

Bachelor's Thesis

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

Author

Your Name Here

your-email-address-here

Supervisor

Zarah Leonie Weiß

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

August 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.

Contents

1	Introduction	1
2	Related Work	2
2.1	Bechdel-Wallace Test	2
2.2	Co-reference Analysis	2
3	Data	2
3.1	Collection and Types	2
3.2	Data Wrangling	5
4	Methods	6
4.1	Gender Analysis	6
4.2	Co-reference Analysis	8
4.3	Statistics	11
5	Analysis	11
6	Conclusion	11

List of Tables

1	A metadata table of the analyzed television shows.	4
2	Results comparison between the three co-reference models.	11

List of Figures

1	An in-line formatting of the source data.	3
2	A table formatting of the source data.	3
3	The formatting used for coreference resolution.	6
4	The co-reference model not properly accounting for speaker annotations. .	10
5	A proportional percentage of words spoken by each gender.	12
6	A proportional percentage of speakers of each gender.	13
7	A proportional chart of words spoken by each individual speaker.	14
8	A co-reference graph, speaker to addressee gender.	14

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 spread the values and ideals embedded in them onto its consumers, and at the very least give those values space in the public discussion, 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, 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 helps 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* (?), 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 an end-all-be-all 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.

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). 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 1 mentions related work with regards to both the Bechdel-Wallace test, and the natural language processing methods used in later sections. Section 2 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 3, and its results and analyses are shown in section 4. Finally, section 5 concludes the thesis.

2 Related Work

2.1 Bechdel-Wallace Test

2.2 Co-reference Analysis

3 Data

3.1 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 paper, 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.

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 1 under their specific 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 3.2). 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 1), 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 2).

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 1: 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 2: A table formatting of the source data.

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 1: A metadata table of the analyzed television shows.

3.2 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 1. 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 1, the data was thus ready for gender analysis described in subsection 4.1 – the reason for why this format was chosen for this particular analysis will also be described in this subsection, as well as the specific benefits to each individual method. However, the data to be used for 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 3), 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 3: The formatting used for coreference resolution.

4 Methods

4.1 Gender Analysis

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 "he" 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.

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.

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 (Recasens et al., 2013; Lee et al., 2011; Raghunathan et al., 2010). 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 and 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 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 4). 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 4: 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. 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 2.

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

Table 2: 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 better at producing results, it was also the fastest of the three, and it was thus chosen as the model to perform the full-scale analysis with.

4.3 Statistics

5 Analysis

6 Conclusion

References

- Clark, K. and C. D. Manning (2015). Entity-centric coreference resolution with model stacking. In *Association for Computational Linguistics (ACL)*.
- Lee, H., A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics* 39(4).
- Lee, H., Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky (2011). Stanford’s multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Conference on Natural Language Learning (CoNLL) Shared Task*.
- Manning, C. D., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60.
- Raghunathan, K., H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning (2010). A multi-pass sieve for coreference resolution. In *Empirical Methods in Natural Language Processing (EMNLP)*.

