Universität Zürich[UZH]

Bachelorarbeit
zur Erlangung des akademischen Grades
Bachelor of Arts
der Philosophischen Fakultät der Universität Zürich

Adapting Gender-Inclusive Rewriting Models to Gender-Neutral German

Verfasserin: Renate Hauser
Matrikel-Nr: 18-101-238

Referent: Prof. Dr. Rico Sennrich

Institut für Computerlinguistik

Abgabedatum: 1. Juni 2023

Abstract

Concurrently with an ongoing societal discourse in German-speaking societies on gender stereotypes that are perpetuated through language use, communication of public institutions as well as industry companies is moving towards more gender-inclusive language. In recent work Amrhein et al. (2023) present a neural rewriter for German as an assistive tool, which reformulates gendered terms with a marked gender-inclusive suffix. However, in the context of corporate or public communication, strategies to gender-inclusive language that are gender-neutral are more desirable. With this motivation in mind, this thesis investigates the applicability of the approach presented in Amrhein et al. (2023) to the case of gender-neutral reformulation. I find that while the approach proves more difficult to be applied to the gender-neutral case and does not generalise as well, the underlying concept of exploiting biased language models for artificial data creation still proves promising and remains to be further explored.

Zusammenfassung

Parallel zu einem aktuellen gesellschaftlichen Diskurs in deutschsprachigen Gesellschaften über Geschlechterstereotypen, die durch die Verwendung von Sprache aufrechterhalten werden, bewegt sich die Kommunikation öffentlicher Institutionen sowie Unternehmen aus der Industrie hin zu einer genderinklusiven Sprache. In jüngerer Arbeit präsentieren Amrhein et al. (2023) einen neuronalen Rewriter für das Deutsche als unterstützendes Tool, das geschlechtsspezifische Begriffe mit einem markierten, gendergerechten Suffix umformuliert. Im Kontext der Unternehmens- oder öffentlichen Kommunikation sind jedoch Strategien zur gendergerechten Sprache, die geschlechtsneutral sind, wünschenswerter. Vor diesem Hintergrund untersucht diese Arbeit die Anwendbarkeit des in der Arbeit von Amrhein et al. (2023) vorgestellten Ansatzes auf die geschlechtsneutrale Umformulierung. Es zeigt sich, dass der Ansatz keine vergleichbaren Ergebnisse für diese Umformulierungsstrategie liefert. Dennoch bleibt das zugrunde liegende Konzept der Ausnutzung von Gender-Bias in Sprachmodellen zur künstlichen Datenerzeugung vielversprechend und bedarf weiterer Untersuchungen.

Acknowledgement

I would like to thank my supervisor Rico Sennrich for his valuable inputs at the beginning of my project.

Also, I would like to express my gratitude to the developers and maintainers of Genderapp, who gave me access to their crowd-sourced terminology.

Further, I would also like to thank my friends Patrick and Simon and my sister Maya for their support.

And finally, I thank my supervisor Chantal Amrhein for her constant support, invaluable inputs and feedback at every step of the process.

Contents

List of Tables

List of Acronyms

GPU	Graphics Processing Unit
NLP	Natural Language Processing
NMT	Neural Machine Translation
URL	Uniform Resource Locator
WER	Word Error Rate

1 Introduction

1.1 Motivation

In German-speaking societies, there has long been a societal and political discourse, which has intensified in the past few years, on how the usage of language perpetuates gender stereotypes(Schach, 2023). The main debate evolves around the usage of the so-called "generic masculine" which means the general practice in German to use the masculine grammatical gender to refer to an individual or a group with unknown gender (Kotthoff, 2020). The main criticism raised by advocates of inclusive language, and the subject of numerous studies about this linguistic practice is, that contrary to what is often claimed, people of other genders than male are not or less included in the cognitive representation of a masculine word (Kotthoff and Nübling, 2018). While the political debate is ongoing, the concept of gender-inclusive language has already found its way into communication guidelines of institutions such as the public administration[1] but also companies from the industry (Schach, 2023). Consequently, there is an increasing demand for gender-inclusive text.

The need for inclusivity extends beyond human-generated text and encompasses the field of Natural Language Processing (NLP). In NLP, large-scale neural models have been found to reproduce or even amplify biases inherent in the extensive amounts of human-written texts on which they are trained (Sheng et al., 2021; Blodgett et al., 2020). These biases can manifest in various ways within downstream systems, including lower performance for underrepresented social groups and the reinforcement of stereotypes, which can perpetuate negative generalizations about specific social groups (Blodgett et al., 2020). Gender bias is a well-researched form of bias, which is prevalent in many NLP models. It can create representational harm by perpetuating gender stereotypes. With the significant advancements made in various NLP tasks in recent years, generative tools have become increasingly utilized in practical applications. With the growing amount of machine-generated texts that humans are exposed to in everyday life, the propagation of biases by these generative tools

[1] `https://www.bk.admin.ch/bk/de/home/dokumentation/sprachen/hilfsmittel-textredaktion/leitfaden-zum-geschlechtergerechten-formulieren.html`

become increasingly problematic.

As discussed in detail in Chapter 2, recent research has presented various approaches to develop rewriting systems capable of transforming gendered input sentences into gender-inclusive output. Rewriters offer an elegant solution for enhancing the gender inclusivity of texts, as they can be applied to any type of input text, regardless of whether it is generated by machines or humans. This flexibility enables their utilization as a post-processing step for mitigating biases in generative language technology, as well as assistive technology for individuals, facilitating the composition of inclusive texts.

