**Thesis Draft**

<aside> 💡 **Notion-Tipp:** Verwende diese Vorlage, um schnelle Notizen zu machen, auf die du später zurückgreifen kannst. Erstelle im Handumdrehen umfangreiche Dokumente. Du kannst Links, Bilder, To-do-Listen und mehr einbetten. [Hier](https://www.notion.so/help/guides/types-of-content-blocks) erfährst du mehr über die verschiedenen Arten von Inhaltsblöcken.

</aside>

**1. Introduction**

bla bla bla

**1.1 Research questions**

Previous research focussed mostly on english articles and the short introductions of wikipedia biographies. I want to take a closer look at specific subsections of wikiepdia texts that seem most relevant for the analysis which are the subsection about the person’s private live and their career. Further, I propose a methodology including matching the data and check different cohorts/timeslots which previous research did not do and thus try to come closer to actual biases in the texts opposing to biases that mirrir real world differences in male and female lives. I aim at answering the following Research Questions:

1. How do Wikipedia biographies of male and female german politicians differ in descriptive attributes and in the language used?
2. Has there been a shift in the gender differences of the biographies of politicians over the last ten years?

In order to answer those questions the thesis will be structured in the following way: The first Chapter will lead the reader through the already existing research of gender bias and wikipedia. the second chapter will provide insights on the data I am using and chapter three will outline the methodology used. Chapter four then guides through the results of my analyses, starting with the descriptive analysis and the whole sample to then narrow it down to the matched corpus of text. The results for different subsets like party, age group, old text etc. will be dicussed in thsi section. This section is followed by a discussion of the results including policy implications. Last but not least I will provide some insights on the limitations of my research and future challenges.

**2 Theoretical framework and literature**

**2.1 Conceptualization of Gender bias**

First, I want to give an overview of gender bias and present what previous research has shown. Gender bias is a very broad term and there are a lot of different meanings to it depending on the context. This paper deals with gender bias in the linguistic context. Beukeboom [reference] defined this as “a systematic asymmetry in word choice as a function of the social category to which the target belongs.” (p.XY) stemming from “essentialist beliefs about social categories”. Whereas Beukeboom denotes that the biases do not have to be harmful he also notes that they potentially lead to discrimination if people are assessed through the lens of stereotypes instead of their individual characteristics.

Glick and Fiske’s [reference] theory of benevolent and hostile sexism showed that sexism in language does not have to be openly hostile, often women are presented with a tone that feels subjectively positive. However, this can result in harmful outcomes, such as the stereotypical portrayal of women in limited roles which can have damaging and disadvantageous consequences for women.

In the context of linguistic gender biases, research has been focussing on the concept of “markedness” describing asymmetries in how male and female individuals are referred to, influenced by gender stereotypes (Beukeboom). This leads back to the idea of the masculine form as the neutral form as it has been uncovered by Simone de Beavoir in her oeuvre The second sex [reference]. Kirgios und Fellbaum point out htis concept by writing: “The maleness of a male candidate is not surprising; rather, it is expected.” [reference *The Lenses of Gender* ]. This is why they theorized that the ocurence of gendered language is higher for female texts - as being female is the deviation of the norm.

In the media, this asymmetric way of presenting females vs males helps to maintain a hegemonic worldview and emphasizes stereotypes concerning women’s place, roles and lives. [reference Gendered Media: Women men and identity politics]. This issue has been discussed more recently after the publication of the Finkbeiner Test, first theorized by the science journalist [Ann Finkbeine](https://en.wikipedia.org/wiki/Ann_Finkbeiner)r and taken up by Aschwanden [reference Aschwanden, Christie (5 March 2013). [The Finkbeiner Test: What matters in stories about women scientists?](https://web.archive.org/web/20170312212947/http:/www.doublexscience.org/the-finkbeiner-test/). *Double X Science*. Archived from the original on 12 March 2017. Retrieved 31 March 2013.] This test raises awareness for stereotypical categories that should not not be used when writing about women in science: To pass the Finkbeiner test, the text must not mention: “The fact that she’s a woman; Her husband’s job; Her child care arrangements; How she nurtures her underlings; How she was taken aback by the competitiveness in her field; How she’s such a role model for other women; How she’s the “first woman to…”. “This scheme (originally developed for women in science) is transferable to the field of politics where I expect the wikipedia articles of women containing information about these categories.

