Bachelor's Thesis

Moving beyond the Bechdel-Wallace test: using NLP to study gender representation in animated television shows

Author
Emma Sophie Vališová
emma.sophie.valisova@gmail.com

Supervisor
Zarah Leonie Weiss, M.A.
zweiss@sfs.uni-tuebingen.de

A thesis submitted in partial fulfilment
of the requirements for the degree of

Bachelor of Arts
in
International Studies in Computational Linguistics

Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

September 2021

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst habe, dass ich keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe, dass ich alle wörtlich oder sinngemäß aus anderen Werken übernommenen Aussagen als solche gekennzeichnet habe, dass die Arbeit weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen ist, dass ich die Arbeit weder vollständig noch in wesentlichen Teilen bereits veröffentlicht habe, dass das in Dateiform eingereichte Exemplar mit dem eingereichten gebundenen Exemplar übereinstimmt.

Tübingen, September 15, 2021

Abstract

This thesis analyzes the transcribed dialogue of 16 different animated television shows in order to gauge the state of their gender representation on both a surface level and an in-depth level. A corpus of 1 803 individual speaker-annotated episode transcripts was collected and appropriately transformed for the purpose. The first part of the analysis uses statistical metrics to gauge the quantity of representation such as the proportional number of words spoken by each gender, the proportional number of speakers, an average number of words spoken by a speaker, and a variance of such statistics over the course of individual episodes. The second part focuses on natural language processing, namely co-reference analysis of the source material, in order to gauge the quality of representation. Additionally, as the majority of the analysis was performed using self-developed automatic tools, or is achievable through publically available tools such as the CoreNLP pipeline (Manning et al., 2014), this thesis forms the basis for further analyses of larger scope.

Acknowledgements

First and foremost, I would like to thank my supervisor Zarah Weiss, not only for her great advice and guidance throughout writing this thesis and making me interested in the topic in the first place, but also for her understanding and for making me feel the safest I have ever felt during my university years.

I would also like to thank my partner and my friends for supporting me all the way through this process, being interested in my work, happily listening to my enthusiastic ramblings, and cheering me on when I needed it the most. Without you, I never would have had the strength to do this.

Finally, I would like to express my gratitude to my family, without which I never would have had an opportunity to study abroad and I never would have had the chance to write this thesis. Thank you for backing me up in my dreams.

Contents

List of Tables

List of Figures

1 Introduction

Art in its many shapes and forms is and always has been a core pillar of the world, serving as a form of expression, entertainment, respite and activism all at once. All art inherently reflects a set of cultural, moral and political values, either intentionally or not (as a lack of active statements in a work of art is a statement within itself), as art is a reflection of the artist, and no human being is apolitical or lacking any sort of values or convictions. It is therefore clear that works of art carry the values and ideals embedded in them into the public sphere, at the very least give those values space in discussions, make them heard, and potentially even spread them and influence the world further. Such ideals are not necessarily overtly shown, and can manifest themselves in relatively minute or inconspicuous aspects of the work of art. With one of such values comes the idea of representation, the extent to which different groups of characters or different cultures appear, or are represented, in a work of art (such as gender, race and sexuality representation), and the way in which such characters or cultures are painted – its quality (positive, neutral or negative representation of a group). It is important to note that lacking, insufficient or negative representation is not necessarily a product of malicious intent, and can simply be caused by misinformation or lack of education on the specific topic – and in such cases, providing accurate statistical analyses on the subject can influence future representation in the matter, and is a form of education, not necessarily criticism. A representation of a group can influence the extent to which such a group is normalized over time by the majority population (its appearance being viewed as standard or normal, not a purposeful inclusion or a deviation from norm), and by extension can help members of such group feel included in the larger body of society. One can study many distinct representation groups, however, this thesis focuses on the idea of gender representation (the amount different genders make an appearance in a work of art, and the quality of such appearances).

First making its appearance in a 1985 comic strip "The Rule" in *Dykes to Watch Out For* (Bechdel, circa 1985), the Bechdel-Wallace test reads as follows: does a piece of media have at least two women in it who talk to each other about something other than a man? This question can be further separated into three sub-questions that have to be further analyzed separately: 1) does the work in question contain at least two women, 2) who talk to each other, 3) about something other than a man? The test aims to investigate a more complex issue with regards to gender representation than simple surface statistics, as a work of art can feature many non-male characters while failing to provide them with any depth, complexity, goals or stories of their own, which is a flaw a more superficial test might not be able to uncover. It is important to note that the Bechdel-Wallace test, while measuring the interactions of female characters more in-depth, is in no shape or form meant to be a conclusive evaluation of a work of art (Zeisler, 2016). Since it serves more of a first impression purpose, or a basis for further examination, this thesis aims to take over the concepts of the test and apply them to the work of art as a whole, or more specifically, to extend the simple yes/no result of the Bechdel-Wallace test into a more nuanced and detailed analysis across the whole work in question and to extract statistics more relevant to a proper evaluation of gender representation.

In addition to providing statistics about gender representation, this thesis also helps broaden the field of computational linguistics in this specific area. A speaker-annotated corpus was composed for the purposes of this thesis, and different statistical methods

were explored with regards to gender analysis. Natural language processing tools were used in order to analyze the co-reference of the corpus – which additionally included an evaluation of several types of models with regards to the highly specific format of the source data. Such gained knowledge has the potential to be used in further projects, especially in those closely related in topic, and to the potential future expansion of this research.

With countless works of art created every single day, this thesis settles on a narrow selection of works to focus on: animated television shows (further referred to simply as television shows, shows, or cartoons). Given the origins of the Bechdel-Wallace test, which spoke of movies, animated television shows align in spirit with the test and its original intent. In addition, while gender representation is without a doubt an important topic in any medium, it is arguably the most essential with art primarily aimed at, or consumed by, children who are easily impressionable. While this thesis cannot accurately represent and evaluate data from a massive sample of shows to capture a highly accurate trend across the entire spectrum (as the needed source data is largely not publically available for a critical mass of cartoons, and is time-consuming to collect from the raw footage), it aims to create an in-depth analysis of a diverse sample, and to provide automated tools able to largely handle further analyses on their own given the appropriate data.

This thesis aims to answer the first two sub-questions of the Bechdel-Wallace test for several animated television shows and dives deeper into the complex aspects of gender representation in the medium. The television shows in question were hand-picked with a focus on diversity in airing times, creators and age ratings, aiming to cover a spectrum as wide as possible within the scope of the thesis. It is structured as follows: section 2 mentions related work with regards to both the Bechdel-Wallace test and the natural language processing methods used in later sections. Section 3 explains the data selection, gathering and wrangling process, and shows an overview of the data and its metadata used in the analysis process. Methods as well as automated tools used to gather important statistics about the data are subsequently reported in section 4, and its results and analyses are shown in section 5. Finally, section 6 concludes the thesis.

2   Related Work

2.1   Bechdel-Wallace Test

What has later become known as the Bechdel-Wallace test, an idea put into paper by Alison Bechdel thanks to the inspiration from her friend, Liz Wallace, almost certainly has its origins in a quote from Virginia Woolf (Bechdel, 2013):

"All these relationships between women, I thought, rapidly recalling the splendid gallery of fictitious women, are too simple. ... And I tried to remember any case in the course of my reading where two women are represented as friends. ... They are now and then mothers and daughters. But almost without exception they are shown in their relation to men. It was strange to think that all the great women of fiction were, until Jane Austen's day, not only seen by the other sex, but seen only in relation to the other sex. And how small a part of a woman's life is that ..." (Woolf, 1929)

The comic strip (as displayed in Figure 1), has since its first publishing inspired a large phenomenon that has not only attracted praise and criticism, but also further derived

tests, both for the representation for women as well as other groups, such as LGBT characters (Ferraro & Kane, 2013) and people of color (Zeisler, 2016). The Bechdel-Wallace test has made a large cultural impact, especially after decades of its initial publication, with media outlets taking notice and frequently using or mentioning the test in their publications (McGuinness, 2013). However, mainstream usage was not the only area the test was used in, as seen by the Swedish Film Institute incorporating the test into some of their ratings (European Women's Audiovisual Network, 2013).

Figure 1: The first appearance of what was later dubbed the Bechdel-Wallace test. (Bechdel, circa 1985)

However, the Bechdel-Wallace test has never meant to be an all-encompassing test of the inclusivity of a piece of media, and as stated in Zeisler (2016), the test's utility has been elevated beyond its original intention. While it is possible to apply the rules as a type of litmus test to pieces of media to gauge the surface level of inclusivity, it is important to note that passing the Bechdel-Wallace test does not equate with being inclusive and feminist. It is reasonable to object to the attention the test is given and the status of a conclusive test that may be attributed to it, but it would be unjust to judge the test itself for it, as it was never designed to fill such a role.