One possible approach to rewriting makes extensive use of rule-based linguistic analysis tools to detect and reformulate gendered terms in a gender-inclusive way. For German, such an approach has been proposed by Diesner-Mayer and Seidel (2022). There are also productive systems in the form of online tools like the web application Genderapp[2], which also mainly work in a rule-based manner.

In recent work, Amrhein et al. (2023) have proposed the first neural model for German that does not rely on complex linguistic analysis to tackle the task of gender-inclusive reformulations. They argue that their approach has the advantage over the rule-based approach that it generalizes to other languages easily, as there is no need for sophisticated language-specific expertise or NLP tools and therefore is more suitable for being used in production.

In German, there are several strategies to make a text more inclusive in terms of gender. Some strategies assume gender to be a binary category and make male and female members of a group explicit by either using pair forms that name both or by using the so-called Binnen-I, which appends the female suffix with the first letter of the suffix capitalized. Other strategies append the female suffix but use special characters such as an asterisk (the so-called "gender-star") or a colon to make it explicit that all genders (e.g. including non-binary and agender) are meant. This strategy is often used to give gender-diversity more explicit visibility and is therefore subject of an emotionally led political debate. Yet another strategy is to avoid mentioning gender altogether and to use gender-neutral terms or gender-avoiding sentence structures instead. Examples for these strategies can be found in Table 1.

In their work Amrhein et al. (2023) focus on the production of gender-inclusive language using the strategy of a marked gender-inclusive suffix with a special character. However, in the context of corporate communication of large companies, it is often

[2] `https://genderapp.org/`

Generic Masculine	Grundsätzlich sind die Mitarbeiter der Firma zufrieden.
Pair Form	Grundsätzlich sind die Mitarbeiter und Mitarbeiterinnen der Firma zufrieden.
Binnen-I	Grundsätzlich sind die MitarbeiterInnen der Firma zufrieden.
Gender-Star / Colon	Grundsätzlich sind die Mitarbeiter*innen der Firma zufrieden.
	Grundsätzlich sind die Mitarbeiter:innen der Firma zufrieden.
Gender-Neutral	Grundsätzlich sind die Mitarbeitenden der Frima zufrieden.
Gender-Avoiding	Grundsätzlich herrscht ein zufriedenes Arbeitsklima in der Firma.
English Translation	In principle, the employees of the company are satisfied.

Table 1: Examples for common strategies for gender-inclusive formulations in German.

preferred to use gender-neutral formulations, with which the politically charged debate on the gender-star can be avoided while still moving towards more inclusive communication (Schach, 2023). Also, as stated in the Regulations on Linguistic Equality of the city of Zurich[3], in the context of public administration, there are cases such as legal documents where the gender-star can not be used. However, it is also stated that all texts can be formulated in an inclusive way. As a means for this, gender-neutral formulations are suggested. Finally, in the answers to the free feedback question of the human evaluation that Amrhein et al. (2023) performed and to which I received access, there are also voices from the LGBTQ+ community, that prefer formulations where no gender is mentioned.

1.2 Research Question

With the motivation described in the above section in mind, this thesis attempts to create a neural rewriting model that can reformulate gendered sentences in a gender-neutral way. With this, I try to answer the following research question:

> Can the approach to gender-inclusive rewriting proposed by Amrhein et al. (2023) be applied to the strategy of gender-neutral reformulation with comparable quality?

[3] `https://www.stadt-zuerich.ch/portal/de/index/politik_u_recht/sprache/sprachliche-gleichstellung.html`

1.3 Thesis Structure

To answer this research question, I reproduce the data creation pipeline proposed by Amrhein et al. (2023) with adaptations to account for the use case of gender-neutral reformulations instead of explicit gender-fair reformulation as described in chapter 3. Based on this data, I train several models and evaluate them in terms of their ability to perform reformulations and in terms of the quality of their reformulations. The experimental setup is described in chapter 4. The results of the experiments are presented in chapter 5. In chapter 6, the results of the experiments are discussed and the research question is answered and possible future work is proposed. Chapter 7 wraps up the findings of the thesis. All code used for this thesis can be found on GitHub[4].

[4] `https://github.com/renatehauser/gender-neutral-rewriter`

2 Background

Many Natural Language Processing (NLP) tasks and especially generative tasks have been shown to exhibit societal biases that can be rooted or amplified in every step of the pipeline from data collection to deployment in the creation of an NLP tool (Sheng et al., 2021; Blodgett et al., 2020). Consequently, much work has been dedicated to making such applications less biased, especially with regard to gender biases: Bolukbasi et al. (2016) propose a method for debiasing word embeddings. Escudé Font and Costa-jussà (2019) apply word embedding debiasing techniques to the field of Machine Translation. Other approaches focus on creating more balanced training data, e.g. Lu et al. (2020) propose using counterfactual data augmentation to create pairs of ground truth and a copy with targeted words being replaced with their gendered counterparts. Yet another approach in the field of Neural Machine Translation (NMT) which is similar to an approach for politeness control (Sennrich et al., 2016a), incorporates metadata in the form of a gender tag in the source segment and has proven to produce more accurate translations of gendered input (Vanmassenhove and Hardmeier, 2018; Vanmassenhove et al., 2018).

