Data

The dataset consists of 617 movies released between 1927 and 2010, from the Cornell Movie-Dialog Corpus in collaboration with Cristian Danescu-Niculescu-Mizil and Lillian Lee. These were selected by the authors on the basis of a minimum of 5 IMDb ratings and the match between the script and the actual movie. Furthermore, the movies’ provenance is made available through the URLs link of the raw data source.

The dataset contains metadata about the movie characters, movie titles, movie lines, movie conversations, and raw scripts. Among this big amount of data, this research focuses on the movie genre *family*, looking at the characters’ gender, movie lines and date of release.

Before being able to use such data, the process of data cleaning was needed. Indeed, after having selected only movies from the genre “family”, some data turned out to be incomplete, as some of the characters’ gender values were missing. This was solved through the manual assignment of gender through Internet research. Furthermore, not all movies were assigned to the right genre, such as the film “Napoleon”, which was categorized as a “family” and “adventure” movie even though it was an unproduced historical script by Stanley Kubrick. The movie was then excluded from the analysis. Another data limitation was related to the assignment of the gender to the cartoon characters, which was done on the basis of the gender of the respective voice actors. Furthermore, some data were precluded because they referred to various people talking at the same time, to which it was impossible to assign a specific gender, making them not suitable for the research.

The analysis of the proportion of male and female speaking time through years also required some data cleaning, as the dataset did not mention on which country the provided dates were based. This was solved through the manual assignment of the dates based on their publication in the Netherlands through Internet research. Furthermore, the month of release was also added to the data, as multiple films were published in the same year and necessitated a more precise time to be compared.