A Statistical Analysis of Bias in Film

Sara Jensen Brett Grimes Ryan Mahoney
Zachary Erickson

October 6, 2016

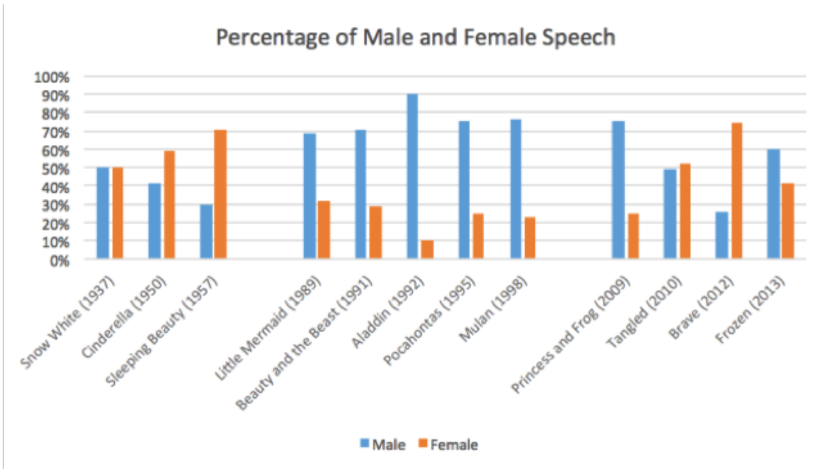
Abstract

We investigate many types of potential bias in films via their scripts. Most results concern gender bias as defined by number of lines spoken by characters. In addition to a random sample, we look more closely at childrens films and films nominated for Academy Awards. Preliminary results are also presented regarding ageism, racism, and sizeism.

1 Introduction

Criticism of the movie industry, mainly concerning racism and sexism, has been a recent discussion among mainstream media as well as in some academic circles; just a few of these articles include [2], [7], and [5]. Mostly, these studies have been conducted by those in humanities fields such as linguistics and gender studies, or they are popular articles written by journalists. The prevailing rhetoric is that older white males dominate film. However, most of the discussion begins and ends with that rhetoric; supporting data is not merely ignored, but rather significant information has been largely unavailable to be analyzed. The purpose of this study is not to objectively confirm or deny any claims of racism, sexism, etc., but rather to provide real, unbiased data in order that the larger community may base their claims on quantitative, in addition to the perceived qualitative, information. To achieve this aim we set out to create a series of programs that can efficiently and accurately analyze movies as well as gather data on select actors and actresses.

Very few similar studies have been previously conducted, but one of note is a study entitled *A Quantitative Analysis of Gendered Compliments in Disney Princess Films.* This research was conducted by professor of linguistics Carmen Fought and graduate student Karen Eisenhauer. The surprising results of the study, which in large part inspired our current research, found that while most Disney films star women, “they are populated overwhelmingly by male characters” (Fought & Eisenhauer, 2016). Their paper cited earlier studies by DeRozario (2004) and Baker-Sperry (2007) to support the importance of this information, stating how they found that children use films to construct their ideologies about language and gender, including gender identity. Fought and Eisenhauer examined two sets of data, percentage of lines spoken by characters (broken down by gender) and the quality of “compliments” given to female characters. Our study attempts to replicate the former data set on a much larger and broader (i.e. films across all genres) scale.



Data courtesy of Karen Eisenhauer and Carmen Fought.

Figure 1: A figure taken from [4].

The importance of studies such as ours and Fought & Eisenhauer's cannot be overstated. Film-making is, at it's most basic, an art of storytelling. If these stories do not properly reflect the truths of reality, then they can negatively affect the way society perceives the world. Discussing their findings in [4], Eisenhauer states, "It says a lot about what we're saying about femininity."

It is important to note the emphasis on quantitative over qualitative data in our study. We almost entirely ignore a character's role in a film. That is, we do not take into account important but mathematically indefinable qualities such as cultural context within a film, personality traits of characters, and "strength" of dialogue within a film, which is present in the Fought and Eisenhauer study. We do, however, provide a mathematical definition of "main" or "leading" character(s) in order to infer some of these qualities which can be found below in Subsection 2.1.

Section 2 contains four subsections. The first provides the necessary definitions for various parts of our study. The subsequent subsections provide details of each of the three computer programs used to analyze films. Section 3 contains numerous subsections. Our main result concerns percentages of lines categorized as being spoken by a male, female, or unidentifiable gender in a random sample of films, accounting for film genre. The remaining results vary, but a brief description of the type of analysis is included in the subsection title. It should be noted that all of our work is being presented, regardless of whether or not the statistical test showed a significant result. In Section 4, we discuss the various shortcomings of our research, while Section 5 provides just some of the many avenues for continued research.

2 Methodology

As stated in Section 1, our aim is to analyze movies efficiently in order to gather as much data on them as possible. The most reliable way to study the dialogue spoken throughout a film would be to watch the movie and record how often the characters speak. However, most films take far too long to personally watch for data collection; collecting data in this way would limit the number of films we would be able to study. We decided instead to find the film scripts and use those to analyze the dialogue.

A huge benefit of using a film's script is the ability to code a computer to analyze the data present in the film. Our initial research goal was to count the number of lines in a film spoken by each gender. Given a movie script, a computer is able to perform this task in seconds, allowing us to analyze an unprecedented number of films. Subsection 2.1 provides the precise definition of a line, and it also provides more descriptions of the type of data being collected by the computer from the movie scripts.

Regarding our choice of programming language and computer software, we needed a quick, relatively portable, flexible coding language. Each researcher is familiar with C++, and because it also met the requirements of the project, we decided to use it as our coding language. We were able to use `http://c9.io`, a web based IDE, to work on the code together without worrying about losing our individual changes. We also used Microsoft Visual Studio, various editions.

2.1 Definition of Terms

Much of our analysis depends on counting the number of lines or words spoken by characters in a movie script. Although a word is a standard definition, our definition of a line must be made precise.

Definition 1. We consider a *line* any amount of spoken dialogue by a single character until either another character speaks or stage direction takes place.

Note that by this definition, both a single word and several sentences constitute a line. While it may initial seem unfair or inaccurate for these to carry the same weight, we have reason to believe that this does not significantly affect our results. (see Section 3, where we compare percentage lines and percentage words). While long monologues and soliloquies may be iconic for certain films, they are not nearly as common as one might think. We should also note that occasionally, scripts have a character speaking, a pause for stage directions, and then the character continues to speak. This situation register as two lines spoken by that character; the initial line and the continued line.