Nevertheless, work has also been made in the field of automating the application of the Bechdel-Wallace test, one example of which is shown in Agarval, Zheng, Kamath, Balasubramanian, and Dey (2015). This paper aimed to create an automated way to test whether or not a movie passed or failed the three parts of the Bechdel-Wallace test. In addition to databases containing gender data, this research has used a gender guesser method similar to the one used in this thesis and described in subsection 4.1. However, the paper only uses machine learning for topic detection, and detects characters interacting with each other by the position of their lines relative of each other, with care

put into separating different scenes from each other. With this approach, the paper mentions the test for characters interacting with each other does not reach high levels of recall.

Additionally, computational methods have been used to assess the matters of gender equality in other spaces as well. An interesting research has been conducted in Garcia, Weber, Rama, and Garimella (2014), in which social media posts involving trailers of different movies were analyzed and the statistics of the numbers, genders and interactions of the users who shared such trailers were related to whether or not said movie passed the Bechdel-Wallace test. Another research has also been conducted in Wagner, Garcia, Jadidi, and Strohmaier (2015), in which computational tools were used to find differences in gender portrayal on Wikipedia, and finds that while women are sufficiently represented in Wikipedia articles, their portrayal is often significantly different from male figures.

2.2 Co-reference Analysis

In the instance where two different words or expressions in a text refer to the same object – referent (e. g. in the sentence "*Jane* said *she* is ill," the words Jane and she refer to the same object, Jane), we are speaking about co-reference (Crystal, 1997). As such a relationship between expressions can be complicated and not always clear (and not even always unambiguous), co-reference resolution attempts to find such bound expressions, either for the purpose of further analysis, or for the purpose of annotation. Its uses are wide, as many linguistic tasks depend on it, ranging from the basic such as correctly understanding the source text in the first place, to more complex analyses and statistics, some of which will be explored later in this thesis. It is important to note that automatic co-reference analysis relies on other annotations in order to function. While this varies between models, in general co-reference analysis requires a type of parses, be it dependency or constituency parses.

As with many corpus analysis tasks, while it is possible to annotate co-reference resolution manually, many automated tools were developed for the purpose. There are three main branches of different models, namely deterministic, statistical and neural, and each have gone through their evolutions and inspirations. Since a full recount of the history of analyzing co-reference resolution could easily result in a full-length thesis on its own, this section aims to cover the topic briefly with outside sources for providing more in-depth information. Additionally, this section focuses on models trained for English, as that is the language the research of this thesis focuses on.

Described in detail in Jurafsky and Martin (2008), co-reference pairs can be filtered out on the basis of several constraints, among which belong number agreement, where the two expressions should align in number (singular or plural), person and case agreement, gender agreement, syntactic rules, and others. This is only a first step, however, as while these methods can rule out possible expressions that do not align, a pair that passes through every filter is still not guaranteed to be co-referent. This issue is present with rule-based approaches in general, as the English language contains many ambiguities even for a human eye, and absolute precision is thus not always achievable.

Rule-based approaches have gone through meaningful development over the course of its existence. At first, the systems did not achieve good performance due to factors such as hand-tuning the weights of the models (one such early model can be seen in Lappin and Leass (1994)). Further information about early model propositions can also

be found in Baldwin (1997). At first, the realizations of such propositions involved single-pass models, which were later replaced by the use of multi-pass models (Raghunathan et al., 2010). Such models consist of several different rule-based sieves which attempt to find a co-reference pair, and unresolved expressions trickle down into the subsequent sieves. Usually, sieves are sorted in order of coarseness, such that an exact-match sieve (which looks for strings that match each other) is applied first, with more specialized sieves being applied next. Further extensions of the model have been seen in Lee et al. (2011), such as the addition of multiple sieves and a post-processing step. In addition, several researches and improvements such as identifying singleton mentions (Recasens, de Marneffe, & Potts, 2013) have been published since then.

As described in Clark and Manning (2015), a development has been made to the co-reference resolution methods by working with clusters of mentions instead of simple pairs. With such approach, it was possible for earlier co-reference decisions to inform later ones (such as the aforementioned gender agreement, where a co-reference link between a name and a pronoun could establish the gender, and such decision could inform further co-reference links further in the text). Additionally, the model reduces the search space for possible links, and introduces several features which inform the decision to create a link, such as distance, string matching (borrowing from rule-based methods) and lexical, semantic and syntactic features.

With the rise of neural networks, such systems have penetrated even the subject of co-reference resolution. One such model is described in Clark and Manning (2016a), using mention-ranking at its core (scoring pairs of mentions on their likelihood of co-reference). Such a model can yet again be extended by working with clusters, and thus including an entity-ranking model in addition to the mention-ranking one (Clark & Manning, 2016b). Overall, the field of co-reference resolution is a blooming one, and new methods, findings and improvements are frequent.

3 Data

3.1 Corpus Profile

The corpus composed for the purpose of this thesis, the detailed creation process of which is described later in this section, contains episode transcripts (speaker-annotated lines of dialogue complimented by out-of-character descriptions) of 16 animated television shows[1]. In total, 1 803 individual episode transcripts are present, with a total number of 5 309 981 tokens collected, an average episode length of 2 945 tokens, and a standard deviation of 1 494. A detailed breakdown of the total number of tokens, the average number of tokens per episode transcript, and the standard deviation of each individual television show is available in Table 1. A more detailed metadata table of the analyzed television shows is available in subsection 3.2 in Table 2.

---

[1]Adventure Time Wiki (n.d.); Avatar: The Last Airbender Wiki (n.d.); Avatar: The Legend of Korra Wiki (n.d.); DuckTales Wiki (n.d.); Futurama Wiki (n.d.); Gravity Falls Wiki (n.d.); My Little Pony: Friendship is Magic Wiki (n.d.); Rick and Morty Wiki (n.d.); She-Ra and the Princesses of Power Wiki (n.d.); SpongeBob SquarePants Wiki (n.d.); Steven Universe Wiki (n.d.); The Amazing World of Gumball Wiki (n.d.); The Owl House Wiki (n.d.); The Powerpuff Girls Wiki (n.d.); Transcripts Wiki (n.d.); Voltron: Legendary Defender Wiki (n.d.)

Code	Total Token Number	Average Tokens/Episode	Standard Deviation
ADV	490 826	2 080	776
ATL	330 422	5 329	1 460
CLW	36 308	2 421	353
DUC	190 991	3 351	1 333
FUT	531 152	5 107	2 501
GRA	187 396	4 685	762
GUM	620 409	2 607	448
KOR	263 496	4 972	1 602
MLP	632 587	2 787	281
OWL	93 512	3 896	1 928
PPG	444 325	3 220	882
RAM	122 362	4 894	1 392
SHE	195 008	3 824	1 106
SPO	677 610	2 104	1 066
STU	411 959	2 203	839
VOL	81 618	3 401	1 930

Table 1: A breakdown of the total token number, average token number per episode, and standard deviation of each television show in the corpus.

3.2 Collection and Types

In order to gather a sufficiently sized sample of data from a wide variety of television shows, it was deemed impossible to manually transcribe a statistically significant number of episodes and provide additional required metadata, such as speaker annotation. Additionally, official transcript data (a transcript being a rewriting of an episode's dialogue with speakers assigned to their respective lines, often with out-of-character descriptions included) is rarely available and in the case of select streaming platforms which offer their own subtitles, an export of such files is not possible. To solve this problem, data had to be gathered from fan-made wikis (a wiki being a topic-specific encyclopedia website). Due to not having access to official or self-produced data, special attention had to be put into assessing the reliability and accuracy of the given information provided by a non-official source.

Accounting for such time-intensive process of collecting data, and with the number of animated television shows existing in the world greatly exceeding the scope of any one singular research, it was necessary to focus on a limited number of cartoons, and thus, for as broad of a scope as possible within the limited data availability, metrics for selection had to be created to cover as many different types of television shows as possible. Surprisingly, a crucial metric such as popularity was hard to access, as the process was hindered both by the age of the cartoon – and thus incomplete or missing viewer data, or data which fails to capture the general viewership trends across the years – and also by the recent shift in television as a whole – with many cartoons running in parallel or exclusively on streaming services instead of traditional television. With this in mind, while it was possible to use viewership and popularity to some extent to determine the shows for analysis, it was necessary to supplement those with other metrics. Determining it was important to focus on diversity (as a large number of television shows exploring similar space would provide skewed information about the cartoon scene as a whole), the metrics devised to capture a wide range of shows were as follows: 1) the main author, 2) the original network, 3) the original airing time, 4) the genre(s), and 5) the age rating.

