIVIO)	-0.143	-0.657	0.029

Table 7: Spearman's Rank correlation results between human judgment and linguistics metrics on the strongness of female characters on **James Bond** Movies. The columns, diff(abs), are the metrics score difference (absolute value) between female and male's dialogues. Female and male columns represent metrics score from female and male's dialogues.

	diff(abs)	female	male
IMAGABILITY	-0.224	-0.527	-0.733
AGE OF ACQUISITION	-0.067	0.152	-0.115
VALENCE	-0.442	-0.539	-0.612
CONCRETENESS	-0.782	0.176	-0.297
CONTEXT AVAILABILITY	-0.345	0.055	0.127
FAMILIARITY	-0.188	0.782	0.830
PLEASANTNESS	-0.406	-0.297	-0.164
PRONOUNCEABILITY	-0.503	-0.636	-0.685
DOMINANCE	-0.164	0.818	0.709
GENDER LADENNESS	-0.370	-0.479	-0.624
MEANINGFULNESS (COLORADO)	-0.309	-0.248	-0.030
AROUSAL	-0.248	0.442	0.067
MEANINGFULNESS (PAIVIO)	-0.418	0.527	0.770

Table 8: Spearman's Rank correlation results between human judgment and linguistics metrics on the strongness of female characters on **Star War** Movies. The columns, diff(abs), is the metrics score(absolute value) difference between female and male's dialogues. Female and male columns represent metrics score from female and male's dialogues.


VALENCE
AOD
year
(a)
(b)
PRONOUNCEABILITY
(c)
(d)
PLEASANTNESS
(e)
(f)
MEANINGFULNESS (PAIVIO)
(g)
(h)
AOD
FM(ABS)
FM
F
M


IMAGABILITY
AOD
year
(j)
(k)
GENDER LADENNESS
(l)
(m)
FAMILIARITY
(o)
(p)
DOMINANCE
(q)
(r)
AOD
FM(ABS)
FM
F
M

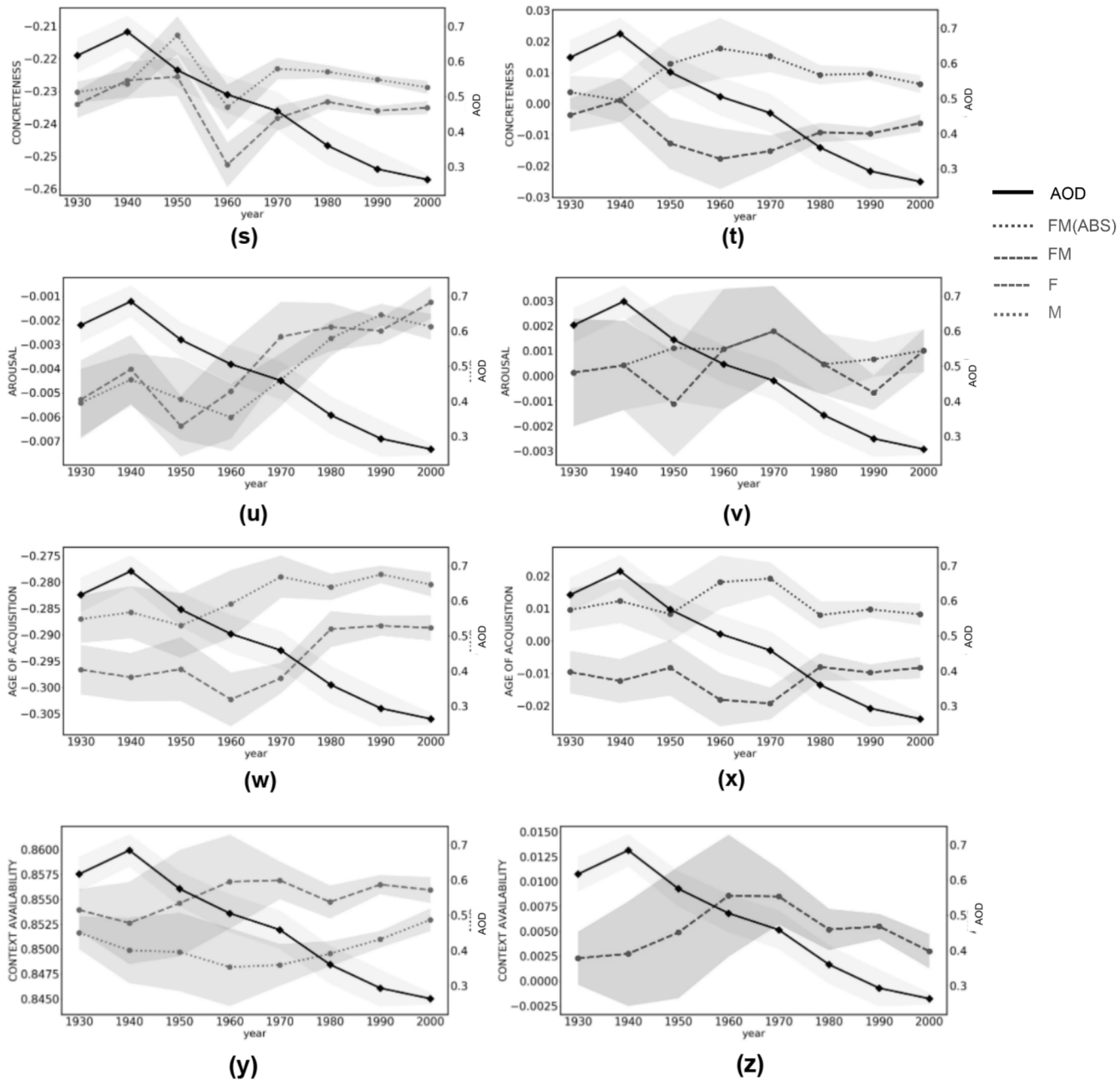


Figure 1: Average linguistic metrics score over time in male and female's dialogues vs. the average percentage of difference. Meaning of positive or negative value in linguistic metrics score is associated with specifics metrics. (check data description). The dotted grey link represented linguistic metrics score extracted from male's dialogues over time and the dashed grey link is from female dialogues. The black solid line is the average percentage of difference in occupation between gender. Each shaded region is the SE interval.