The format of most scripts allows us to easily determine lines as most scripts alternate stage direction and dialogue. For more information on how our program is able to make this distinction, see Subsection 2.2.

We now define the different categorizations of characters used in our analysis and also in common speech. Although the definition of a bit character is widely accepted in the film industry (it was introduced as a category of actor by SAG-AFTRA[3]), the definition of a main character in the literature is ambiguous; we provide our own definition here.

Definition 2. A character who speaks 6 or fewer times is called a *bit character*.

Definition 3. A character who speak more lines (or words) than the arithmetic average of the top 5 characters' lines (or words) is considered to be a *main character*.

Definition 3 is a result of an algorithm we created to determine main characters. This allows flexibility in terms of number of main characters. For example, in a film such as *Forrest Gump* where one character overwhelmingly dominates the dialogue, this definition gives one main character. Conversely, a film such as *The Avengers* may have a group of main characters with roughly equal speaking shares. In this case, 3 or possibly even 4 main characters may arise. Note that the character in the film with the most lines (or words) is automatically a main character by this definition.

It should also be made clear that a character who plays a major role in terms of the plot is not guaranteed to be a main character by this definition. A good example would be Princess Buttercup (Robin Wright), whom the title of the film (*The Princess Bride*) references. According to our definition, she does not qualify as a main character.

Finally, we define a supporting character.

Definition 4. A character who has more than 6 lines but do not qualify as a main character is considered a *supporting character.*

For those who find these sort of things interesting: Frodo Baggins is, by our definition, a supporting character in both *The Two Towers* and *The Return of the King* (we have not yet analyzed *The Fellowship of the Ring*). Interestingly, this view of Frodo is accepted by many Lord of the Rings fans as well as Tolkien himself, who considered Sam to be the true hero of the story[9].

Lastly, we introduce some terminology that will be present in Section 3.

Definition 5. We say that a film is *male dominated* if more than 50% of the cast is male. Similarly, we say a film is *female dominated* if more than 50% of the cast is female. In the event that exactly half of the case is male, we say that a film is *balanced.*

Also in Section 3, we will analyze other ways in which films can be male/female dominated, such as *main male dominated* or *male line dominated.* Keeping with the spirit of definition 5, a movie is said to be *main male dominated* if at least 50% of the main characters in the film are male. In the event that a movie is *main balanced*, we break the tie by looking at the actual percentages of lines spoken and percentage of words spoken by the main characters. The gender of the character with the greater percentage of lines and words spoken is attributed to this film. In the event of a discrepancy between greater line and greater word percentage, the main dominated characteristic is given the gender attributed to the greater percentage.

For example, consider the film *Sweeney Todd: The Demon Barber of Fleet Street.* According to our data, both Johnny Depp (Sweeney Todd) and Helena Bonham Carter (Mrs. Lovett) are main characters in the film. The remaining information can be summarized in the following table.

Character	% Lines	%Words
Sweeney Todd	33.48	25.77
Mrs. Lovett	29.74	36.68

As Sweeney Todd has a higher percentage of lines in the film but Mrs. Lovett has a higher percentage of words, we must look at the difference between percentage of lines and percentage of words. Because the difference in percentage of words is greater than the difference in percentage of lines, we say that *Sweeney Todd: The Demon Barber of Fleet Street* is a main female dominated film. In the event that a movie was perfectly balanced in some component, we would identify it this way (e.g. the movie is main balanced).

2.2 Script Data Collection Program

The first step in our process involves gathering scripts for the movies we wish to analyze. Scripts, also called screenplays, are usually easy to find on the internet. While not every studio releases their scripts, there are hundreds that can be found on websites like `IMSDB.com` and `SimplyScripts.com`. We locate the script of our desired film and check to make sure it follows the format of a “standard” script. Most scripts are written so that each character’s name is fully capitalized when he/she speaks, and his/her line of dialogue is ended, either by stage directions or a change in dialogue, by a blank line. Figure 2.2 gives an example of a script formatted in this way, demonstrating both ways to end a line of dialogue.

```
                    YOUNG ANNA (CONT'D)
            This is amazing!

                    YOUNG ELSA
            Watch this!

  Elsa stomps her little slippered foot and a layer of ice
  suddenly coats the floor, forming a giant ice rink. Anna
  slides off, laughing.
```

Figure 2: Excerpt Taken from *Frozen*, script available from `IMSDB.com`.

Using this format, we are able to separate one character’s lines from those of other characters and stage directions.

After the script has been found and saved to a shared Dropbox folder (`http://dropbox.com`), we go to the film’s IMDB (Internet Movie Database) profile on `http://IMDB.com` and find the full cast list. We then copy the cast list from the website and saved it as a plain text file (.txt). When our program reads this file, the structure of the text document is updated so that each actor/actress’s name is followed by the name of his or her character in the film.

Our program reads each line of text in the script and determines if a character is speaking. In the standard format of scripts, the character’s name is always on a line by itself in all capital letters. Therefore, when the program gets to a line that meets those criteria it stores the line as a “potential character”. It then reads the following lines of text as dialogue, terminating at a blank line. The dialogue is stored as a vector of strings. From this vector we determine number of lines and count the number of words spoken. If the program comes across a character name that has already been found, it adds the proceeding lines of dialogue to that character.

As demonstrated in Figure 2.2, one complication is that stage directions are occasionally included in a character name. That is, the *Frozen* script has potential characters “YOUNG ANNA” and “YOUNG ANNA (CONT’D)”. Our program removes any parenthetical comments from a potential character’s name before searching for the character, this associating “YOUNG ANNA” and “YOUNG ANNA (CONT’D)”, or in the case of voice overs in *Forrest Gump*, “FORREST” and “FORREST (V.O.)”. Our program does not associate “YOUNG ANNA” and “ANNA” due to the potential to mistakenly associate different characters (as is the case in *Forrest Gump*, where “YOUNG FORREST” is truly Forrest’s son).