ID	Code	Full Name	Genre(s)	Airing Time
1	ADV	Adventure Time	Adventure, Comedy Science Fantasy	2010 - 2018
2	ATL	Avatar: The Last Airbender	Action, Adventure Fantasy	2005 - 2008
3	CLW	Star Wars: The Clone Wars	Action, Adventure Science Fiction	2008 - 2020
4	DUC	DuckTales	Action, Adventure Comedy	1987 - 1990
5	FUT	Futurama	Comedy, Sitcom Science Fiction	1999 - 2013
6	GRA	Gravity Falls	Adventure, Comedy Mystery	2012 - 2016
7	GUM	The Amazing World of Gumball	Comedy, Fantasy Sitcom	2011 - 2019
8	KOR	Avatar: The Legend of Korra	Action, Adventure Steampunk	2012 - 2014
9	MLP	My Little Pony: Friendship is Magic	Fantasy	2010 - 2019
10	OWL	The Owl House	Fantasy Horror Comedy	2020 - present
11	PPG	The Powerpuff Girls	Action, Adventure Superhero	1998 - 2005
12	RAM	Rick and Morty	Adventure, Comedy Science Fiction	2013 - present
13	SHE	She-Ra and the Princesses of Power	Action, Adventure	2018 - 2020
14	SPO	SpongeBob SquarePants	Comedy	1999 - present
15	STU	Steven Universe	Action Science Fantasy	2013 - 2019
16	VOL	Voltron: Legendary Defender	Action, Adventure Science Fantasy	2016 - 2018

ID	Age Rating	Creator(s)	Network(s)	Episodes
1	PG	Pendleton Ward	Cartoon Network	236
2	Y7	Michael Dante DiMartino Bryan Konietzko	Nickelodeon	62
3	PG	George Lucas	Cartoon Network Disney	15
4	Y7	Jymn Magon Brad Landreth	Syndication	57
5	14	Matt Groening	Fox, Comedy Central	104
6	Y7	Alex Hirsch	Disney	40
7	Y7	Ben Bocquelet	Cartoon Network	238
8	PG	Michael Dante DiMartino Bryan Konietzko	Nickelodeon	53
9	Y	Lauren Faust	Discovery Family	227
10	Y7	Dana Terrace	Disney	24
11	Y7	Craig McCracken	Cartoon Network	138
12	14	Justing Roiland Dan Harmon	Adult Swim	25
13	Y7	Noelle Stevenson	Netflix	51
14	Y7	Stephen Hillenburg	Nickelodeon	322
15	PG	Rebecca Sugar	Cartoon Network	187
16	Y7	Lauren Montgomery Joaquim Dos Santos	Netflix	24

Table 2: A metadata table of the analyzed television shows.

The unavailability of official data and the need to rely on fan-made content has noticeably reduced the possible range of television shows, however, the selection still offered a large amount of variety. All 16 television shows used for the analysis along with their details can be found in Table 2 under their assigned three-letter codes, which are further used to uniquely identify them instead of their names. While the number of analyzed episodes per television show varies greatly, it was considered more beneficial to work with proportional analysis rather than gather a substantially smaller equally sized sample of each television show. Overall, a total of 1 803 episode transcripts were collected, with each individual transcript representing 10-25 minutes of raw episode footage. Individual episode transcripts were manually checked and deemed acceptable for use in the analysis depending on their completeness, accuracy and consistency. Unfinished transcripts or transcripts with partially or fully missing speaker labels were not considered for the analysis, but transcripts with missing out-of-character descriptions were deemed acceptable, as lack thereof only resulted in a larger amount of further manual work (specifically automatic gender determination, described further in subsection 4.1), not a decrease in accuracy.

All source data in the form of individual transcript websites was manually gathered in the HTML format, as such collection followed by automatic text conversion and transcript extraction proved to be more efficient than manually locating transcript data within a webpage and exporting it into a TXT format. During the process, several different types of transcripts were also identified that had to be handled differently in the upcoming data wrangling process (described in subsection 3.3). Including only the data deemed fit to be analyzed, meaning complete transcripts with fully annotated speaker information, the types appearing were as follows: 1) an in-line speaker annotation separated by a colon from their respective lines, with out-of-character descriptions sitting on a separate line in either round or square brackets (displayed in Figure 2), and 2) a table format, with the first column listing the speaker (or being empty in case of an out-of-character description), and the second column listing the respective lines (displayed in Figure 3).

Speaker 1: "A sentence spoken by speaker 1. A second sentence spoken by speaker 1."
An out-of-character description. A second line of the description.
Speaker 2: "A sentence spoken by speaker 2. A second sentence spoken by speaker 2."

Figure 2: An in-line formatting of the source data.

Speaker 1	A sentence spoken by speaker 1. A second sentence spoken by speaker 1.
	An out-of-character description. A second line of the description.
Speaker 2	A sentence spoken by speaker 2. A second sentence spoken by speaker 2.

Figure 3: A table formatting of the source data.

3.3   Data Wrangling

Data wrangling describes a process of transforming and filtering the collected raw data into data ready to be used for further tasks. It is crucial to determine the desired end-point and format of the data before beginning the wrangling process itself. For the purposes of this thesis, two different data wrangling processes were applied to the source data, with the latter process using the data from the former as its source. This was necessary, as the different analyses implemented further down the line required a

separate format in order to function properly. However, even before such split could be made, the first step to be taken was a transformation of the source data to a unified format. With the high number of the source files, it was determined the best approach was transforming the data automatically using a self-developed script tailored for the purpose.

As the desired output format was a TXT file, it was determined the unified format should follow the one shown in Figure 2. Other available options for a TXT file (since a table format is not supported) was a comma-separated file (which would have the potential for a high level of disruptions, since the source data features a large number of commas), and a tab-separated file (which features the least amount of potential faultiness). It should be acknowledged that the chosen format has a slim potential for faultiness since colons do appear in the source format even in other contexts, however, due to the further handling of the data, it posed no risk of affecting the outcome of the analysis. A tab-separated format would be a better formatting choice from the point of additional workload in a vacuum, however, as the majority of the source data already corresponded to the colon-separated format, the minor faultiness would have already applied were the data transformed to a tab-separated format beforehand, and it was thus chosen to stay with the colon-separated format.

In order to change the formatting of the transcript, it was first required to isolate it from the rest of the webpage, as it featured additional information pertaining to the transcript but not being a direct part of it. As the individual wikis shared little consistency between each other, it was necessary to manually identify the markings used to differentiate transcripts from additional information for each television show separately, and to then subsequently divide the shows to smaller groups to be handled separately. In total, three such groups were devised, and each individual show was additionally checked to be converted into the proper format if necessary. Since a table format converted into plain text was represented as separate lines, special care had to be put into assigning each line to its rightful speaker, and to not mix those with out-of-character descriptions, other speakers' names, or other speakers' lines.

With all the individual episode transcripts following the format shown in Figure 2, the data was thus ready for gender analysis described in subsection 4.1. However, co-reference resolution, described in subsection 4.2, has a need for a more specific format of data, as the implemented automatic tools required a specific formatting in order to acknowledge for the present metadata, such as the annotated speakers. The reasoning for the used format (displayed in Figure 4), as well as the process that led to devising such a format, will be described in the appropriate section, as it is necessary to first explore how automatic co-reference analysis and the specific tool used in this project work.

Speaker 1 says: "A sentence spoken by speaker 1."
Speaker 1 says: "A second sentence spoken by speaker 1."
An out-of-character description. A second line of the description.
Speaker 2 says: "A sentence spoken by speaker 2."
Speaker 2 says: "A second sentence spoken by speaker 2."

Figure 4: The formatting used for coreference resolution.

4 Methods

4.1 Gender Analysis

Gender Definition and Data Collection: In order to perform gender analysis, it was integral to at first establish the definition of gender used in this project, and to subsequently assign a gender group to each speaker. There are three groups of genders classified for the purpose of the following analysis: male, female, and non-binary (encompassing all genders falling outside of the gender binary, as well as genderless characters). It should be noted that sexless characters are not necessarily genderless and vice versa – while gender and sex overlap for the majority of characters in the selected television shows, the analysis focuses on gender and ignores sex in case those two characteristics do not align with each other. As there rarely exists definitive data for each character, this thesis follows these guidelines in order of importance: 1) canonical gender provided either out-of-universe by the television show's creators, or in-universe inside of the show itself, and 2) if no such canonical information is provided, it is assumed pronouns he/him/his refer to male characters, she/her/hers refer to female characters, and singular they/them/their/theirs refer to non-binary or genderless characters. The names of all speakers were automatically gathered from all episode transcripts along with the total number of words spoken by that speaker, and subsequently exported into a separate file for further use.

To analyze any meaningful statistics about gender, the necessary first step was to devise a method to either manually or automatically assign a gender group to each individual speaker. Initially, it was planned to gather gender information by automatically searching the associated webpages of each individual character on their respective fan wikis. However, this process was hindered by several complications and eventually was deemed too faulty to use reliably. The largest flaw was the frequent mismatch of website names and character names used in the source transcript. In many cases, a character was referred to only by a part of their name, e.g. "Anakin" in place of "Anakin Skywalker". Additionally, many significant characters appeared in the transcript who either did not have an associated webpage (due to the incompleteness of the wiki itself, the character page policy of the given wiki, or simply due to several different mutually related characters sharing a single webpage), their gender was not listed on their website, or their name was in other ways different from the name of the webpage their character profile was located at. It was decided that while this method would be capable of producing sufficient results, it would also require heavy manual correction following the automatic search part (and the whole process would possibly become even more time-intensive than a fully manual approach), and thus a different approach was more likely to be suitable.