As an attempt to create a tool, which can both fix biased outputs produced by NLP systems but also assist humans in writing gender-inclusive texts, rewriters have been proposed that are able to reformulate any input text in a gender-inclusive way.

A rule-based approach can be employed in order to detect terms that need reformulation in the source and alter the segment to make the term gender-inclusive and the surrounding context grammatically correct. Diesner-Mayer and Seidel (2022) present a rule-based system that detects human referents in the generic masculine and proposes reformulations either with the pair form or with the gender star. Their approach makes heavy use of morphological and dependency parsing, co-reference resolution and a word inflexion database.

Other works argue, that directly applying a rule-based system in production is expensive, because of the high computation costs of sophisticated linguistic tools (Vanmassenhove et al., 2021; Sun et al., 2021). Both propose to use a rule-based rewriting approach to synthetically create parallel data and train a neural rewriting

model for English. On the one hand, this has the advantage of faster and easier inference (Vanmassenhove et al., 2021). On the other hand, it is argued that neural models have the ability to generalize to unseen examples (Sun et al., 2021). In consequence, this approach can make such systems more scalable.

Similar approaches are applied to morphologically more complex languages that have grammatical gender. Jain et al. (2021) train a neural model to create gender variants for gendered input for Spanish. While during training it is known, which samples are re-genderable and which are not, this is unknown during test time. Similar to Habash et al. (2019), who propose a system to identify and reinflect gendered first person singular in Arabic, Jain et al. (2021) propose a gender classifier to label sentences as "re-genderable" or "neutral" and append the resulting tag to the source. However, their approach struggles to identify neutral segments correctly. Alhafni et al. (2020) build upon the work of Habash et al. (2019) in creating a gender-reinflection model for Arabic by successfully incorporating linguistic features in their neural model. However, their model is only able to reinflect gendered first person. Alhafni et al. (2022) extend the previous work to a model that is able to rewrite sentences with first and second persons with independent grammatical gender preferences involved.

The aforementioned systems use a forward augmentation approach to produce synthetic data: they debias biased text using rules and word lists and thus create artificial target segments for training neutral rewriters. Especially for English, as in the work of Vanmassenhove et al. (2021) and Sun et al. (2021), this is relatively easy to do, as it is a morphologically simple language and does not exhibit grammatical gender. English only expresses social gender on pronouns and on a closed group of words for professions. Therefore, identifying gendered human referents and also creating rules for debiasing is relatively straightforward.

Neural sequence-to-sequence models are more sensitive to target-side noise in the training data than to noise on the source-side. Following this argumentation, Sennrich et al. (2016b) reverted the augmentation direction in work on data augmentation for machine translation. This backward augmentation was shown to be more beneficial for the training of MT models than forward augmentation (Khayrallah and Koehn, 2018; Bogoychev and Sennrich, 2020).

Amrhein et al. (2023) adopt this argumentation. They propose to change the augmentation direction as compared to previous work on gender rewriting and instead of debiasing biased text, they suggest artificially biasing gender-inclusive text, in order to obtain target-original training data. They also show, that in German, identifying gender-inclusive forms instead of gendered forms can be achieved with relatively simple regular expressions and therefore mitigates the problem of differen-

tiating between human referents and general nouns. This makes intricate linguistic rules unnecessary, even for morphologically complex languages like German. Amrhein et al. (2023) successfully reproduce Sun et al. (2021)'s forward augmentation model for English with a backward augmentation approach and report no loss in quality.

However, while the identification of human referents can easily be achieved in German when reverting the augmentation direction, expert language-specific knowledge is still needed to define rules for biasing the original unbiased text. To make the production of synthetic training data not reliant on such sophisticated linguistic knowledge or tools, Amrhein et al. (2023) propose to leverage the social bias that machine translation models exhibit and create round-trip translations - meaning translating to a pivot language and back to the original source language using an NMT system - from gender-inclusive texts in order to artificially bias it. With this, they make use of a technique that has already proven useful in other NLP tasks such as Grammatical Error Correction (Lichtarge et al., 2019), Paraphrasing (Mallinson et al., 2017), Summarization (Fabbri et al., 2021) and Automatic Post-Edition (Junczys-Dowmunt and Grundkiewicz, 2016; Freitag et al., 2019), which leverage grammatical and fluency errors stemming from machine translation.

While there have been various successful attempts to gender-fair rewriting for several languages that either focus on generating gender-variants or language-specific gender-inclusive patterns, there is no work that focuses on debiasing rewriters that make text inclusive by avoiding gendered formulations altogether to the best of my knowledge.

3 Data

3.1 Data Creation

3.1.1 Real Data Extraction

In their methodology, Amrhein et al. (2023) employed a backward augmentation approach wherein they artificially generated biased segments from real unbiased data. This approach has the advantages that firstly, gender-fair forms in the German language can be easily identified based on their patterns, while determining gendered terms would require an animacy test to ascertain if a noun refers to a gendered human being or a regular noun. Secondly, the authors demonstrated that training the model with target-original synthetic data reduces the likelihood of learning "false" patterns.

To construct their dataset, the researchers utilized OSCAR (Abadji et al., 2022), a large multilingual web corpus, and filtered for gender-fair segments using regular expressions designed to identify common German gender-fair patterns. Subsequently, they subjected the obtained unbiased segments to synthetic biasing. This was achieved by round-trip translation of the segments using biased machine translation models. To prevent the model from acquiring patterns arising solely from the translation behavior of the round-trip translation models, the researchers merged the round-trip translated segment with the original segment. Consequently, the only alteration from the source to the target segment was the gender-fair reformulation.