After we read through the script and gather a list of all the speaking characters, we read in the cast list. Here the program links the actor/actress’s name to the official character’s name. We do this by taking the name of the character according to the script and then trying to find its match using the official names. Most of the time the script name is a substring of the official name (e.g. “FORREST” in the script and “FORREST GUMP” on `http://IMDB.com`), so by comparing the strings to see if the script name can be found within the official character’s name, the program can usually link the two names together. However, we had to set some limitations to this automated linking. If the program is unable to find a link then the user is prompted to input who the character is supposed to be, given a list of possible names from IMDB. If, as is the case for a few characters in every movie, an actor or actress cannot be determined, the user may give the scripted character one of the following attributes: Unidentified Male, Unidentified Female, or Unidentified Neutral. The Unidentified Neutral is used only in instances where:

- The program is unable to automatically link the character with the actor/actress.
- The user is unable to determine the actor/actress that played the character.
- the script uses an unspecific name like “TROLL” or “CAB DRIVER” for a character.

As is shown in Section 3, unidentified neutral characters are quite rare.

At this point the program knows what actors and actresses appear in the film, which characters they play, and the lines that they speak.

Next came the process of assigning each one of these actors/actresses a gender. We sought a way to do this automatically, as each new film usually added 50 new actors to our database. We put together two lists of the top 100 gender specific names (removing any redundancies e.g. Taylor) and then ran the names of the actors across the corresponding gender list. That method made gender assignment almost entirely automatic. If we are unable to find the actor/actress’s name in a name list, the program prompts the user to enter a gender. If we determined that this new name is also gender specific, just unpopular, the program allows us to append the name to the list so that the actor/actress may have his/her gender automatically assigned if he/she appear again.

After a film has been analyzed the data we gather from it is exported to a series of files for the next steps of analysis. Each movie has its information exported to a plain text file (.txt) and a comma separated values sheet (.csv). Microsoft Word and Excel allow us to compare the data from each of these files and begin gathering statistics on the sample (see Section 3 for more information on our findings).

2.3 Actor Data Program

We sought to study other types of bias in films, so we began work on two additional studies using other programs. One study aimed to analyze the characteristics of each main character (See Definition 3) and the other analyzed the characteristics of the movies.

The actor study program begins by reading in the file containing all of the results on each movie, compiled by the program discussed in Subsection 2.2. The file is read line by line until it reaches the actor/actress information on the sheet. From there it is able to gather information on the actor's name, gender, the number of movies he/she has been in, words spoken, lines spoken, and percentage of lines and words in each movie. Afterwards the user is prompted to begin entering in additional information on the particular actor or actress, including his or her age, height, weight, net worth, and race(s).

This process requires a great deal of legwork because most actors are not absolutely open with characteristics like weight and net worth. Using the internet, we are able to gather most of the information we need; however we have no standard source for the information. Additionally, it becomes increasingly difficult to gather information on an actor/actress the longer he/she has been gone. Surprisingly, race is often an easier characteristic to gather information on, despite the fact that any individual may be more than one race. After gathering all the available information on an actor/actress the data is exported to a file where we keep a comprehensive list.

2.4 Movie Analysis Program

The movie analysis program deals with filtering and analyzing the data that has been output by the primary program discussed in Subsection 2.2. This program begins by reading in the csv file that contains the data on all of the films studied. This includes the title of the movie, the year it was made, the director(s), the number of Academy Award nominations the film received, the genres of the film, and the data pertaining to the characters. This data includes the names of the characters, the genders of the characters, the percentage of lines and words spoken by the characters, and the categorization of characters as main, supporting, or bit characters.

After reading in and storing data, this program asks the user if he or she would like to sort or filter the data. If the user chooses to sort the data, the user may do this by the film's year, genre, or by number of Academy Award nominations. If the user chooses to filter the data, he or she must select a filtering option from the menu below.

1. Genre
2. Director
3. Actor

4. IMDB top 250
5. Disney Movies
6. Academy Award Movies
7. Female Leads
8. Random Sample
9. By List

After an option is chosen, an appropriately named csv file is created and only the movies that fit the selected category are output to that file. For example, the Female Leads option above filters only movies that have a female main character. The random sample option selects a sample from all of the films that is balanced with respect to representing the genres of the films. More detail about this selection process can be found in Subsection 3.2. The final option allows the user to enter a text document containing a list of movies and the program finds all movies in our comprehensive database that also appear on the list. This allowed us to study other subsets of our movies, such as movies with a rating of PG or below and movies nominated for an Academy Award between 2013 and 2015.

3 Results

3.1 Lines vs. Words

In our analysis of the data, we will almost exclusively be referencing the percentage of lines spoken by characters, rather than words. There are several reasons for this, the first of which concerns the accuracy of counting number of words. While most scripts clearly differentiate between dialogue and stage direction, there are some that do not. That is, our program would add stage direction to a character's lines, which would not affect line count, but would affect word count. However, since we take each character relative to their own movie, percentage-wise things should, for lack of a better term, "even out". In fact, when we run a t-test comparing the sample of line percentages and word percentages, we get an insignificant t-score of 0.55. Furthermore, the means for the two are remarkably close, 71.36% for the average percentage of male line spoken and 72.51 % for percentage of male words spoken. We should note, however, that 3 of our movies are defined as male dominated (see Subsection 2.1 for the definition of a male dominated film) by words but not by lines.

3.2 Random Sample

For our analyses, we wanted to have a random sample of movies that was as balanced according to genre as possible. As described in Section 2, our main source of scripts was `http://IMSDb.com`. This website classifies movies according to 18 genres, 16 of which had a sufficient supply of scripts to be considered(The Film-Noir and Short Film categories were not used as they each contained fewer than 5 scripts). Each film entered into our program is classified according to these 16 genres, where the genres attributed to a film are determined by IMSDB or IMDB, where the latter is used when a film script is taken from an alternate source.

To create our random sample, we found the genre represented by the fewest scripts; say this genre contained n movies. We aimed to select a subset of our 200 films with the properties that:

1. The sample contains at least n films of each genre.
2. The movies selected to represent a genre are selected randomly.

To create the random sample, all movies are considered. They are read in from our common output file and placed in a vector, which is shuffled according to the C++ algorithm library's random shuffle function. The program then goes through the shuffled vector and examines each movie in order. If there is a genre attributed to that movie for which the sample has fewer than n movies, the movie is included in the sample. If not, the movie is not included in the sample.