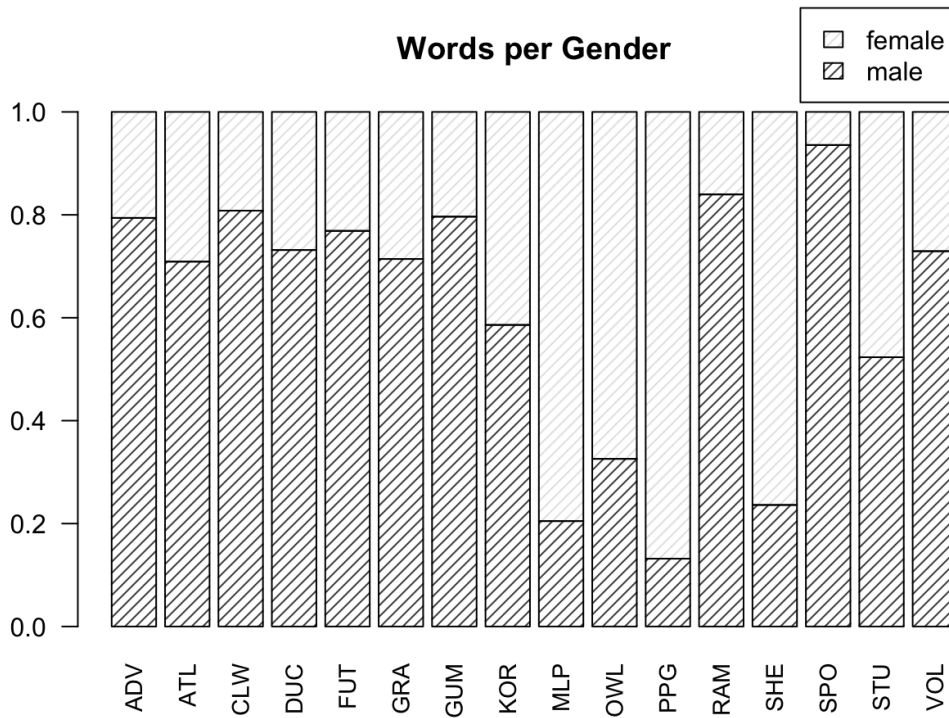


Figure 5: A proportional percentage of words spoken by each gender.

Recasens, M., M.-C. de Marneffe, and C. Potts (2013). The life and death of discourse entities: Identifying singleton mentions. In *North American Association for Computational Linguistics (NAACL)*.

Zeisler, A. (2016). *We Were Feminists Once: From Riot Grrrl to CoverGirl, the Buying and Selling of a Political Movement*. PublicAffairs.

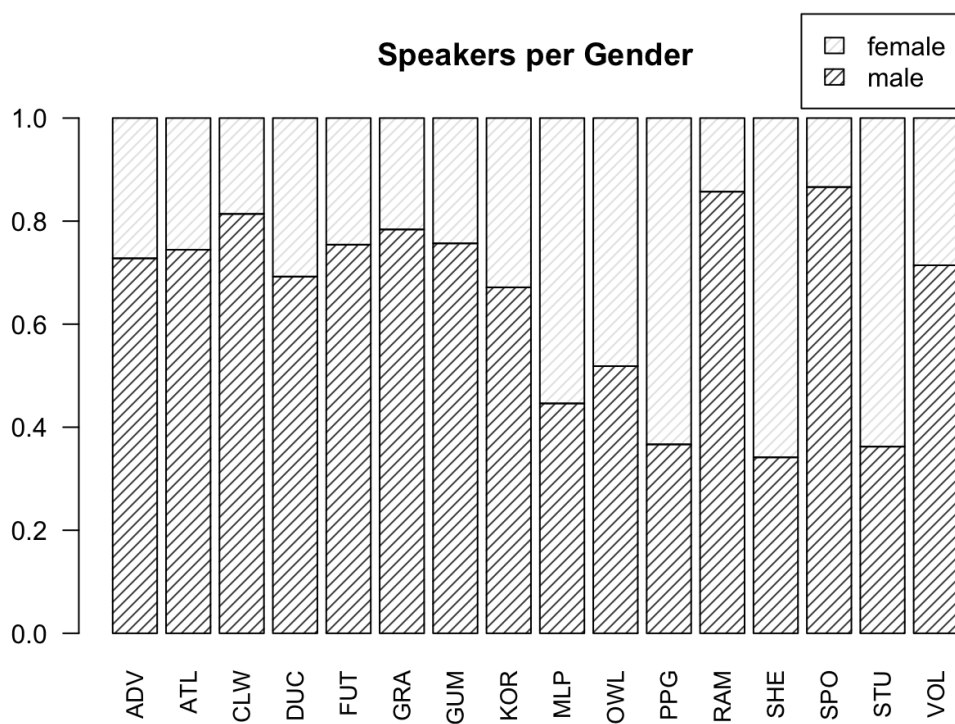


Figure 6: A proportional percentage of speakers of each gender.