After this general perspective on gender bias, I now want to give have a closer look at gender bias in online settings and the specific characteristics and research for this.

**2.2 Gender Bias in Online Settings/Wikipedia**

Wikipedia offers a great base to study gender biases in text. The specific nature of the text, being crowdsourced offers interesting insights which is why wikipedia text and text of similar crowdsourced nature have been studied before by a number of researches. This thesis focusses on the supply side of wikipedia’s marketplace - which encompassed all the contributions that provide content and maintenance for the site including article creation and editing of the content. Otterbacher et al [reference] studied linguistic bias in Collaboratively Produced Biographies and discovering that the pervasive presence of linguistic biases in social media serves to reinforce societal stereotypes. They examined the use of abstract vs concrete language as a measure of bias in biographies of African American and Caucasian actors at the Internet Movie Database. Otterbacher et al. highlight that even though factual information on Wikipedia is usually corrected quickly the real reason for concern is the nuanced ways of communicating that contribute to the formation or reinforcement of stereotypes. [reference]

Wagner et al [reference] identified and studied different types of bias in wikipedia texts, namely notability, topical focus, linguistic bias, structural properties, and meta-data presentation. They found that firstly, women featured on Wikipedia tend to have higher notability. Secondly, biographies of women are more likely to focus on family, gender, and relationships. Thirdly, there's a linguistic bias where men's biographies use abstract terms for positive traits, whereas women's biographies use them for negative traits. Lastly, differences in metadata and hyperlinks in biographies suggest gendered patterns in how information is connected and sought on Wikipedia. This Thesis will have a closer look into the second of Wagner’s findings and transfer its approach to the specific field of politicians and specific sections of wikipedia biographies, using an analogous approach of measuring word frequency with pointwise mutual information and annotating the results in categories.

First Women, Second Sex: Gender Bias in Wikipedia [referecne] have applied a similar approach, comparing meta-data, language, and network structure. As most other papers they work with english wikipedia data and focus on the introduction text of the biographies. The authors acknowledge that some of the detected differences are mirroring real world circumstances and do not originate from writing biases; however, other differences can be explained with gender bias within Wikipedia's content. This spotlights the difficulty of such an analysis: differentiating where the asymetries are rooted and if they can be assigned to the writing or if they reflect teal world structural differences which exist as well. [almendinger reference]. Previous research has mainly compared all entries for both genders, not narrowing it down to more comparable cases. Brun et al [reference] has used a matching strategy by balancing the data for each occupation proposing that future work should introduce a more elaborate matching strategy including other features and thereby controlling for more confounding variables - which is exactly the goal of my thesis.

The topic of different kind of biases has become more spotlighted in recent decades. Movements like Me too and other attempts to fight inequlaities has raised awareness for such different representation of genders - also on wikipedia. This wikipedia page addresses gender bias and also presents efforts that have been taken to address gender bias: <https://en.wikipedia.org/wiki/Gender_bias_on_Wikipedia>. According to the article, the wikimedia foundation tried to incentive more women to edit pages by dedicating funds to build a more encouraging environment for diversity and by closely monitoring statistics. On the user side there have also been attempts to make wikipedia more diverse for example by organizing edit-a-thons dedicated to women editing wikipedia

In 2022, an article in [*VICE*](https://en.wikipedia.org/wiki/Vice_(magazine)) magazine detailed how British scientist [Jessica Wade](https://en.wikipedia.org/wiki/Jess_Wade) has created more than 1,700 Wikipedia entries on women scientists since 2017, as many women whose contributions have gone unnoticed.[[95]](https://en.wikipedia.org/wiki/Gender_bias_on_Wikipedia#cite_note-95)

**2.3 Gender Bias von Politiker:innen (in online Setting)**

Now I want to highlight the theoretical background of politicians and gender bias (on wikipedia/in online settings). Munzert et al.[reference] have analyzed the political use of wikipedia and concluded that politicians use wikipedia strategically to increase their visibility and increase political power. its specific to their field that they let their pages be edited for them highlighting their political visibility.