This process resulted in a random collection of 94 movies, with the following breakdown of genres:

Genre	Number of Films
Action	19
Adventure	22
Animation	13
Comedy	21
Crime	15
Drama	53
Family	9
Fantasy	12
Horror	6
Musical	9
Mystery	11
Romance	17
Sci-Fi	11
Thriller	15
War	8
Western	7

In this random sample, the average percentage of lines spoken by males is 70.58%. The average percentage of lines spoken by females is 29.12%, leaving an average of unidentified character genders of only .3%. Similarly, the average percentage of male words in the films is 71.05% while the average percentage of female words in the films is 28.64%. On average, a film contained 19.15 male characters and 7.04 female characters, with only 1.25 gender neutral/ undetermined characters. This means that the average film has 69.79% of its cast consisting of males.

Note: This type of aside should be separated for a more colloquial paper. Imagine being given a coin and wanting to know whether or not the coin is biased. The typical way to test this is to flip the coin many times, and record the number of times the coin comes up heads. You expect the percentage of heads to be nearly 50% (e.g. 5 out of 10 coin flips heads), but you wouldn't be so worried if the coin came up heads 4 out of 10 tosses. Perhaps you would be worried if the coin came up 9 out of 10 heads, even though this could (and in fact should) happen occasionally. To account for this, we compute the probability that a coin flipped 10 times comes up heads 9 times. It is customary to call results significant if there is less than a 5% chance of the event happening randomly.

As males and females are equally distributed in the general population, there is a 50% chance that a given person is male and a 50% chance that a given person is female. Removing neutral characters from scripts, our work has an average movie containing about 26 characters, 19 of which are male. If we think of a character cast in a movie as a coin flip, casting an average movie is like performing a sequence of 26 coin flips, where heads corresponds to "a person is male" and tails corresponds to "a person is female". We could then ask what number of males/females corresponds to a 5% chance that the theoretical casting coin is fair? For a cast of 26 characters, 19 or more male characters corresponds to a 1.4% chance, while 18 or more male characters corresponds to a 3.78% chance. Given that a coin can be biased in either direction, this means that there is a 2.8% chance that a movie has either more than 18 males or more than 18 females and a 7.56% chance that a movie has either more than 17 males or more than 17 females. With a 5% significance level, our data does indicate statistical significance of gender bias when casting an average movie.

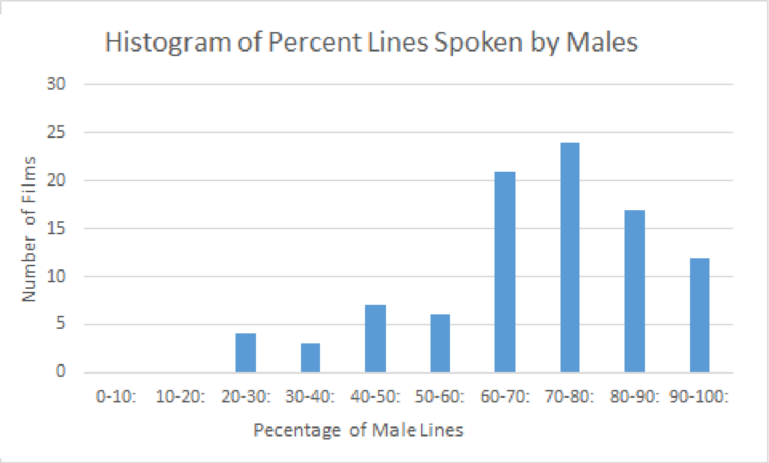
Begin editing again here.

Similarly,our random sample considers 94 scripts. We can think of tossing a coin 94 times, where heads corresponds to "a film is male line/word dominated" and tails is "a film is female line/word dominated". For 94 total films, having either more than 56 male dominated line/word films or fewer than 38 male dominated line/word films would register as significant. Our random sample had 80 male line dominated films and 81 male word dominated films, both of which certainly qualify as significant results. Note that an alternative method of performing this analysis is via z-scores. The former result gives a z-score of 6.81.

Another analysis performed on our random sample was an appropriateness of fit for the normal distribution. That is, instead of only accounting for if a movie was male dominated or not with respect to a particular attribute, we keep in mind how much a movie is male dominated. For this, we compared our random sample to a theoretical sample of the same size that would fit a normal distribution of the data. For the theoretical random sample, we assumed a mean and median of 50% male lines and a standard deviation of 12.5% lines. Figure 3.2 provides a histogram of the percentage of male lines in the script.

When performing a one sample z-test comparing our random sample with the theoretical normal distribution, the resulting z-score is 11.93, where a z-score of 2.56 would be significant at the 5% level. More simply, as one can see from Figure 3.2, a normal distribution is not an appropriate fit for what was observed in our random sample.

The last analysis done on our random sample was a more detailed look at representation of males and females within the categories of main characters, supporting characters, and bit characters to see if women were more well represented in any of these categories. The following table summarizes the number of male and female dominated films in each of these categories. The definitions for these terms are provided in Subsection 2.1.



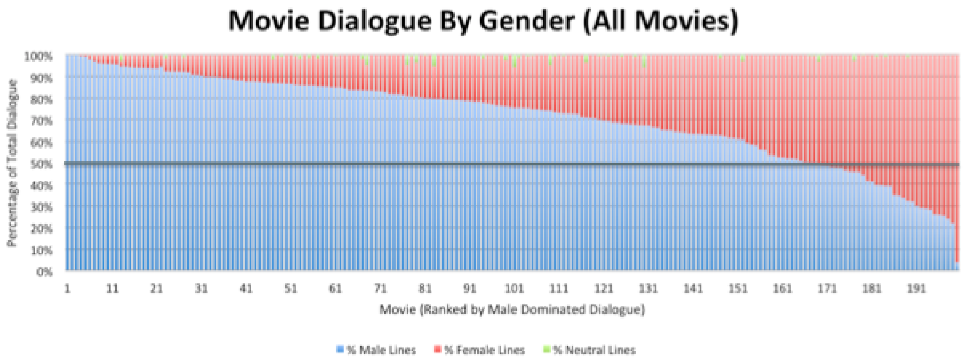
	Main Character	Supporting Character	Bit Character
Male Dominated	74	79	80
Female Dominated	20	15	12

Note that there are only 92 movies in the bit character column because 2 movies did not have any bit characters.