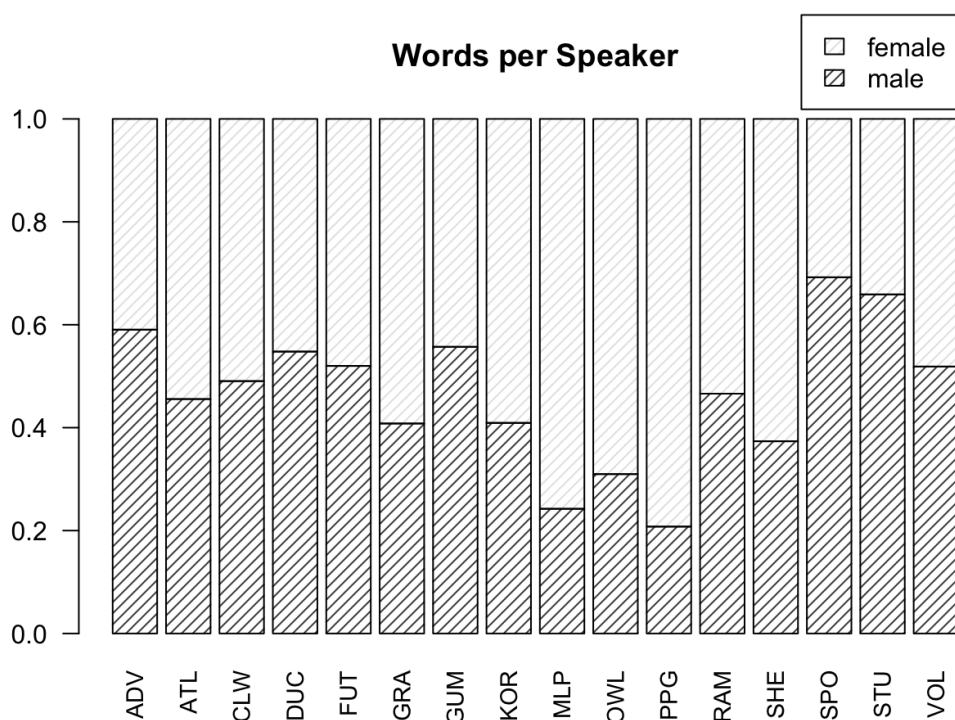


Figure 7: A proportional chart of words spoken by each individual speaker.

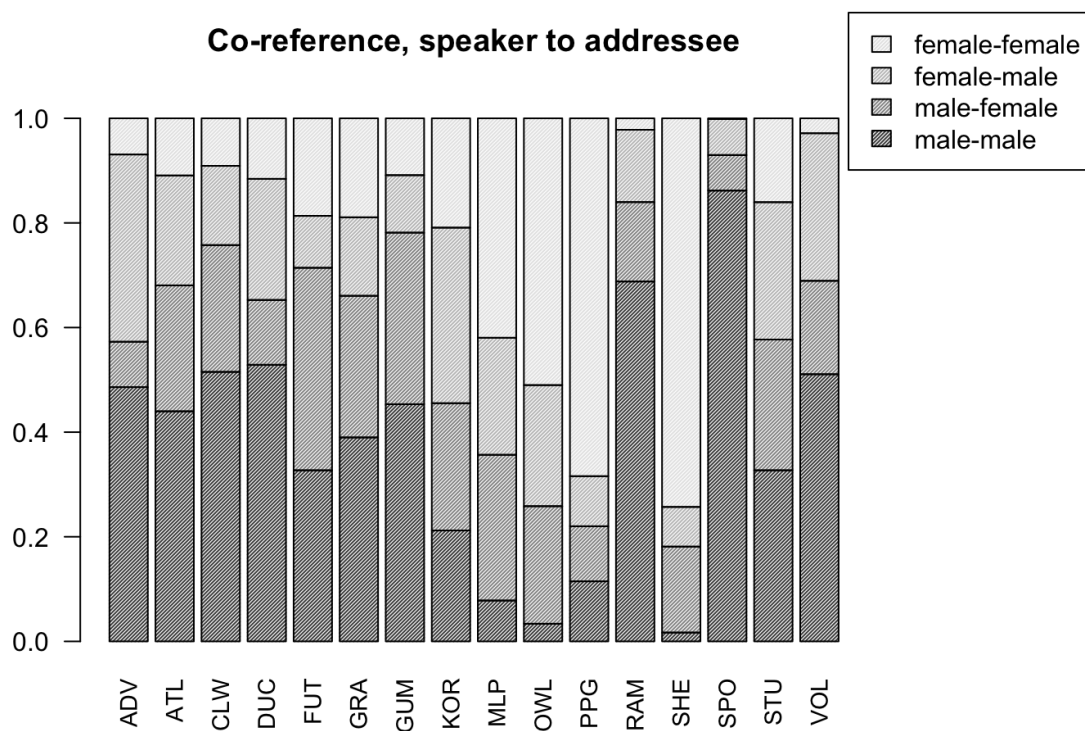


Figure 8: A co-reference graph, speaker to addressee gender.