Still, stereotypical presentations are present in those texts - Pradl [reference] analyzed that these stereotypes can even be used strategically and purposely to target specific voter groups. She theorized that there is an interplay when it comes to party and use of stereotypes in texts - with right parties relying more on biased traditional images of women as opposed to left parties.

There is an extensive research on the social expectations and stereotypes of female politcians. This stems from the fact that women are still universely underepresented in the field of politics. DAWN LANGAN TEELE [reference] argues that this will not change as long as societal expectations regarding women's family responsibilities conflict with the requirements of a full-time political career, leading to the continued underrepresentation of women in politics.

Monica C. Schneider [reference] researched the stereotypes towards female polticians and found that female politicians cannot win with male-associated qualities and also don't benefit from female-associated qualities.

**2.4 Why is this asymmetry a problem?**

An asymmetry itself is not a necessarily problem but the previous research shows that the linguistic bias concerning women on wikipedia highlights stereotypical aspects and traditional gender depictions. Gendered Media: Women men and identity politics [reference] argues that the media's continuous portrayal domestic of female parliamentarians clearly communicates to the audience that these are the main concerns of our women politicians. Thus such depictions push women in traditional roles and perpetuates this specific view on women in society. The importance of online media to create common believes and set a norm is not to be underestimated. Otterbacher [reference] puts it as follows: “The information shared represents “what is known” or “what we believe to be true” at the collective level, at a given point in time.”. Olga Zagovora [reference] suggests that when it comes to careers and gender, a possible consequence is that especially young individuals might be influenced negatively in their choice of profession after encountering such biased information.

**Data**

I am using the comparative legislator Database [CLD, reference] for my analysis, encompassing political, sociodemographic, career, online presence, public visibility, and imagery data for more than 45,000 modern and past politicians across ten nations. I am working exclusively with the german data. Looking at this specific group using the database is attractive as we there is coverage, all politicians are included in the data and we do not deal with coverage bias or selection effects. The database includes the name of the wikipedia page of each politician which is used as base to extract the politicians wikipedia article using the wikipediR package in R. The data is then cleaned so that pages with redirect or scraping issues are removed. This leaves me with 4259 politicians out of each of the 20 sessions of the Bundestag in my dataset. Other variables like text length /total service or seubsets of the text data (live subsection/career subsection) are created out of the existing/scraped data as a next step. This is the base for the following more extensive analysis of the text in the text content.

The cleaned dataset consists of 4259 politicians with their respective wikipedia section texts and other information about them. Reflecting the above mentioned gender gap in politics, 963 of them are female and 3296 are male.

For the following analysis it is relevant to know that some of the variables are correlating. Figure XY gives an overview of the correlation in the dataset.

**Methodology**

The method used in this thesis builds strongly upon wagner et al [reference] but It adds an important step that makes it unique and goes a step further than previous research by conducting a matching strategy with the data. The next chapter will provide first the results for the complete dataset to then provide the matched version. I use propensity score matching, condensing multiple covariates into a single propensity score for each unit. This first step of my methodology makes the analysis more meaningful as the big challenge of the bias analysis is to figure out if differences can be attributed to the writing process or to the offline structural and historic differences that exist between women and men. So far, literature has mostly left out this aspect but the propensity score matching allows to focus only on comparable cases. For politicians the career level they are at and many other aspects influence the length of texts as well as the amount of available personal information. As the CLD includes a lot of interesting variables to be matched on this is the first step of my analysis. I match on the following variables: Birth year, duration of service, page views, important office. I also apply exact matching on the session meaning that only politicians out of the same session will be matched, as I expect this to be a big confounder, adding a time variance to the data.