The method used in the final version of the project was an automated gender guesser accompanied by manual checking and filling in of missing information. Since a large portion of the source data contained out-of-character descriptions, it was possible to automatically extract an exact number of gender or pronoun mentions occurring in all instances of descriptions accompanying a specific character's lines using a self-developed script. With the guidelines established above, a statistic was established for every character and their gender was established based on the prevalent gender or pronoun usage. However, it is important to note this process is not definite, and requires subsequent manual check-up, as descriptions occasionally include mentions of genders or pronouns belonging to characters other than the speaker. While the guesser was mostly accurate

(as mentions of the speaker occurred in larger numbers than mentions of other characters), several errors had to be manually corrected. In addition, the guesser was implemented solely for male and female genders, as the combination of non-binary genders being too infrequent in the source material and the large amount of noise generated by mistaking the plural "they" for the singular "they" has proven to decrease the accuracy of the guesser. A manual check-up was performed on the output data, correcting the gender of non-binary characters and inserting gender information for television shows whose transcripts lacked out-of-character descriptions. Additionally, it was chosen to only include named characters with more than 100 words (words being all alphanumeric sequences of characters separated by a whitespace) spoken across the entire television show, as characters below this threshold form a large percentage of speakers and a minuscule percentage of on-screen appearance, and have thus a high potential to bloat the amount of data needed to be manually inspected, and to skew the following statistics. The exclusion of non-named characters was made on the basis of both the high chance of smaller importance, as well as the possibility of one label belonging to more than one character, possibly to characters of differing genders. For example, a speaker referred to as "Guard" can possibly refer to several different guards even in the span of a single television show or a single episode, and it is possible those characters do not share a gender. Additionally, even if such a label only referred to a singular character, it would require a large amount of manual work examining the video footage of the given episode to determine the speaker gender (as non-named characters do not possess a webpage of their own on their respective wikis) for comparably minuscule gain (non-named characters tend to have a significantly smaller number of words spoken than named characters, as characters with a significant presence in the source material have most often been given names).

While non-binary representation is an aspect of gender representation worth investigating, the resulting data have shown very few results on the matter. No individual television show has contained more than a handful of non-binary or genderless characters according to the previously established criteria, and the majority of television shows contained none. With the acknowledgement non-binary representation is an important subject to study, and with the admittance the source material has proven to not contain enough data to include such a study in the following analysis, this thesis decides to only focus on the representation and relation of male and female genders from this point onward, and excludes all non-binary or genderless characters from its statistics.

Metrics and Statistics: Gender analysis can be performed in several ways, each one yielding a different type of information about the source material. This thesis aims to analyze the data from multiple different angles to form a larger picture about the state of gender representation and to answer several questions that can help shed more light into the exact specifics of the issue. It is important to note that because this subsection deals with no natural language processing tools and only works with raw numbers, the data gathered in this part of the analysis largely focuses on quantity over quality (the latter of which will be explored more thoroughly in subsection 4.2). The first and most basic statistic that is possible to extract from the source data is comparing the total number of words spoken by all speakers of a given gender in one specific television show to the same statistic of a different gender, and providing the results for every cartoon separately (further referred to as *words per gender*). With access to transcript data and without using actual episode video footage and complex automated tools, this metric is the closest

approximation to the amount of time characters of different genders spend speaking in their respective television shows. However, as the total number of words spoken in each analyzed cartoon differs drastically, in order to extract meaningful comparisons, it was decided to handle such analysis proportionally, and simply display the percentage of words spoken instead of raw numbers.

Additionally, it is possible to analyze this data in a more thorough way by calculating the statistic for each individual episode of each individual show, and plot the ratios between male and female speakers separately for each cartoon. With the help of such plots, it is possible to get a more accurate look at gender representation and the spread and distribution of the data. While the total ratio between words said by male and female speakers does share important information, it lacks the ability to distinguish between shows with a consistent level of representation close to the average in each episode, and shows with highly fluctuating ratios. In addition, the distribution of the individual ratios does not have to follow a linear pattern, and it is possible for some shows to follow a more complex distribution.

With the proportional number of words spoken by speakers of each gender, it was possible to observe gender representation in terms of total words or time spoken by a specific gender. However, this single statistic is not all-encompassing, and has to be accompanied by others in order to get a clearer picture about the state of gender representation. The focus of the second statistic was to capture the number of speakers themselves that a given television show contained (further referred to as *speakers per gender*). Its focus is on showing how many characters appear of different genders overall, and it thus distances itself from the length of their presence, their importance, or any other characteristic of a given speaker (aside from the earlier character filters applied, namely the character must have been named and must have spoken at least 100 words throughout the entire length of the television show). It illustrates an idea of how many characters can a person of a given gender relate to in a given cartoon.

Another important metric chosen for the analysis was an average number of words spoken by an individual speaker (further referred to as *words per speaker*). However, it was necessary at first to consider what method of averaging was the most representative of the source data. As supported by the gathered data, television shows feature a small number of main characters recurring in most, if not all, episodes of the given shows, and their individual word counts are many times larger than those of any other character, and thus an arithmetic mean value of words spoken per speaker would be disproportionately influenced by those main characters. In order to better represent the data and to achieve more meaningful statistics with regards to all investigated speakers of a given show, a median value was chosen for the analysis.

4.2 Co-reference Analysis

The process for handling co-reference resolution went through several iterations in order to develop the best approach for the project. As the large amount of source material proved difficult to manually handle, it was devised to handle this process through automated tools. Initially, it was proposed to make use of the CoreNLP pipeline (Manning et al., 2014), a customizable tool capable of many different aspects of natural language processing, as it contained the required tools for co-reference resolution (Lee et al., 2011; Raghunathan et al., 2010; Recasens et al., 2013). The tool was implemented through

the use of the command line, and while the specifics of its usage have changed over the course of the project, the tool choice remained the same throughout it. It was assumed from the beginning that due to the non-standard formatting of the source files (as they have contained in-line speaker annotation and consisted of mostly or purely dialogue), a degree of experimentation and adjustment had to be accounted for and implemented in order to get the correct results.

There are three possible systems to be used within the co-reference analysis annotator of the CoreNLP pipeline – some of them are only available in a limited selection of target languages, however all three of them are available for use with English. The first one is a multi-pass sieve rule-based system (further referred to as *deterministic*), more thoroughly described in Lee et al. (2013). The basis of this system is a three-stage process: first is mention detection, in which a collection of sieves (an ordered group of filters that each try to extract a mention in a different manner ranging from an exact string match to pronominal co-reference resolution) attempt to single out specific words or word groups that refer to an entity. The second stage is co-reference resolution by itself, which attempts to link references together, and is then followed by the last stage, post-processing, which adjusts the output based on the desired parameters. This system is the fastest at co-reference annotation, however it requires constituency parsing data as a source in order to function, which take longer to produce.

The second system is statistical, and it consists of a machine learning model with a large set of features including rule-based ones such as in the deterministic model, but also distance, syntactic, semantic and lexical features (Clark & Manning, 2015). It works by iterating through the document multiple times, documenting mentions and deciding on adding co-reference links between the current and past mentions. It is customizable in several degrees, such as the greediness of the model or the maximum distance a link will be considered for. While this system takes longer to create co-reference links than the deterministic system, it only requires dependency parsing to function properly, and thus handles the entire process faster.

The third system is based on a neural network (Clark & Manning, 2016a, 2016b) consisting of two different encoders: a mention-pair encoder containing an input layer and three hidden layers of rectified linear (ReLU) units, which is subsequently connected to a cluster-pair encoder. While this model has a high chance to be the most accurate, it is also significantly more time-intensive than the previous two models, as it requires both the same constituency parsing data as the deterministic model, and the co-reference resolution by itself is significantly more time-intensive as well.

Before an analysis of the models could have been performed, it was first important to establish the method for selecting and extracting results, and the process that led to devising such an approach. The co-reference models output their results in the same format, and claim that a reference in the given sentence is pointing to the same object as another earlier established reference. This action can present equality between different types of objects, not all of which are compatible with the analysis. The end goal of the analysis is to count the occurrences of characters speaking to each other, and such can be expressed and captured by the system in two different ways: a character speaking out any of the pronouns "you, your, yours", or a character speaking out the addressee's name. However, while the first way is relatively unambiguous, as speaking out the mentioned pronouns in most cases signifies speaking to a second person, the other way is not, as a mention of a name can both be used as a method to address a person, and a way to

mention the given person without necessarily addressing them. As such, this thesis has chosen to dismiss name mentions for their ambiguity, and has chosen to only focus on second person pronoun mention. In addition, in order for a co-reference mention to yield useful information to the analysis, it was necessary for the earlier established reference to be a character name, and not a pronoun. Thus, the only co-reference links that were considered were ones in a format such as "(you) -> (<name>)" or similar.