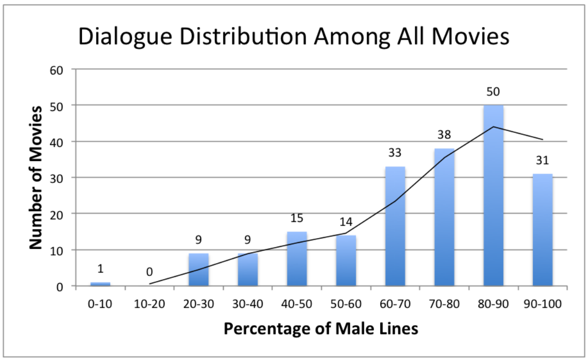
For this analysis, we performed three χ-squared tests each with one degree of freedom. For 5% significance, a χ-squared value of 3.84 is sufficient. The resulting χ-squared values for our tests were 31.02 for main characters, 43.57 for supporting characters, and 49.19 for bit characters. These very high χ-squared values indicate that the number of male dominated films present in our sample is almost certainly due to bias as opposed to chance.

3.3 All Movies

We were able to collect data on exactly 200 movies in total. The following figure shows the breakdown of lines by gender, ranked highest to lowest in terms of male percentage. Those movies with a blue bar above 50% are considered male dominated.



We had 166 Male Dominated Films, giving a t-score of 9.33. The following shows how these movies were distributed.



The range with the highest frequency is 80-90% male lines.

Average Year	2001
% Male Lines	71.36
% Female Lines	28.23
% Neutral Lines	0.40
% Male Words	72.51
% Female Words	27.09
% Neutral Words	0.40

*Note that Percentage Lines sums to 99.99 due to rounding.

Male Main Characters	1.58
Female Main Characters	0.54
Number Males	19.4
Number Females	7.2

Notice that our definition of Main Character gives an average of 2 Main Characters, which seems to validate our definition of Main Character, if you assume a protagonist and an antagonist in each movie.

Our data indicates that a male actor is about 3 times more likely to get a main role in a movie than a female actress.

3.4 Gender and the Academy Awards

Another grouping of movies that we studied was Academy Award Nominated films, to see just how much credence the recent media frenzy over the bias in the Oscars. The majority of our sample was from the last three years worth of Oscars, which are from four years ago, due to the way the academy awards work. That being the movies of each year are only nominated for the following years Academy Awards, an example being *The Revenant*, which came out in 2015 was nominated for and won Best picture in 2016. Our results, as seen below, show some fairly significant outcomes:

Percent Lines	Average
% Male Lines	72.20%
% Female Lines	27.41%
% Neutral Lines	0.39%

The statistics show just how much bias exists in the Academy Awards, with males receiving just under 75% of all lines on average. To give some context to this, that means that whenever a character speaks, 3 out of 4 times, they are a male. This is not better than any statistics generated by our random sample, which indicates that the Oscar

The next section to look at is main character distribution between genders. The table below indicates the percentages of main characters, according to Definition 3:

Percent Main Characters	Average
% Male Main Characters	74.81%
% Female Main Characters	25.19%

The data shows that some of the bias in quantity of lines most likely has a strong correlation to the percentage of main characters associated with that particular gender, with a 2.6% difference between the percentage of lines and the percentage of main characters. This does not dismiss the bias, however, but more clearly demonstrates where it lies. That being in casting, and again the Academy Award winning movies are no less biased than their not nominated counterparts.

We then progressed on to look at supporting characters, and their gender distribution, which is shown in the following table:

Percent Supporting Characters	Average
% Male Supporting Characters	71.84%
% Female Supporting Characters	28.16%

while supporting characters are technically more balanced than the main characters were, it was only by a measly 3%, which considering the fact that it still leaves the percentage of male parts at over the 70% mark, leaves allot to be desired. This is still no better than the bias found in the rest of the filming industry.

Finally we took a look at the simple average of quantity of roles, by gender for each movie. The results are shown in the following table:

Totals	Average
Total Number of Males	20.20
Total Number of Females	7.33

These results show that there is a vast difference in sheer quantity of roles with an average difference of almost 13 parts per movie. This is saying that if life were a movie, males would outnumber females 20 to 7, or nearly 3 to 1. This is obviously not the case, and just further goes to demonstrate that the Academy Awards are no less biased than any other films.

3.5 IMDB Top 250

IMDB.com (The International Movie Database) is a website that keeps track of a wide range of movie and actor information. IMDB allows its users to find films and leave their ratings and reviews, which are then averaged across the database. IMDB keeps track of these average scores, and has compiled a list of the top 250 films. Using the data gathered from our main study we compiled a list of 52 films from the public's top 250 list.

On average each of these films was nominated for 6 Academy Awards, indicating that the public and the academy have the same taste in films. We also see that, on average, men spoke 80% of the lines in these films. Only two films had more female lines than male lines, and no film had more than 75% female lines (as opposed to 37 films with 75% or more male lines).

3.6 Kids Films

The criteria for classifying our films as a "Kids" film came from Wikipedia's article "List of Children's Films" [6]. We took exception the *Star Wars* films, which Wikipedia included. This gave us 38 films in total.

These films have the following breakdown.

Percent Male lines	Percent Female lines	Percent Neutral Lines
67.69%	32.22%	0.088%

Male Characters	Female Characters
13.92	6.81

The demographic breakdown of main characters in Kids movies is as follows:

Average Male Main Characters	Average Female Main Characters
1.66	0.71

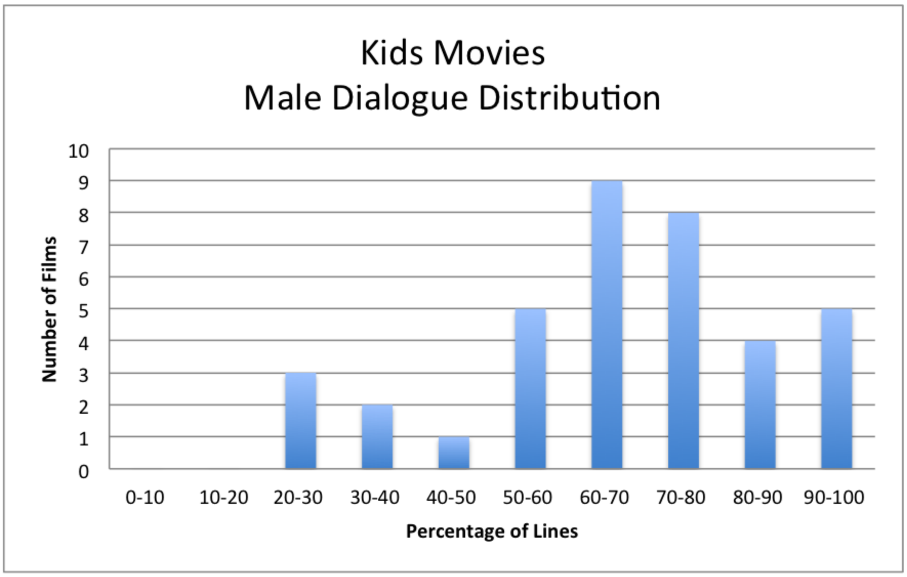
Whites	Non-Whites	Blacks	Asians	Hispanics	PI/Hawaii	NA/AK	Mixed
80	8	2	3	1	0	1	1