As a first part of the analysis, I will present some descriptive figures for the female and male politicians. I will focus on text length, number of links and number of page edits, keeping in mind that they are somewhat correlated (as shown in figure XY). Text length is a common feature to combare by in the literatue about gender bias. Previous research has shown that female texts tend to be longer [reference]. I also put spotlight on the length of the live and the career section as I am particularly interested how the personal information about female/male politicians is distrubuted. The method applied further to the matched data serving to detect gender asymmetries in word use is inspired by (Wagner et al. 2016) and has been used by other authors as well to assess lexical bias (Graells-Garrido, Lalmas, and Menczer 2015) For this method, Pointwise Mutual Information (PMI) is used to find out which words are strongly associated with articles of females/males. Previous research has often focused on the Wikipedia introduction text to detect bias. In my analysis, I apply this method to the live subsection of the texts as well as to the career section. First step of this approach is to tokenize the wikipedia articles and to create the vocabulary, with “gender”, “word” and “Number of biographies” containing this word. Next, stopwords are removed, as well as the names of the politicians and other words that are not removed by the stopword remover but are disturbing the analysis. Further, specific to the german language, there are a lot of gendered ccupational titles in the data which distort the analysis so I manually encoded the top occupation titles to be the same for both genders. Also, I only keep words for my analysis that are present in both genders allowing for a meaningful comparison of their association and ignoring gender specific noise. Then, the PMI scores for the vocabulary are calculated. PMI is expressed as: PMI(c,w) = log p(c,w) p(c)p(w) where C represents the gender and w represents the word. Further, the PMI score needs to be normalized for further analysis. Also, I follow the approach of Wagner et al. in setting a threshold for words to be included in my analysis as PMI gives extra weight to words with very low frequencies - only words that appear in 0,1% of the biographies are considered for the analysis. PMI helps identify words that tend to occur together with a specific gender term more frequently than expected by chance - so, higher PMI values indicate a stronger association between a word and a specific gender. The next step is thus to sort the resulting PMI values decreasingly for each gender. Following the approach implemented by (Wagner et al. 2016), the top 100 words are manually annotated and put in one of these four categories: Family, Relationship, Gender, Other. To assess differences between the genders, Cohens\_w is used to assess if there are significant differences for the categories in the groups of men and women.

**Results**

The are slight differences between the text length of females and males surprisingly with male texts having on average longer live sections and shorter career sections. Overall male articles are a bit longer than females’. Conducting t-tests for the three groups shows that none of these differences are significant.

For the matched dataset, the average length shows significant differences for the career section and for the full articles. In both categories, male articles are longer than female articles. Previous research found no meaningful differences in article length of biographies on wikipedia (JOSEPH REAGLE, 2011) or small differences with female articles being longer (Graells-Garrido, 2015). Having this in mind and looking at the very small effect size of the length differences of politician’s articles,

The mean article length is 5,955 characters for men and 6,013 characters for women (a significant difference according to a t-test for independent samples: p < 0.01, Cohen’s d = 0.01). The mean out-degrees (number of links) of 42.1 for men and 39.4 for women also differ significantly (p < 0.001, Cohen’s d = 0.06). T

For the number of edits of the articles of males nd females, there are no significant differences between the genders, neither in the whole sample nor in the matched data.

discussion

“Editing Wikipedia and NPOV. Critics may rightly say that by relying on secondary sources, Wikipedia just reflects the biases found in them. However, editors are expected to write in their own words, “while substantially retaining the meaning of the source material”9, and thus, the differences found in terms of language that objectify women are chosen explicitly by them. In this aspect, Wikipedia should provide tools that help editors to reduce sexism in language, for instance, by considering already existing manuals like [5]. Furthermore, their neutral point of view guidelines should be updated to explicitly include gender bias, because biased language is a clear violation of their guidelines.” [graells first sex second gender]