At first, the input format used was identical to the input format for gender analysis, and a sample input was analyzed through all three systems. However, it was clear from the results that while a comparison of the three models could be made at that point, the pipeline clearly did not handle the source data in the intended manner, and thus a data transformation was in order before any model comparisons would yield results. The models at first split the source text into individual sentences and creates dependency and constituency parses for those. In the intended way, the model includes information about the annotated speaker in order to take the data into account while performing co-reference analysis. However, several errors occurred that necessitated a change in approach to the source data format. Firstly, the split sentences failed to account for the annotated speaker data in many cases. This occurred due to the way the annotations were presented to the model, as they appeared in front of the entire dialogue line separated by a colon, and thus this information only appeared in the first sentence of a given line (see example in Figure 5). Additionally, since the sentence split also split apart the quotation marks belonging to each other, the model attributed the singleton quotation marks to other sentences. This has among others caused the model to perceive the speaker annotations as dialogue lines, and the lines themselves as descriptions. This issue on its own has caused the need for a different format, however, it was also determined that while the model did account for the speaker annotations in some cases, it was clear the model needed a clearer guidance on the type of sentence it was supposed to be analyzing (a dialogue line).

Source data:
Speaker 1: "The first sentence. The second sentence."
Speaker 2: "The first sentence. The second sentence."

Analysis:
Sentence #1:
Speaker 1: "The first sentence.
Sentence #2:
The second sentence.
Sentence #3:
"Speaker 2: "The first sentence.
Sentence #4:
The second sentence."

Figure 5: The co-reference model not properly accounting for speaker annotations.

Based on these two flaws in the input data, a new format had to be devised and properly tested. In order to fix the first issue, it was planned that while handling the source data, individual sentences inside one line would be split into separate lines, with the speaker annotation appearing in front of all of them. However, for this approach to be implemented, it was at first necessary to test which characters or sequences of characters cause the co-reference model to detect a sentence boundary. While the task seems simple at first, several oddly-behaving patterns were discovered that had to be

accounted for, such as an ellipsis (. . .) or a dash (—), that did not cause a sentence break trigger neither when occurring in the middle of a sentence (correct behavior), nor at the end of a line (incorrect behavior). Such characters had to be followed by a simple period when occurring at a line break in order to allow the model to treat the current and following sentence separately. In addition, the word "says" has been added to each speaker annotation (such as "Anakin says:" in place of "Anakin:") in order for the model to recognize the meaning of the annotation more clearly.

After applying the necessary changes to the source format and performing the test analysis again, it was apparent that the format was appropriate and the models handled the input data correctly. The subsequent step was to make a comparison of the three models on a small sample size, compare their speed and results, and select the best-performing model for the entire analysis. A sample size of 16 episode transcripts, one from each television show, was randomly selected and a test was performed. Since the speed of the models is affected by other processes running on the same device at the same time, all testing was conducted in as equal conditions as possible, with as few other processes running as possible and with no human interference. The metrics measured were the speed of the model (taking into account both the time of the co-reference annotator itself and the other required annotators), as well as the number of results acquired, and are displayed in Table 3.

Model	Time (s)	Result Count
deterministic	1013	21
statistical	610	198
neural	2061	76

Table 3: Results comparison between the three co-reference models.

While the number of results returned by itself does not mean a better accuracy, the results from the models were manually checked in addition to producing the aforementioned statistics. It was determined that the statistical system was not only significantly faster than the other two models and better at producing a large number of results, its results were of no worse quality than the other models, and it was thus chosen as the model to perform the full-scale analysis with.

5   Study

5.1   Setup

The analysis was performed on the corpus introduced earlier in the thesis, and made use of both self-developed scripts and the CoreNLP pipeline's statistical co-reference resolution model. The first part of the study, gender analysis, has utilized data in the format displayed in Figure 2, while the second part, co-reference analysis, has utilized data in the format displayed in Figure 4. Each animated television show has been handled separately.

For gender analysis, several different statistics were produced: first, the total percentage of words spoken by each gender over the course of the entire show, as well as ranked graphs covering the percentage of words spoken by each gender in each individual episode of each individual show. Additionally, statistics were produced for the total percentage of speakers of each gender, as well as the average number of words spoken by
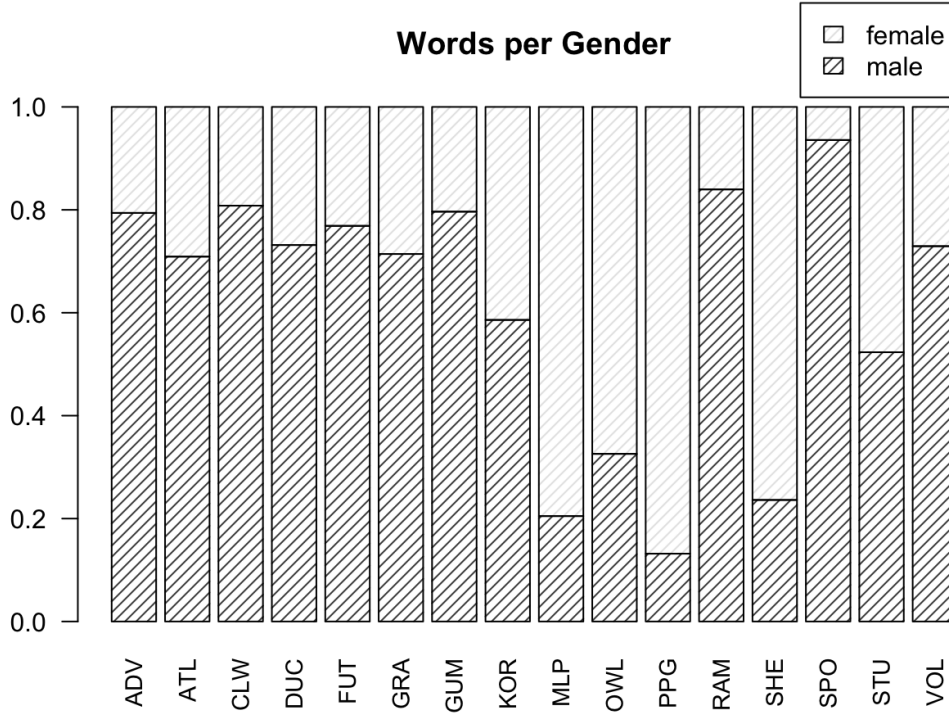
Figure 6: A proportional percentage of words spoken by each gender.

each speaker. Only named speakers with over 100 words spoken over the course of the entire show were considered for the analysis.

In the latter half, the study focuses on co-reference resolution, using automated tools to capture all co-reference links throughout the data. Afterwards, the links are filtered to only include relevant second-person data, and the results are presented for the total percentage of gender pairs, i. e. measuring the percentage of words that speakers of a gender speak addressing speakers of another gender (male -> male, male -> female, female -> male, and female -> female).

5.2   Results and Discussion

Gender Analysis: Firstly, an analysis was made of the number of words spoken by each gender. However, as the total number of words varies between episode shows, the statistic is presented proportionally, i. e. the percentage of words speakers of one gender say compared to the total number of words said. The results can be seen in Figure 6, and each television show is presented in its own column, independent of the others.