In this study, I replicate the filtering approach proposed in the previous work by Amrhein et al. (2023) using a terminology-based approach instead of a pattern-based approach. The pattern-based approach is not suitable for the gender-neutral case due to the varied and intricate nature of the patterns required to identify gender-neutral forms and to create gendered corresponding forms for them. To address this limitation, spaCy's[1] PhraseMatcher[2] was utilized to extract gendered segments

[1] `https://spacy.io/`

[2] `https://spacy.io/api/phrasematcher`

and gender-neutral segments from the OSCAR dataset, based on a terminology further described in section 3.1.1.1. The objective was to obtain segments that were either consistently gender-neutral or entirely gendered throughout the entire segment. Consequently, segments where both gendered and gender-neutral terms were identified were discarded.

3.1.1.1 Terminology

For this work, I used a terminology with entries that map gendered to gender-neutral terms. It is a curated crowd-sourced resource collected mainly via a web application called Genderapp[3]. I received access to an export of the terminology by the developers and maintainers of the respective website.

Each entry within the terminology maps a gendered term in both masculine and feminine forms and in singular and plural number to a corresponding gender-neutral term in singular and plural. Not all fields within an entry are required to be filled. Moreover, many-to-many relationships exist within the terminology, meaning that a single gendered term can have multiple corresponding neutral terms, and vice versa. A total of 2345 term entries gendered to gender-neutral mappings are contained in the terminology.

3.1.1.2 Term Matching Paradigm

Performing term lookup in a language with limited morphology is relatively straightforward (Vanmassenhove et al., 2021; Sun et al., 2021). However, this task poses a challenge when dealing with morphologically complex languages like German. One potential approach is to employ orthographic matching, as this method is sufficient in morphologically simple languages such as English. Nonetheless, this approach may overlook segments that contain inflected terms, thereby presenting a potential drawback. To address this issue, another possibility is to match terms based on their lemmas, thereby accounting for variations in inflexion. However, lemmatization not only results in the loss of case information but also number information. This is problematic because certain terms exhibit gender neutrality in their plural form but not in their singular form (e.g., "die Studierenden" (PL) vs. "der Studierende" (SG)). Thus, I hypothesized that using a lemmatized matching paradigm would yield a higher number of segments matched as "neutral," while in reality, they contain gendered terms.

[3] `https://genderapp.org/`

	Gendered	Neutral	Inanimate
Overlap	7.84%	45.10%	47.06%
Lemmatized Matching	5.88%	39.22%	54.90%
Orthographic Matching	19.61%	51.96%	29.41%

Table 2: Results of Manual Evaluation

To determine the most suitable matching paradigm for the objectives of this study, I conducted a preliminary investigation. From a subset of the OSCAR corpus, I extracted gendered and gender-neutral segments using both orthographic matching and lemmatized matching. Counting the extracted segments reveals that lemmatized matching resulted in approximately 19% more extracted segments for gender-neutral terms and an even greater increase of 36% for gendered terms compared to orthographic matching. However, there is a large overlap between the two paradigms, with only about 9% of the segments uniquely matched through orthographic matching and 24% uniquely matched through lemmatized matching.

	Gendered	Neutral	Inanimate
Lemmatized Matching	7.37%	43.69%	48.94%
Orthographic Matching	8.93%	45.69%	45.38%

Table 3: Results Extrapolated to Total Number of Matches

To evaluate the hypotheses described in the preceding paragraph, I performed a manual analysis of the unique neutral segments identified by orthographic matching, lemmatized matching, and the overlap between the two paradigms. A total of 103 segments were annotated to determine if they were gendered, neutral, or inanimate. For a segment to be classified as "neutral," the entire segment needed to be neutral, while a segment was labelled as "inanimate" if it lacked any animate referent. Such segments are extracted due to the ambiguities of terms in the terminology, which refer to humans in certain contexts, but not in others. An example for this can be seen in Table 4. The results are presented in Table 2. Subsequently, I compared the proportions of gendered, neutral, and inanimate segments in the overlap of the two paradigms and the proportions in the uniquely matched segments of each paradigm.

Contrary to the initial hypothesis, the analysis reveals that lemmatized matching does not result in a higher number of segments being falsely labelled as "neutral."

Neutral Match	*Die eingebaute Entkalkungsanzeige sorgt zudem dafür, dass die Kaffeepadmaschine genau die richtige **Pflege** bekommt, um die Lebenszeit so lang wie möglich zu halten.*
Gendered Match	*Webshop: Sie können Ihren **Anhänger** in unserem Webshop selber konfigurieren.*

Table 4: Examples for ambiguous terms that can be inanimate in certain contexts.

In fact, the opposite is the case: substantially more gendered segments are found as "neutral" by orthographic matching. Examining the data provides an explanation for this observation. As previously mentioned, segments containing both gendered and neutral terms were excluded from consideration. However, orthographic matching occasionally fails to identify gendered terms due to their inflexion. Consequently, these segments are not filtered out despite not being entirely neutral or gendered. These findings support the use of lemmatized matching, as it yields a substantially greater number of matched segments and mitigates the issue of gendered segments being falsely labelled as "neutral."