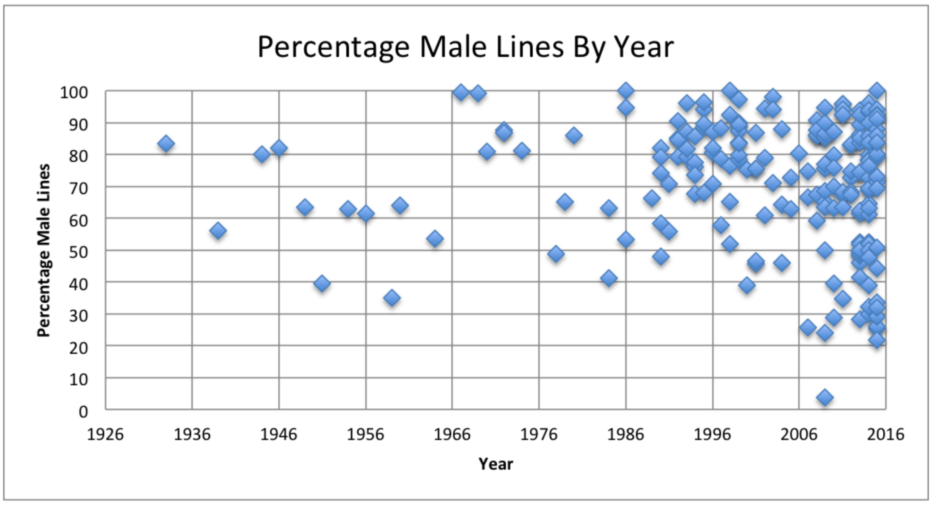
Note that the following figures are represented as averages.

Age Male	Age Female	BMI Male	BMI Female
42.84	37.78	24.60	20.33

Perhaps the most important statistical test run on our kids films is a one-tailed t-test comparing our kids films to a random sample of non-kids films. When performing this test, we obtain a t-score of 1.68, giving a p-value of .047. At the 5% significance level, this indicates that there is a significant difference between kids films and non-kids films, where kids-films are less biased towards females.

In general, it appears that childrens films are subject to more scrutiny regarding gender bias than films for the general population. Although still gender biased, the fact that kids films are significantly less biased indicates that paying attention to gender bias in films can decrease over bias of the film.





3.7 Through The Years

Since we gathered data from movies spanning a large period of time, we were interested in looking for any patters or correlations over time.

Due to the fact that recent films are more likely to have their script available online, our data is slightly saturated with films released within the past 30 years. It is difficult to find any patterns from this graph. One may notice that between 1966 and 1976 that there are 5 films at or above 80% male lines and none below, but we would need a larger sample size to make any substantial claims. Similarly, it is only after 2006 that we have any movies below the 30% line (that is, 60% or more female lines). But again, that time frame from 2006 and 2016 is where a large percentage of our movies were released, so it is difficult and uncertain to compare to the previous time frames.

We ran a linear regression analysis on percent movie lines and the year in which they were released and got an R value of 0.2379 (R-squared = 0.0566) with the following equation

$$\%\text{Male Lines} = 3.21398 + 0.03404 \cdot \text{Year}$$

With such a low R value, this model is certainly inaccurate and insignificant, yet that very fact makes this data interesting. In this case, accepting the null hypothesis that year has no effect on percentage of male lines is perhaps more insightful than rejecting it. Many people who are aware of such discussions of gender bias in films may be inclined to believe that, despite it's faults, the movie industry is progressing towards gender equality. Our data gives no evidence to support such a claim.

3.8 Race and the Academy Awards

A lot of controversy surrounded the 2016 Academy Awards as all of the nominees for Best Leading and Best Supporting Actor and Actress were white [8]. This is not the first time the Academy Awards have been criticized for a lack of diversity; our aim was to gather some data to asses this criticism quantitatively.

Using our Actor Data program, we compiled information on all actors and actresses that qualified as a main character according to our definition (for more details, see Subsection 2.3). From this data, we filtered according to two data sets; all films that received at least one Academy Award nomination (in any category), and films that were nominated for an Academy Award in the past three years. Films nominated for an Academy Award in the past three years (films from 2013 - 2015) were actively sought out in order to ensure a reasonable sample size, and all available scripts were used in our analysis. In total, our sample for the latter data set contains 43 films. With the exception of foreign films and film shorts, all films nominated for Academy Awards from 2013 - 2015 are included in this sample except for 12: *Wild, Mad Max, Room, Cinderella, Selma, Two Days, One Night, The Judge, Blue Jasmine, August: Osage Company, The Grandmaster*, and *Inside Illewyn Davis.*

We then ran a chi-squared distribution on the main actors and actresses from these films, where actors and actresses were sorted by race into one of the following categories.

1. White
2. Black
3. Asian
4. Hispanic
5. American Indian / Alaska Native
6. 2 or more races

These races were chosen to correspond with the racial listings available via the Henry J. Kaiser Family Foundation website, `http://kff.org`. From this website, we were able to gather information about the racial demography of the United States to compute expected values for the number of people in each

	White	Black	Asian	Hispanic	AI / AN	2 or more races
Observed	265	14	4	4	2	3
Expected	181	35	17	52	2	5

Number of main actors/actresses of each race in all Academy Award nominated films.

	White	Black	Asian	Hispanic	AI / AN	2 or more races
Observed	78	6	0	2	0	1
Expected	53.76	10.62	5.04	15.30	.68	1.60

Number of main actors/actresses of each race in 2013-2015 Academy Award nominated films.

racial category. The observed/expected values for each data set are summarized in the following figures.

For both tests, a chi-squared value higher than 11.07 signifies statistical significance at the 5% level. The chi-squared value for our sample containing all Academy Award nominated movies regardless of year is 107.41 while the chi-squared value for Academy Award nominated movies from the past three years is 30.45. That is, both of our samples demonstrated racial bias in a statistically significant way. By examining the data, we can see that each race other than whites is significantly underrepresented, strongly justifying the claim that at least Academy Award nominated films are dominated by whites.

