
Investigating gender differences in Episode IV and VII of the Star Wars movies

Samuel Effler
Matrikelnummer 5929373
samuel.effler@student.uni-tuebingen.de

Sarah Dockal
Matrikelnummer 5626979
sarah.dockal@student.uni-tuebingen.de

Abstract

In this paper gender differences regarding the amount of speech in two star wars movies are investigated. After some consideration we decided to work with movie scripts instead of subtitle data. We gathered the information from the scripts with the help of a self-implemented parser. The results are illustrated via pie charts, scatter plots showing the correlation of the number of words and how many times a character has spoken and lastly word clouds for female and male characters.

1 Motivation

We looked at the lecture slide with proposed data sets. Since we both enjoy watching movies, the data set including subtitles sounded very promising so we agreed on doing our project with this data set. In the next step, we thought about what information we could extract from this data set. Rather fast, we decided to investigate gender differences especially the representation of women. For example how big the share of speech of both female and male characters is. How many female characters appear was the second questions we asked ourselves.

Next, we thought about which movies we want to examine. Firstly, we thought about marvel movies since we both are huge fans of the movies. But than, the idea entered our minds that we could investigate at least two movies with some time apart so we could see whether there were any changes in this period of time. Most of the Marvel movies were all made in a relatively short period of time, compared to the star wars movies. Here, there are three times three films and a certain amount of time has passed between the trilogies. Perfect for our research. To reduce the scope of this work, we only analyzed the first (Star Wars: Episode IV – A New Hope) and the seventh (Star Wars: The Force Awakens (Episode VII)) Star Wars movie.

2 Data collection

We looked at the data set with the subtitles and decided that they were not enough for our vision of the project. There was too much information missing, most importantly the information who was talking. So we searched for scripts of the movies and used those from [imsdb](https://www.imsdb.com/)¹. The next step included thinking about a way we can scan the documents and extract the information like who is talking and what are they talking without doing it manually. Therefore, we programmed a parser, that can be found in the [git-repository at github.com/SamuelEffler/UniTueDataLiteracy](https://github.com/SamuelEffler/UniTueDataLiteracy).

Parser The required information is extracted from the respective HTML file of the script web page. For this purpose, we saved the pages locally as HTML file. Initially, the HTML code was read using Python and searched for specific string patterns that occur in spoken sentences in the script. However, we changed the procedure in a second iteration. Here, the search for string patterns was replaced by regular expressions (regex). This made searching more efficient and comprehensible.

¹<https://www.imsdb.com/>, accessed on 2022-02-06

Furthermore, this approach allowed to quickly change the searched patterns. Since the scripts are not always formatted in the same way, different patterns were necessary to analyze different movies. The exact procedure of the algorithm is as follows: First, the places in the screenplay where a character name appears are searched. A distinction is made between searching for all occurring names based on the formatting only and searching for a match with a list of given names. In some scripts it was not possible to distinguish names from other text passages, such as stage directions, so a list of the first ten characters in the end credits of the movie was created. Once a character name is found, it is saved and the subsequent spoken sentence is searched for. Before the phrase is saved, stage directions and punctuation marks are removed to simplify later analysis. Before the record of name and phrase is added to the CSV file, the number of spoken words is counted and added. Finally, the data set is saved and the next name is searched. The final output is a CSV file that contains the names, phrases and number of words spoken. Characters such as droids who do not speak English are not considered, either in word count or in times spoken.

3 Data preparation

From everyone who spoke in the movies, we selected the first 10 because they make up the most in the movies. An example for this is the following: In Episode IV, Luke has the biggest impact regarding speech. He spoke 252 times. The tenth person is Wedge who spoke 14 times. This is why we decided to cut the data to the first 10 persons. Also, we dropped characters like 'leader' or 'trooper' because there are not bound to a specific person.

After this, we added the column 'gender' to the data frame. We divided the data into the categories 'female', 'male' and 'none'. 'none' is included since C3PO makes up a large share of speech. We then summarized the number of words as well as how many times were spoken per gender. The visualization of this can be found in section 4.

Further more, we collected all the words per gender to generate a word cloud for female as well as male spoken words.

4 Data visualization

For the visualization, we decided to use a pie chart to illustrate the ratio between female, male and none characters since this visualization gives a good overview.

After this, we implemented a interactive chart to see which character talked the most and to differentiate between gender. To experience it, go to the git repository and hover over each dot.