According to the results, the ratio fluctuates significantly between different shows, reaching from 0.1 (10% of words spoken in the television show are spoken by men) to 0.95 (95% of words spoken in the television show are spoken by men). However, the data does not seem to follow a linear pattern when ranked, and many of the shown bars are clustering around specific values. The most notable would be a value of 0.75, around which many of the ratios congregate. The other, albeit less frequent value, is around 0.2,
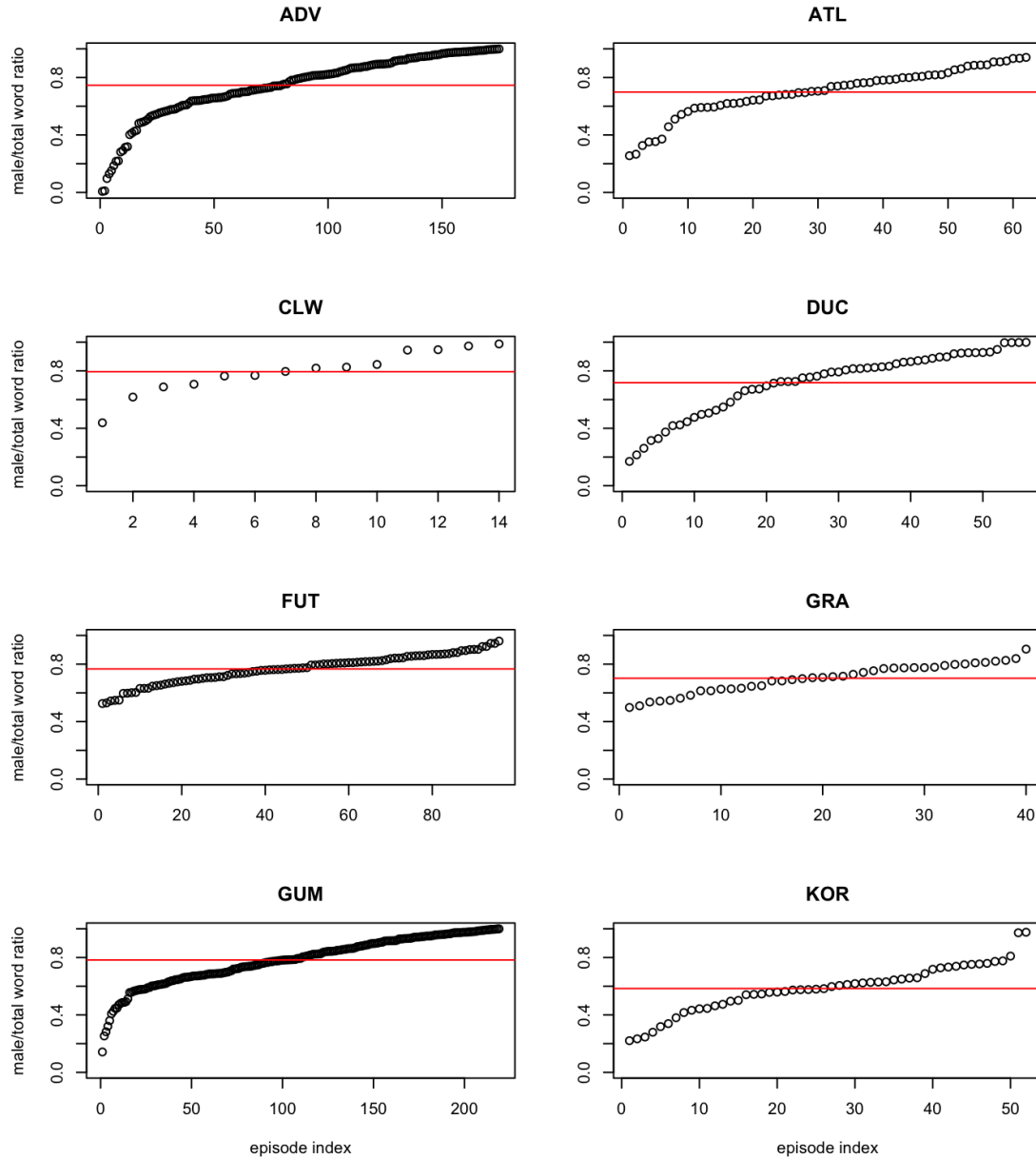
Figure 7: The ratio of words spoken by male speakers to the total number of words spoken per episode. Individual ranked charts for each cartoon and individual points for each episode with a mean ratio line. Part 1 of 2.

with only a handful of shows that do not belong to either of those two clusters. This finding is interesting, because while a wide range of ratios could be expected beforehand, the clustering is rather jarring. It is possible that this phenomenon could be caused by the occurrence of main characters in almost every television show – characters recurring in almost every episode and thus having a large impact on the total number of words. Unless the collection of main characters – the main cast – is perfectly balanced in terms of gender, it is possible this has a large influence. To some extent, however, it is possible to measure this influence, specifically with variance (displayed in Figures 7 and 8).

For each cartoon, a ranked plot was made with the ratios of male to total words spoken
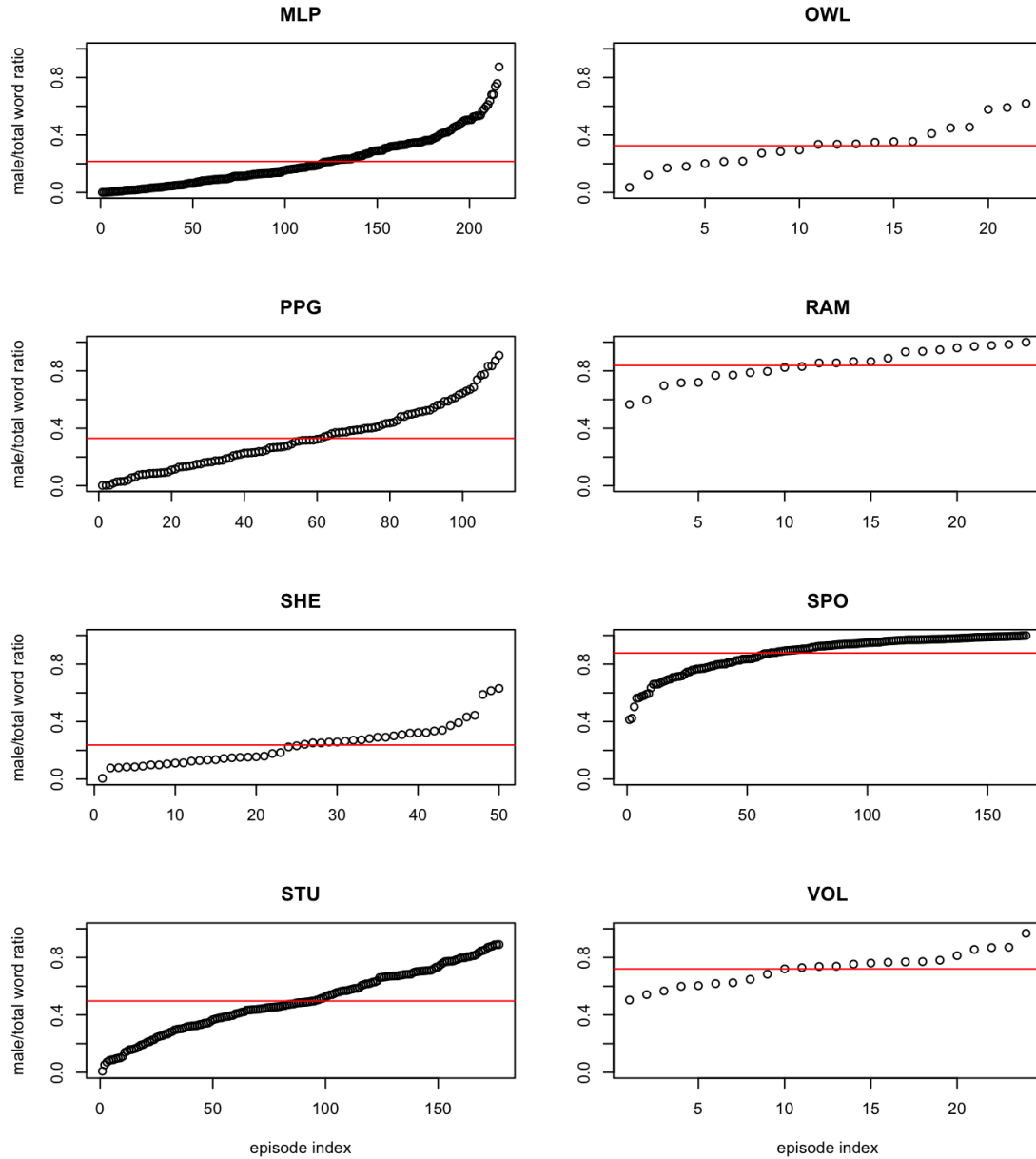
Figure 8: The ratio of words spoken by male speakers to the total number of words spoken per episode. Individual ranked charts for each cartoon and individual points for each episode with a mean ratio line. Part 2 of 2.

in each individual episode. This statistic is thus an extension of the graph in Figure 6, merely showing the ratios episode by episode instead of presenting a single value. With a more detailed graph, it is possible to see if individual episode values follow a specific trend – however, it is important to note that the episodes are ranked by value, not by time of airing, and thus do not speak about the evolution of gender representation over time.