However, lemmatized matching does exhibit a higher tendency to match terms that are actually inanimate instead of gender-neutral such as in the Example in Table 4. Approximately 55% of the matched segments were classified as inanimate, compared to 47% in the overlapping segments. Nevertheless, from Table 3 we see, that when extrapolating this finding from the limited sample to the segment counts of the OSCAR subset utilized in this investigation, it is evident that this difference has a negligible negative impact on the overall proportion of inanimate matches in a larger dataset and can therefore be disregarded.

Based on the findings of the preliminary study described above, I decided to adopt lemmatized matching for the filtering step, as it is expected to yield higher-quality training data. However, it is important to note that these results are specific to the German language and may not hold true for other languages. Furthermore, it should be emphasized that lemmatization relies on advanced language-specific resources, including part-of-speech taggers and dependency parsers such as those provided by spaCy. Utilizing orthographic matching instead of lemmatized matching would increase the language independence of the approach, as it does not depend on such sophisticated language-specific resources.

3.1.2 Forward Augmentation

In the gender-fair case, Amrhein et al. (2023) argue in favor of using backward augmentation by artificially biasing gender-fair texts, because firstly, it is easier to identify animate gendered entities using a pattern-based data extraction method. Secondly, using target original data has been demonstrated to be more effective, as it prevents the model from learning false patterns. To investigate whether this argumentation holds true for the gender-neutral case when employing term-matching-based data extraction, a manual analysis was conducted on 105 randomly selected segments from the extracted gendered and gender-neutral segments. The results indicate that the first argument does not apply to the gender-neutral case. In fact, the opposite is observed: about 58% of the gender-neutral matches are actually inanimate, while this is only the case for about 10% of the gendered matches. This finding can be attributed to the fact that many strategies used to make sentences gender-neutral involve terms that are typically not used for referring to human beings (and therefore are not gendered). Instead, these terms are understood as animate only in the given context, as was seen earlier in the example in Table 4. However, with the terminology-based data extraction approach employed in this study, it is not possible to distinguish between these two cases.

To assess the data quality that could be expected from a forward augmentation approach, wherein gendered terms are simplistically replaced with gender-neutral terms from the terminology, a detailed manual evaluation was conducted as outlined in section 4.2.2.2. The results, as presented in Table 9, indicate that only approximately 18% of the modified segments are grammatically correct in terms of gender and number agreement.

I additionally evaluated the performance of Genderapp translate[4], an online tool which employs more sophisticated term insertion techniques involving rule-based and statistical methods to accurately inflect the inserted terms. This approach demonstrates great improvement, with approximately 56% of the resulting segments being grammatically correct. However, even with this enhanced approach, it is still not a viable method for creating training data for a neural model with a forward augmentation paradigm, as such a model would not be able to effectively learn the task with almost half of the target segments in the training data being ungrammatical.

Despite the problem of inanimate term matches, I align with the approach proposed by Amrhein et al. (2023) and opt to use a backward augmentation approach for training data creation.

[4] `https://genderapp.org/translate`

3.1.3 Backward Augmentation

In the subsequent section, I provide a detailed description of three data creation methodologies that were employed in my experiments with the aim of reproducing the data creation pipeline employed by Amrhein et al. (2023) and achieving a similar level of quality as their resulting gender-fair model.

3.1.3.1 Term Replacement

The first approach utilizes backward augmentation, wherein gender-neutral terms are simplistically replaced with gendered equivalents from the terminology. To create a balanced dataset in terms of gender, a random subsample of neutral segments was selected from the extracted segments of the OSCAR corpus and the neutral terms within the chosen segments are substituted once with exclusively masculine terms and once with exclusively feminine terms.

3.1.3.2 Round Trip Translation

As described in section 3.1.2, simplistic replacements lead to many segments with ungrammatical and/or unnatural or meaningless contexts. This is even more the case when replacing neutral terms with gendered terms as compared to the other direction because many matched neutral terms are actually inanimate nouns and do not reference human beings. An example of this can be seen in Table 5. I hypothesize that round-trip translation of the replaced segments will on the one hand make the synthetic source sentence more grammatical without the need for complex language-specific rules and statistics. On the other hand, I expect that this creates a more natural context for the inserted term, as translation models are trained to produce not only grammatical but also fluent output.

Gendered Replacement	*Handgefertigte Lampen aus Trompeter und anderen gebrauchten Instrumenten.*
Neutral Replacement	*Webshop: Sie können Ihren Unterstützende in unserem Webshop selber konfigurieren.*

Table 5: Examples of how simplistic replacements can lead to ungrammatical or meaningless sentences in both replacement directions.

This step aligns with the approach taken by Amrhein et al. (2023), who also employed round-trip translations to produce their synthetic source segments. In their

study, English was used as a pivot language since it is a non-gendered language. This allowed that gender bias would be introduced when translating back to the gendered German language. Following this rationale, the same Facebook WMT 2019[5] translation models Ng et al. (2019) used by Amrhein et al. (2023) for round-trip translations were leveraged in my work. In their setup, they fine-tuned the English-to-German translation model by appending gender tags to enforce the desired gender in the output during inference. This allowed for the balancing of training data for different genders. They reported that while the tags did not guarantee the correct gender in all cases, only 36% of the segments with feminine tags were produced with masculine terms, compared to 90% when using the model without fine-tuning.

I adopted the same approach and created round-trip translations using the same fine-tuned model, which I obtained through the authors of the paper. This approach proved directly applicable to my data setup, facilitating the production of round-trip translations while maintaining gender-specificity.