Lastly, we wanted to investigate some of the content. This was done by summarizing all the words of each gender and to feed it into a word cloud generator². Larger words represent a more frequent use of them.

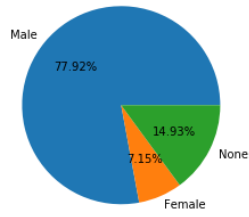


(a) **Episode IV.** Overall 9148 words were spoken. 6846 were spoken by male characters, 2220 by female ones and 82 by C3PO.

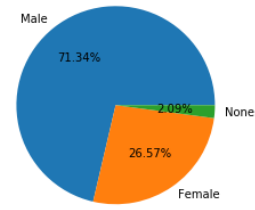
(b) **Episode VII.** Overall 5574 words were spoken. 4197 were spoken by male characters, 1198 by female ones and 179 by C3PO.

Figure 1: How many words did male, female and none speak in both Episode IV and in Episode VII.

²Created with <https://www.wortwolken.com/>, accessed on 2022-02-06



(a) **Episode IV.** Overall characters spoke 797 times. 590 words were spoken by male characters, 193 by C3PO and 14 by female ones.



(b) **Episode VII.** Overall characters spoke 670 times. 478 words were spoken by male characters, 178 by female ones and 14 by C3PO.

Figure 2: How many times did male, female and none speak in both Episode IV and in Episode VII.

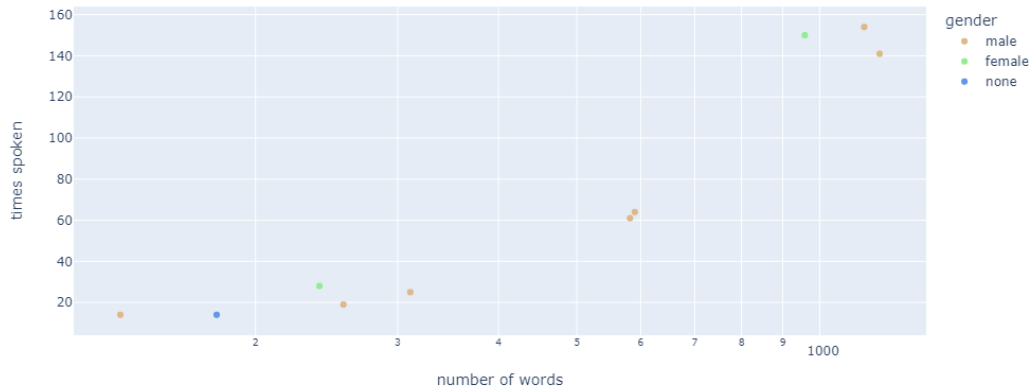
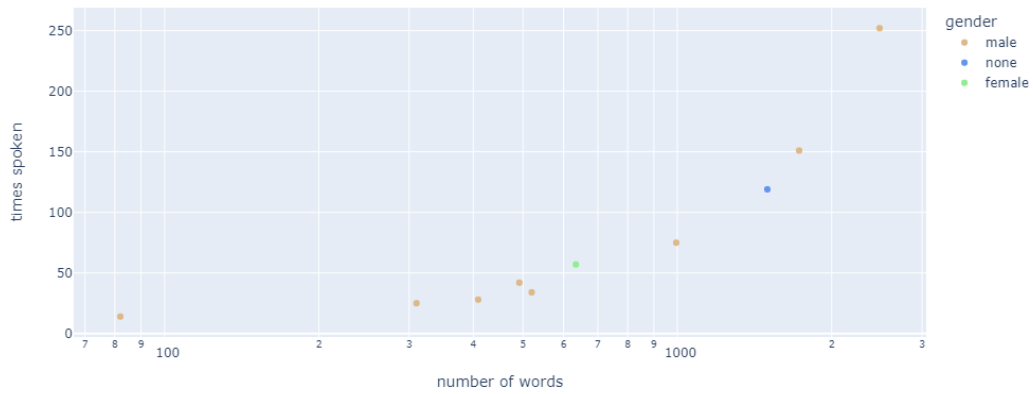


Figure 3: Correlation of number of words and times spoken for each of the ten characters in the fourth (first figure) and in the seventh (second figure) movie. To experience the interactive graphic, go to the git repository. There you can hover over each scatter point to see which dot is which person and the exact number of words and spoken times.