The results are to an extent surprising, as several different patterns can be observed in the different graphs. First off, many of the graphs are relatively linear, but with vastly different skews. It can be deduced that even if two different shows have a similar
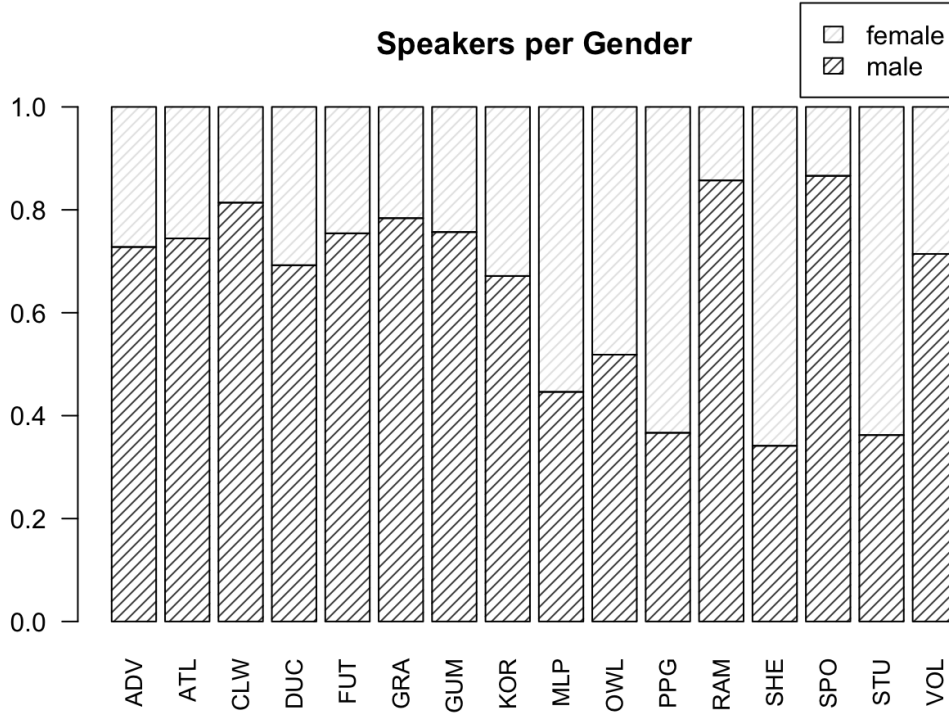
Figure 9: A proportional percentage of speakers of each gender.

total ratio, the actual episode-to-episode representation can differ vastly, as some shows feature episodes with mostly the same level of representation each episode (see RAM) and thus closer to the average ratio of the entire show – represented by a red line in the graphs, while others feature episodes all across the spectrum (see STU). While it is not possible to say that shows with a small skew are highly influenced by their main cast (as the stable nature of their graphs can either be caused by a large influence of the main cast, or a stable level of named-character gender representation outside of it), it is safe to say that shows with a large skew are either to a large extent unaffected by their main cast, or feature many episodes where only a part of the main cast is playing the main role, thus increasing the variance of its gender representation.

However, this is not the only interesting finding in the graphs: several graphs do not seem to follow a linear pattern at all (see ADV, GUM as an example). Instead, their graphs more closely resemble a logarithmic scale, with a relatively stable ratio throughout the majority of the episodes and a trail of either increasing (see SPO) or decreasing (see MLP) female representation – one that is more significant than an outlier. It is debatable why this occurs, and more data, specifically time of airing data would be required to make any further conclusions. This behavior can be caused by occasional episodes which largely focus on a different set of characters – although this idea would most likely not result in such pattern, as the episode statistics would more likely cause a jump rather than a gradual increase/decrease. It is also possible this behavior is caused by a change in main cast, or – if supported by times of airing – by a gradual move toward a different level of gender representation.
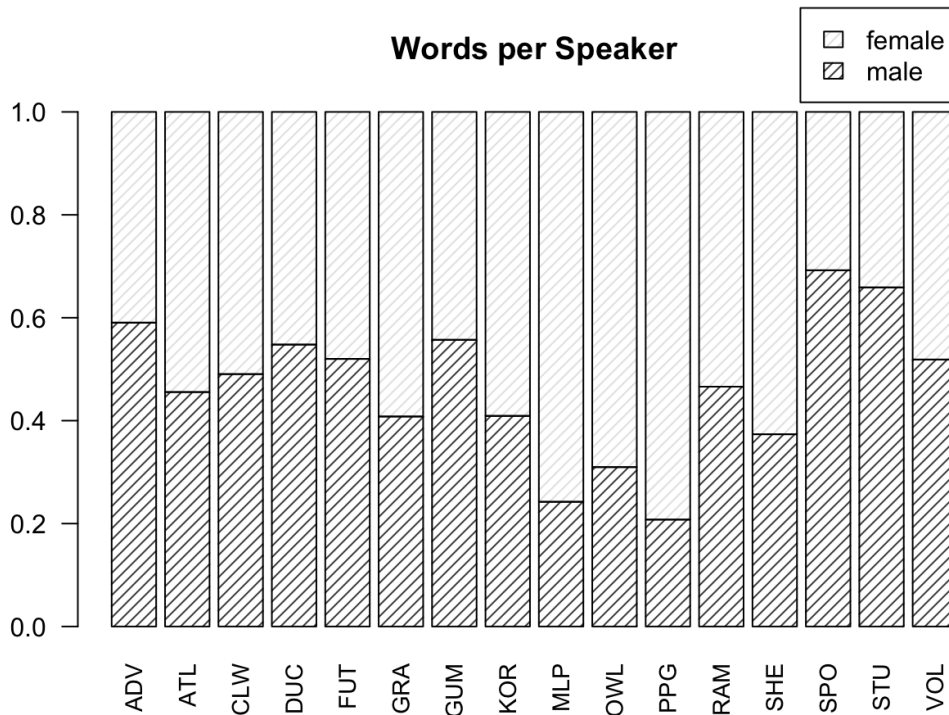
Figure 10: A proportional chart of the average number of words spoken by speakers of each gender.

The next statistic in the study is the one of speakers themselves. Moving entirely away from how many words a speaker says over the course of the cartoon (aside from the limitation of a minumum of 100 total words spoken), this statistic (displayed in Figure 9) covers the ratio of male speakers versus female speakers for each individual show.

As with the ratio of words spoken by different genders earlier, the data in this graph fluctuate as well, although to a significantly lesser degree, only from about 0.35 to 0.9, showing the ratio is less extreme when only taking into account the speaker numbers themselves. Such a statistic shows some other interesting behaviors, such as the fact that while the ceiling for male representation has not moved much from Figure 6 to this one (only from 0.95 to 0.9), the floor has moved to a much bigger degree (from 0.1 to 0.35). From this, it can be deduced that even when shows consist mainly or almost entirely of lines spoken by one gender, there is still a floor for the minumum number of speakers of each gender, and this floor is higher for male characters. Additionally, the clustering pattern observed in Figure 6 is even more apparent here: all of the displayed shows can be split into two clusters, a larger one centered at around 0.75, and a smaller one clustering at around 0.4.

So far, the data indicates that the ratio of words said by speakers of each gender highly correlates with the ratio of the speakers themselves. While this might be obvious, as word numbers rise with the number of speakers, it indicates that the ratio of the average number of words spoken by each individual speaker separated by genders should be significantly more even than the previous two statistics. If that was the case, it would
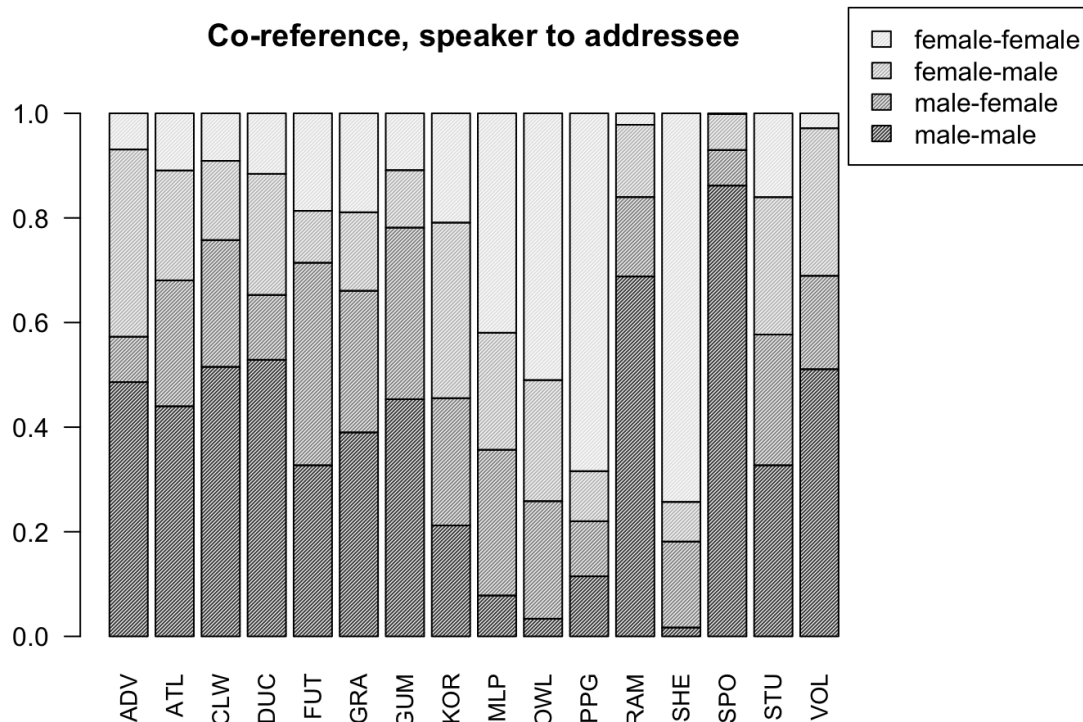
Figure 11: The number of words spoken by speakers of one gender while addressing another, gender presented as speaker-addressee.

hint toward the issue of lacking gender representation not necessarily being a factor of too few words spoken by speakers of one gender, but of the number of the speakers themselves. Such a graph can be seen in Figure 10, although in order to minimize the great influence of the main cast, the median value was used instead of the mean, and as such the data in the chart represents the majority of the characters more accurately.