3.1.3.3 Term Based Filtering

In the data created using round-trip translations, I observed that many of the round-tripped segments no longer contained the inserted term. This occurred frequently when the inserted term did not match the context and was therefore improbable to be produced by the translation model utilized for the round-trip translation. Amrhein et al. (2023) controlled their training data to only contain changes in the gendered terms by merging the round-tripped segments into the original segment. However, this merging approach is not feasible in the gender-neutral terminology-based approach, as the reformulations often require more complex adjustments to the sentence structure compared to the gender-fair approach. Therefore, it is explicitly desired to allow for reformulations that go beyond a simple term replacement.

To address this challenge, I employed a softened version of the merging approach. I filtered the round-trip translated source segments based on the presence of a gendered term in the terminology. This filtering process aims to reduce parallel segments where ungendered source segments are unnecessarily reformulated. Additionally, this heuristic served as a means to control the similarity of meaning between the source and target segments. It is hypothesized that this filtering step would aid the model in learning reformulations that better retain the meaning of the source segment while still allowing flexibility in how a sentence is reformulated.

[5] `https://huggingface.co/facebook/wmt19-de-en`, `https://huggingface.co/facebook/wmt19-en-de`

4 Methods

4.1 Rewriting Models

4.1.1 Model Architecture and Hyperparameters

I followed the training scheme of Amrhein et al. (2023) to reproduce their approach as closely as possible. The models are trained with a transformer architecture (Vaswani et al., 2017) with 6 encoding and 6 decoding layers, 4 attention heads, a hidden layer size of 512 and a feed-forward layer size of 1024. I use a batch size of 10'000 tokens.

I used the Sockeye 3 toolkit (Hieber et al., 2022) for all the trainings with a joint byte-pair vocabulary (Sennrich et al., 2016c) computed with SentencePiece (Kudo and Richardson, 2018) and follow Amrhein et al. (2023) and use Adam for optimization (Kingma and Ba, 2015) with standard hyperparameters and the standard Transformer learning schedule in Vaswani et al. (2017) with a linear warmup over 4'000 steps. However, for faster training, I use a learning rate of 0.004 instead of the 0.0005 used by Amrhein et al. (2023). I used early stopping to end the training after validation perplexity has not improved for 8 checkpoints. I trained all models on NVIDIA A100 PCIe GPU.

I trained four models with different training data described more closely in the next section.

4.1.2 Training Data

For the first model, I used the simplistic backward augmentation approach described in section 3.1.3.1 (Backward Replacement model). To match the model of Amrhein et al. (2023) in terms of training data size, I randomly subsampled 12M neutral segments from the extracted segments of the OSCAR corpus. I also added the same neutral segments that I augmented as copy segments so that the model learns not to change already neutral segments. Also following the data setup of Amrhein et al.

(2023) I added 15M segments without gendered or gender-neutral terms amounting to 30% of the total training data resulting in a total of 51M segments.

From the parallel segments used for the Baseline model, I produced round-trip translations as described in section 3.1.3.2 and trained an additional model on the resulting data (Round Trip Translation model). For a third model, I performed term-based filtering as described in section 3.1.3.3 as I had observed, that many of the round-trip translated segments do not contain the inserted term anymore (Term Based Filtering Model).

Filtering the training data with the term-based filtering approach reduces the number of gendered training samples from which the model can learn to do reformulations to about 18M segments as compared to 24M, resulting in a total training data size of about 42M segments. To check, that potential changes in the performance of the model trained with the filtered data are not only an effect of the training data size, I trained a last model with the same training data size as the Term Based Filtering model by randomly subsampling the gendered training data of the Round Trip Translation model (Subsampled model).

To remove noisy parallel segments, I filtered all training data with OpusFilter (Aulamo et al., 2020) with the following filtering steps:

- LengthFilter: unit=character, min=15, max=250
- LongWordFilter: threshold=40
- AlphabetRatioFilter: threshold=0.5
- LanguageIDFilter fasttext: threshold=0.0

4.1.3 Validation Data

I did not have any genuine gendered/gender-neutral parallel data from which I could build a validation set for the training. However, I wanted to have qualitative good parallel segments for validation, as I did early stopping in the training based on the validation set. The validation set should meet the following criteria:

1. It should have the same distribution of gendered, gender-neutral and ungendered segments as the training data.
2. The target segment should be a valid reformulation of the source in terms of grammaticality and meaning.

3. There should not be unnecessary reformulations.
4. Gender-neutral terms in the gender-neutral copy segments should be genuinely gender-neutral and not denote non-human referents.
5. The target segments of gendered and gender-neutral source segments should be completely gender-neutral.

To ensure that the listed criteria are met by my validation set, I created round-trip translations of backward replacements from gender-neutral segments subsampled from the data extraction from OSCAR and did term-based filtering for masculine and feminine terms. I then manually checked and corrected the feminine and masculine data and the gender-neutral data to produce 150 parallel segments of each category. The resulting validation set was complemented with 200 ungendered copy segments that were subsampled from the ungendered segments from the data extraction from OSCAR.

4.2 Evaluation Setup

I use several evaluations to compare the models of this work to the Bias-to-Debias model presented in Amrhein et al. (2023) with regard to their ability to generalize from the training data, the extent to which the models detect and reformulate gendered terms, the extent to which the models perform unnecessary changes and finally the quality of the performed reformulations. In section 4.2.1 I describe the metrics I use for automatic evaluation. In section 4.2.2, I describe the evaluation setups that I used to assess the models.