3.9 Actor/Actress Bias

As mentioned in 2, we were able to categorize characters mathematically according to their vocal presence in a film. Additionally, we assigned each character the actor or actress who played them on-screen. From the 200 movies we analyzed, we had 241 actors and 102 actresses who played a main character. We attempted to gather as much information on these people as possible. This data comes entirely from a variety of websites, the reliability of which is questionable of course. We recorded information on age, height, weight, race, and net worth. From an actor or actress's height and weight we were able to calculate his or her BMI. We recognize the failures of using BMI as a measure of healthiness, but it is the best measure we could use with the available information.

3.9.1 Age

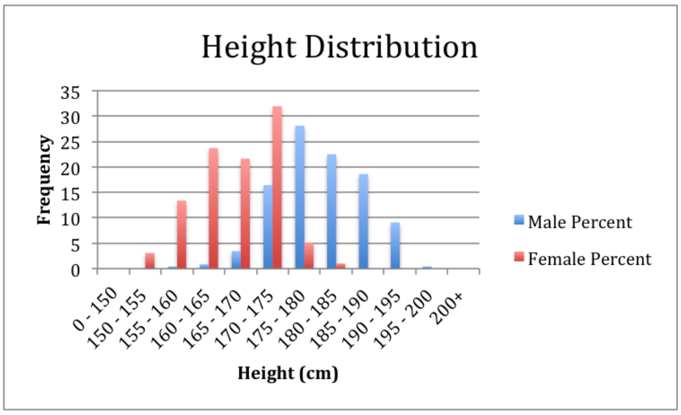
In the following analyses age will refer to the age of an actor or actress at the time the film they appeared in was released.

	Sample Size	Mean	Standard Deviation
Male	185	40.75	10.86
Female	96	35.53	13.92

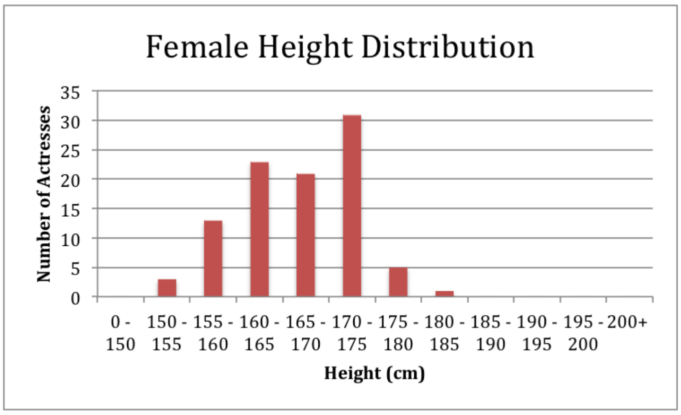
The average age of a male Main Character is a little over 5 years older than a female Main Character. A t-test assuming equal variances gives a t-score of 3.46 ($p = 0.00031$). Therefore, we conclude that there is a significant difference in the ages of leading male actors and leading female actresses.

3.9.2 Height

The following graph shows the distribution of height (cm) among males and females.



Clearly there is a difference between the male and female samples, but that is to be expected. The male sample appears to follow a normal distribution (with perhaps a slight negative skew towards taller actors), but we noticed a few interesting characteristics in the female distribution. The isolated histogram is shown below.



We notice two peculiar qualities: (1) The range with the highest frequency is 170-175 (cm) and (2) There is a remarkable drop off in frequency after that 170-175 (cm) range. Firstly, according to the CDC[1], the average height of a female adult is 162 cm. Our sample's mean is 165.7 which is slightly taller, but it is still odd that there is such a drastic drop off after 175 cm.

One explanation, which cannot be proven mathematically, is that there is an attempt for a female lead to "compliment" a male lead. The peaks for the male and female height distributions are adjacent to each other, with females one range shorter. Perhaps filmmakers (or rather film casters) attempt to pair a male actor with a female actress who is slightly shorter, but almost never taller, than the male counterpart. In this respect, it is possible for an actress to be "too tall" for Hollywood. In fact, Academy Award nominated actress Sigourney

Weaver claims this is true. In an interview, she revealed that early in her career, “hardly anybody wanted to hire a woman who was six feet tall”. She added, “Which man wants to spend their shooting days on the set on an apple box just to be able to look into my eyes?”.

http://www.torontosun.com/2014/07/05/sigourney-weaver-i-was-too-tall-for-hollywood

3.9.3 BMI

In our sample, we found the average BMI for males and females to be 24.33 and 20.98, respectively. Both of these means are in the optimal or ”healthy” range of 18.5-24.9. Using the t-test on these samples gives a t-score of 7.08, showing that there is a significant difference between male and female BMI.

There is some debate on how BMI is distributed among males and females. However, the CDC[1] claims that the average BMI for male and female adults in the U.S. are nearly the same (26.6 and 26.5). Therefore, according to our results, the average leading female actress is approximately 6 kg/m^2 less than the average female adult.

3.9.4 Net Worth

When Forbes released its lists of the highest paid actors and actresses in 2016 there was controversy over the discrepancies between the two groups. For example, the 20th ranked actor (Matt Damon - $25 million) would have ranked 3rd in the female category, above stars such as Angelina Jolie ($15 million) and Sandra Bullock($8 million). Again, we wanted to broaden this sample size.

Since our films span several decades, we decided it would be unreasonable to find yearly incomes for each actor/actress in the year a movie was released. Although it may not be ideal, we used the actor or actress’s current net worth. While this may create problems with actors and actresses who have passed away, or haven’t released movies recently, it is the most widely available standard of income available.

We compiled the following data on the net worth of males and females.

	Sample Size	Mean	Standard Deviation
Male	215	56,276,074	87,627,389
Female	89	32,212,359	51,312,978

Running a t-test on our data (assuming equal variances) gives us a t-score of 2.42 ($p = 0.00799$). Therefore, we conclude that there is a significant difference in the net worth of actors and actresses.

We were not able to create sufficient sample sizes to compare net worth among races.

4 Discussion

While we were as thorough as we could be in pursuing our research, there were some things which limited the amount that we could do. One such thing is the availability of scripts, which is confined to those which other people have already written and made available in a text format online. We chose to use two main websites in collecting these scripts, the first and foremost of which is IMSDb, the International Movie Script Database. We also use simplyscripts as a source, but to a much lesser degree than its counterpart.