The prediction for the most part seems accurate, as many of the bars cluster at around the 0.5 mark. While there are several shows that dip below this value, it is possible to say that the average number of words spoken by speakers is significantly more even than the previous statistics, and even slightly skews toward female. As can be seen, at least as far as the presented sample is concerned, the issue of gender representation does not lie in individual female characters having less opportunity to speak than male characters, but more so the imbalance of the numbers of male and female characters in the first place.

Co-reference Analysis: Moving to the matter of co-reference, the question shifts from the presence of characters by itself to the matter of their words. The graph is shown in Figure 11, and displays all four types of dialogue this study is interested in: male speakers addressing male characters, male speakers addressing female characters, female speakers addressing male characters, and female speakers addressing female characters. The results are interesting especially when compared with the number of words spoken by speakers of each gender (presented in Figure 6). While the general trend still holds for many of the shows – characters speak in similar amounts proportionally to one another –

there are significant disparities and noteworthy occurrences in the graph. At first glance, it can be seen that in several shows, a specific interaction is almost completely absent (best seen with the lack of female-female representation in SPO, but also visible in RAM, SHE, VOL and others).

Curious comparisons can also be made between several shows whose share of gender representation was very similar in terms of total number of words spoken. As one example, DUC and KOR, while almost identical in Figure 6, have significant differences in co-reference. While KOR is possibly the most well-distributed bar in the entire chart, with a share of about 0.25 for each category, over half of the bar of DUC is taken up by male-male interaction, and overall for each gender, their interaction with male addressees is significantly higher than with female ones. Additionally, a comparison can be made with VOL, which features almost no female-female interaction, even though the word share is relatively equal. In contrast, GRA, which features a smaller share of total words spoken by female speakers, has a significantly larger portion of both female-female and male-female interactions in comparison to both DUC and VOL. It is clear that while the inclusion of a larger percentage of characters of one gender allows for greater significant interaction with said characters, it is not equal, and shows with a higher percentage of speakers of a given gender do not necessarily lead to more diverse interactions.

6   Conclusion

For the purposes of this thesis, a corpus of the speaker-annotated transcripts of 16 different animated television shows was gathered and analyzed. Both gender and co-reference analysis were conducted using statistical measures as well as automated natural language processing tools. It has been found that for the sample collected, gender representation issues are significantly more impacted by the disparity of the number of speakers in a given television show rather than the amount that speakers of each gender speak on average. Further patterns (such as non-linear ranked gender representation in individual shows, as well as significantly differing co-reference statistics for television shows of similar gender representation) were observed that would serve as an interesting starting point for further research into more specific patterns. Overall, this thesis serves as a basis for further research into the topic, as the insights uncovered in the study deserve additional exploration with both more depth and a larger sample.

References

Adventure Time Wiki. (n.d.). *Adv transcripts.* Retrieved from `https://adventuretime.fandom.com/wiki/Category:Transcripts`

Agarval, A., Zheng, J., Kamath, S. V., Balasubramanian, S., & Dey, S. A. (2015). Key female characters in film have more to talk about besides men: Automating the bechdel test. In *Association for computational linguistics.*

Avatar: The Last Airbender Wiki. (n.d.). *Atl transcripts.* Retrieved from `https://avatar.fandom.com/wiki/Avatar_Wiki:Transcripts`

Avatar: The Legend of Korra Wiki. (n.d.). *Kor transcripts.* Retrieved from `https://avatar.fandom.com/wiki/Category:The_Legend_of_Korra_episode_transcripts`

Baldwin, B. (1997). Cogniac: High precision coreference with limited knowledge and linguistic resources. In (p. 38-45).

Bechdel, A. (2013, 11). *Testy.* Retrieved from `https://dykestowatchoutfor.com/testy/`

Bechdel, A. (circa 1985). Portion of "the rule" in dykes to watch out for. *Bechdel's Flickr.*

Clark, K., & Manning, C. D. (2015). Entity-centric coreference resolution with model stacking. In *Association for computational linguistics (acl).*

Clark, K., & Manning, C. D. (2016a). Deep reinforcement learning for mention-ranking coreference models. In *Empirical methods on natural language processing.* Retrieved from `https://nlp.stanford.edu/pubs/clark2016deep.pdf`

Clark, K., & Manning, C. D. (2016b). Improving coreference resolution by learning entity-level distributed representations. In *Association for computational linguistics (acl).*

Crystal, D. (1997). *A dictionary of linguistics and phonetics* (4th ed.). Cambridge, MA: Blackwell Publishing.

DuckTales Wiki. (n.d.). *Duc transcripts.* Retrieved from `https://ducktales.fandom.com/wiki/Category:Transcripts`

European Women's Audiovisual Network. (2013). *Swedish cinemas take aim at gender bias with bechdel test rating.* Retrieved from `https://www.theguardian.com/world/2013/nov/06/swedish-cinemas-bechdel-test-films-gender-bias`

Ferraro, R., & Kane, M. (2013, 08). *Glaad introduces 'studio responsibility index,' report on lgbt images in films released by 'big six' studios.* Retrieved from `https://www.glaad.org/releases/glaad-introduces-studio-responsibility-index-report-lgbt-images-films-released-big-six`

Futurama Wiki. (n.d.). *Fut transcripts.* Retrieved from `https://theinfosphere.org/Episode_Transcript_Listing`

Garcia, D., Weber, I., Rama, V., & Garimella, K. (2014). Gender asymmetries in reality and fiction: The bechdel test of social media. In *Icwsm.*

Gravity Falls Wiki. (n.d.). *Gra transcripts.* Retrieved from `https://gravityfalls.fandom.com/wiki/Category:Transcripts`

Jurafsky, D., & Martin, J. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (Vol. 2).

Lappin, S., & Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics.*

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., & Jurafsky, D. (2013).

Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics 39(4)*.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., & Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Conference on natural language learning (conll) shared task.* Retrieved from `pubs/conllst2011-coref.pdf`

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for computational linguistics (acl) system demonstrations* (pp. 55–60). Retrieved from `http://www.aclweb.org/anthology/P14/P14-5010`

McGuinness, R. (2013, July). *The bechdel test and why hollywood is a man's, man's, man's world.* Retrieved from `https://metro.co.uk/2013/07/18/movie-lovers-hit-back-at-hollywood-misogyny-with-the-bechdel-test-3886103/`

My Little Pony: Friendship is Magic Wiki. (n.d.). *Mlp transcripts.* Retrieved from `https://mlp.fandom.com/wiki/Category:Transcripts`

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., & Manning, C. (2010). A multi-pass sieve for coreference resolution. In *Empirical methods in natural language processing (emnlp).*

Recasens, M., de Marneffe, M.-C., & Potts, C. (2013). The life and death of discourse entities: Identifying singleton mentions. In *North american association for computational linguistics (naacl).*

Rick and Morty Wiki. (n.d.). *Ram transcripts.* Retrieved from `https://rickandmorty.fandom.com/wiki/Category:Transcripts`

She-Ra and the Princesses of Power Wiki. (n.d.). *She transcripts.* Retrieved from `https://she-raandtheprincessesofpower.fandom.com/wiki/Category:Transcripts`

SpongeBob SquarePants Wiki. (n.d.). *Spo transcripts.* Retrieved from `https://spongebob.fandom.com/wiki/List_of_transcripts`

Steven Universe Wiki. (n.d.). *Stu transcripts.* Retrieved from `https://steven-universe.fandom.com/wiki/Category:Transcripts`

The Amazing World of Gumball Wiki. (n.d.). *Gum transcripts.* Retrieved from `https://theamazingworldofgumball.fandom.com/wiki/Category:Transcripts`

The Owl House Wiki. (n.d.). *Owl transcripts.* Retrieved from `https://theowlhouse.fandom.com/wiki/Category:Transcripts`

The Powerpuff Girls Wiki. (n.d.). *Ppg transcripts.* Retrieved from `https://powerpuffgirls.fandom.com/wiki/Category:Transcripts`

Transcripts Wiki. (n.d.). *Clw transcripts.* Retrieved from `https://transcripts.fandom.com/wiki/Star_Wars:_The_Clone_Wars`

Voltron: Legendary Defender Wiki. (n.d.). *Vol transcripts.* Retrieved from `https://vld.fandom.com/wiki/Category:Episode_Transcripts`

Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Icwsm.*

Woolf, V. (1929). *A room of one's own.* John Wiley and Sons, Ltd.

Zeisler, A. (2016). *We were feminists once: From riot grrrl to covergirl, the buying and selling of a political movement.* PublicAffairs.