4.2.1 Automatic Evaluation Metrics

Related work has used edit-distance-based metrics such as Word Error Rate (WER) or the originally for Grammatical Error Correction proposed *MaxMatch* (Dahlmeier and Ng, 2012) on parallel test sets for gender-inclusive rewriting to assess the performance of their systems (Amrhein et al., 2023; Vanmassenhove et al., 2021; Sun et al., 2021; Alhafni et al., 2020). However, as this work is the first to investigate gender-neutral reformulations for German, there is no suitable parallel test set, making edit-distance-based metrics unsuitable. I therefore only use WER to assess unnecessary reformulations (section 4.2.1.3) and propose accuracy metrics more closely described in the following sections 4.2.1.1 and 4.2.1.2 to be used for quantitative evaluation.

4.2.1.1 Terminology Match Accuracy

In order to evaluate the ability of the models to reformulate a gendered term in a gender-neutral way, I propose an adapted version of the Exact Match Accuracy presented by Alam et al. (2021). The Exact Match Accuracy is dependent on a reference. As we do not have any original parallel data, we make use of the terminology we have instead and search the target segment for corresponding gender-neutral terms from the terminology. The formula can be seen in 4.1.

$$accuracy(h, T) = \frac{\#\text{ target matches}}{\#\text{ source terms}} \tag{4.1}$$

Where h is the hypothesis and T is the terminology. I perform lemmatized matching with spaCy to find the terms. For each gendered term in the source, I search if there is a corresponding term from the terminology in the target. A gendered term can also have multiple neutral correspondences. A span in the source can also be matched by multiple terms from the terminology (for example with a different number). Only one correspondence of a source match for the same span has to match to count as a correct match. If a span in the target has already been matched, it is ignored in the search for target matches of later source matches. This ensures, that the same span in the target is not counted as a correct replacement for multiple source terms. This approach is based on the assumption, that the source terms and their replacements in the target are in the same order. After a manual investigation of the data, this is a reasonable assumption.

4.2.1.2 Gender-Fair Match Accuracy

For gender-fair reformulations, the terms in the source were matched in the same way as for the Terminology Match Accuracy, so that the reformulation ability of the gender-fair model would be evaluated on the same terms as the gender-neutral models. Instead of matching terminology correspondences in the target, for each token with a "@@GFM@@in" or "@@GFM@@innen" suffix in the target, it is checked whether the lemma is also in the source matches. Gender-fair reformulations that do not correspond to a source match are neglected. The Gender-Fair Match Accuracy is described in formula 4.2, where h is the hypothesis, T the terminology for source term matching and S is the set of suffixes that are considered for target matching.

$$accuracy(h, T, S) = \frac{\#\text{ target matches}}{\#\text{ source terms}} \tag{4.2}$$

4.2.1.3 Word Error Rate

To check whether the models make unwanted changes to already neutral or ungendered segments, I compute tokenized[1] Word Error Rate (WER) on a copy test set more closely described in section 4.2.2.4. For the computation, I use the jiwer Python package[2].

4.2.1.4 Limitations

The above-described metrics are limited by several factors. Firstly, the metrics are reliant on the terminology used for term matching and are therefore only able to test the performance of the systems on the predefined set of terms. As there is a potentially infinite number of gendered terms that could exist in German, this is of course not an exhaustive list and the evaluated contexts therefore inherently incomplete.

Secondly, the metrics are limited by the performance of spaCy's phrase matching. Terms can be missed on both the source and the target side for the Terminology Match Accuracy. Also, false positives in the source segment can lead to "false misses" in the target. The metrics are additionally reliant on the lemmatization capabilities of spaCy, as we compare the lemmas of the target matches against the lemmas of the source matches. If one of the two is falsely lemmatized, this also leads to "false misses".

Furthermore, the metrics are only able to measure if and to which extent a model is able to do reformulations, however, the quality of the reformulation in terms of the grammatical agreement as well as meaning preservation and context fit are not reflected by the metric. Also, reformulations that go beyond changes of single tokens are ignored or even penalized by the metrics.

Lastly, the comparability of the two accuracy metrics is limited, as, even though they are evaluated on the same source terms, the number of target matches is not counted in the same way.

Despite these limitations, the metrics can indicate if the performances of models in changing specific terms differ substantially. Also, they are useful to evaluate large test set sizes as described in section 4.2.2.3 and therefore can complement more fine-grained insights from a manual evaluation, where the sample size is very

[1] `https://github.com/alvations/sacremoses`

[2] `https://github.com/jitsi/jiwer`

limited, as described in section 4.2.2.2. Furthermore, in a more controlled setup of a template test set such as the frequency distribution test set described in section 4.2.2.1, where there is always exactly one source term which is known beforehand, the above-mentioned limitations are less relevant and the direct comparability of the two metrics is not necessary to reveal patterns of model performances.

4.2.2 Evaluations

4.2.2.1 Term Frequency Distribution

In order to investigate, how well the models can learn to generalize from their respective training data, I evaluated the models on their ability to reformulate terms that they have seen a differing number of times during training. I chose to evaluate this in a controlled setup to avoid unpredictable effects for example from the context of the gendered terms to be reformulated.