The first major restriction other than the available scripts that we encountered was that we need the scripts to be formatted in a fairly specific way so that our program can identify the start of dialogue, as well as who is speaking that dialogue. This means that there are many scripts which are unusable by us because of how they are formatted. This subset of movie scripts unfortunately contains most Disney scripts. We managed to find a few Disney scripts that fit the format that we are using. The format we are looking for, which is still very common, is to have the character's name in all capital letters, in its own line, followed by a paragraph of text which is mostly free of stage directions.

A further complication that came up through our research is that musicals and documentaries are particularly uncooperative with both their formatting and content. Musicals are a problem because of the way that they are formatted whenever multiple characters sing at the same time. They are written so that both character names are in the same line, and the words run parallel in the script, separated by a central tab. This is easy enough to understand when looking at it as people, but our program is unable to identify this as multiple characters. Documentaries, on the other hand, are difficult for our study because they have an almost complete lack of characters, as well as scripted dialogue. The only audible content in most documentaries is a voice over, or some other form of third-person narration of the topic. There are a few exceptions to this, but even within these exceptions, quite a few of them are unscripted. This is referring to the fact that what the people being interviewed say is not scripted, and thus would not appear in an official script.

Another complication that we encountered is that we, as a group, only speak English, which meant that we have to limit our database to movies whose scripts are easily accessible in English, so we can tell whether our code is executing properly or not. This is not the most significant restriction, as there are a plethora of movies in all genres in English, but it does limit our results to be relevant to mostly American made films.

Furthermore, even the scripts that fit all of these requirements still have their fair share of problems. One of the largest of these is how much of the editing process the script has been through in its current state. There are a few scripts we analyzed which differed noticeably from their finished product. Differences ranged from as insignificant as an extra scene with a few lines of dialogue between two minor characters, to being as severe as an entire supporting character who was cut from the film. The event that made us notice this was a problem is in the script for *12 Years a Slave*. In the script, there is a character by the name of Celeste, who did not make it into the finished product. We struggled to figure out which IMDb actress was unidentified, until eventually we found out that she had simply been cut.

For comparison's sake, we can look at another study done by Eisenhauer and Fought, who did research on around 20 Disney movies. We only had 5 movies in overlap, due to accessibility of scripts, as well as readability. The table below compares the Disney films for which scripts were available to the results recorded from the actual movie by Eisenhauer and Fought:

Movie	% Female Lines (EGJM)	% Male Lines (EGJM)	% Female Lines (EF)	% Male Lines (EF)
Beauty and the Beast	29.2	70.8	29	71
Frozen	47.3	49.6	41	59
Mulan	34.8	65.2	24	76
Pocahontas	32.2	67.8	25	75
Sleeping Beauty	65.1	34.9	71	29

Within these movies, the only one that has a difference larger than 10% is Mulan, and the difference there was only 10.8%. The reason that we found these numbers reassuring is because we anticipated some margin for error, especially with gender neutral characters. This is referring to the fact that, while looking at the script, a character named bystander#1 could be either a male or a female, whereas while watching the actual film it is quite easy to tell whether they are a male or female character. Additionally, we had difficulty attributing the correct quantity of words with certain songs, which ended up lasting for several paragraphs. In the end, our code was only able to identify the first paragraph of a song, which usually was only the first verse, which helps account for the other 5% of difference between our studies.

One other very significant drawback to our program is that it generates statistics based off of quantity of lines, not quality. That is to say, it cannot tell the difference between a strong character, and a weak one. An example of this shortcoming is in Snow white and the Seven Dwarves. In the movie, a majority of the lines are spoken by females, either by Snow White herself, or the Evil Queen. However, when you look at those characters for their quality of lines, they are extremely stereotyped. This movie is, according to our study a great, and even progressive movie, in terms of female casting, with 65-70% of the lines belonging to females. However, when you actually watch the movie, there is a problem with this, in that the female characters are not strong characters at all. This movie demonstrates why quantity does not always mean quality, which is probably the biggest downfall of our program.

5 Future Research

Future research directions in this area are abundant, easily accessible, and crucial if equality is to be reached in the very prominent film industry. Some of the many questions that occur to the authors of this paper are:

1. How prominent is age discrimination? What is the average age of a male leading character as opposed to the average age of a female leading character? Are men more likely to be a leading character?

2. Is there any correlation with BMI and average number of lines/ role prominence? Are leading roles equally distributed among people of all body types?

3. Are the movies featured in the most recent Oscars ethnically biased? Are the movies ethnically biased?

4. How biased are the top grossing films, taking into account inflation? Are the movies seen by the largest populations the most biased?

5. How strong is gender bias in children's movies; that is, movies that are rated PG or below?

6. How strong is bias (of any type) in the lowest rated movies of all time?

7. Is there a difference in gender bias in films directed by males as opposed to those directed by females?

8. How faithful are movie representations to books? That is, are the main characters in a movie the same as the main characters in the film representation?

9. Is there a difference in gender bias in films directed by males compared

References

[1] *Center for disease control and prevention.* http://www.cdc.gov/nchs/fastats/body-measurements.htm.

[2] *The new york times.* http://www.nytimes.com/2016/07/19/movies/so-thats-who-you-call-the-politics-of-the-new-ghostbusters.html?_r=0. Accessed: 07-20-2016.

[3] *Sag-afta.* http://www.sagaftra.org/home. Accessed:08-23-2016.

[4] *Usa today.* http://college.usatoday.com/2016/02/01/study-what-disney-princesses-arent-saying-hint-its-more-than-you-think/. Accessed: 06-05-2016.

[5] *The washington post.* https://www.washingtonpost.com/news/wonk/wp/2016/01/25/researchers-have-discovered-a-major-problem-with-the-little-mermaid-and-oth
?tid=pm_pop_b. Accessed: 06-03-2016.

[6] *Wikipedia.* https://en.wikipedia.org/wiki/List_of_children%27s_films note=Accessed: 07-12-2016.

[7] *World woman foundation.* http://www.worldwomanfoundation.com/hollywood-ageism/. Accessed: 08-22-2016.

[8] J. Syed, *Oscars so white: an institutional racism perspective*, Counterpunch, (2016).

[9] J. R. R. Tolkien, *The letters of JRR Tolkien*, 2014.