Gender/Number	Template
Feminine Singular	Bislang hat aber noch keine {term} Geld zurückzahlen müssen.
	Wie sieht denn Ihr Idealbild von einer {term} aus?
Masculine Singular	Bislang hat aber noch kein {term} Geld zurückzahlen müssen.
	Wie sieht denn Ihr Idealbild von einem {term} aus?
Plural	Ich denke, es gibt viele {term}, die das noch nicht begriffen haben.
	Das sehen auch die {term} so.

Table 6: Templates for Term Frequency Distribution Evaluation

I, therefore, created a template test set by picking and adapting sentences from real data with non-domain-specific contexts. For each template, I created a version for singular and plural and for the singular also one each for masculine and feminine gender. This resulted in a total of 60 templates, 20 for each grammatical context. Examples can be seen in Table 6. The complete list of templates can be found in Table 14 in the appendix.

For each model, I counted the number of occurrences of the gendered terms in the terminology in the respective training data and created buckets of terms according to the order of magnitude of their occurrences in the respective training data. I created six buckets with their ranges defined as follows:

$$y = 0, \quad 1 \leq y \leq 10, \quad 10^n + 1 \leq y \leq 10^{n+1} \quad \text{where } n \in$$

$\{1, 2, 3, 4\}$ and y is the number of occurrences of a term

To ensure a fair comparison against the Bias-to-Debias model, I only considered terms that were in the same frequency bucket both for a gender-neutral and for the gender-fair model. I created separate buckets for masculine and feminine terms in order to be able to separately evaluate the performances of the models on the two genders. To produce the data sets, I filled the templates with the terms from the buckets for each bucket, resulting in 480 and maximally 3640 samples per bucket test set.

Finally, I computed the Terminology Match Accuracy or Gender-Fair Match Accuracy respectively on the rewritings produced by each model for each frequency bucket test set and both genders.

4.2.2.2 Manual Evaluation

As explained above in section 4.2.1.4, the automatic metrics are not informative of the quality of the made reformulations. To assess the performance of the models in terms of grammaticality and meaning preservation of the reformulations, I perform a manual evaluation.

To create the test set, I randomly subsampled 300 real gendered segments from the data I extracted from OSCAR and filtered out noisy segments with OpusFilter (Aulamo et al., 2020) with the filters described below. For ease of evaluation, I set a restrictive length filter that only allows segments of at most 250 characters.

- LengthFilter: unit=character, min=15, max=250
- LongWordFilter: threshold=40
- AlphabetRatioFilter: threshold=0.5
- LanguageIDFilter fasttext: threshold=0.0

I assess the Bias-to-Debias model and from the models of this work the Forward Augmentation Baseline, the Backward Replacement Model, the Round-Trip Translation model and the Term Based Filtering model. Additionally, I include genderapp in order to compare the neural rewriting approach against a sophisticated rule-based and statistic rewriter. As I do not expect quality gains from the Subsampled Round-Trip Translation model, which I trained to rule out data size effects of the filtering, I do not include this model in the manual evaluation. I annotated each of the 300 test segments with the binary categories described in Table 7. It has to be noted, that

Category	Description and Gender-Neutral and Gender-Fair Examples
Changed	Is the target different than the source?
	Der Lehrer *ist begeistert /* ***Die Lehrperson*** *ist begeistert*
	Der Lehrer *ist begeistert /* ***Der*die Lehrer*in*** *ist begeitstert*
Neutral / Fair	Does the target contain gender-neutral rewriting of gendered terms (or gender-fair respectively)? This is also True, if already neutral/fair terms were correctly left unchanged. This only takes the terms into account but not articles/pronouns/adjectives.
	Die ***Leserinnen*** *und die* ***Studenten*** *sind begeistert / Die* ***Leserschaft*** *und die Studenten sind begeistert*
	Die ***Leserinnen*** *und die* ***Studenten*** *sind begeistert / Die* ***Leser*innen*** *und die Studenten sind begeistert*
Completely Neutral / Fair	Are all gendered terms gender-neutral in the target? This is also True if the source is already neutral and the target still is.
	Die ***Leserinnen*** *und die* ***Studenten*** *sind begeistert / Die* ***Leserschaft*** *und die* ***Studierenden*** *sind begeistert*
	Die ***Leserinnen*** *und die* ***Studenten*** *sind begeistert / Die* ***Leser*innen*** *und die* ***Student*innen*** *sind begeistert*
Correct Grammatical Gender Agreement	Do articles, pronouns, adjectives and nouns agree in terms of grammatical gender? This is also True, if no pronouns/adjectives don't need to be changed for correct agreement.
	Der *Lehrer ist begeistert /* ***Die*** *Lehrperson ist begeistert*
	Der *Lehrer ist begeistert /* ***Der*die*** *Lehrer*in ist begeitstert*
Correct Number Agreement	Do articles, pronouns, adjectives and nouns agree in terms of number? This is also True, if no pronouns/adjectives don't need to be changed for correct agreement.
	Die Leserinnen ***sind*** *begeistert / Die Leserschaft* ***ist*** *begeistert*
	Die Leserinnen ***sind*** *begeistert / Die Leser*innen* ***sind*** *begeistert*
Meaning Preserved	Is the meaning except for gender information preserved? If the output is nonsensical, this is False.
	Die Leserinnen sind begeistert / Die Leserschaft ist begeistert
	*Die Leserinnen sind begeistert / Die Leser*innen sind begeistert*

Table 7: Change Categories

even though the segments were extracted by subsampling segments with matches of gendered terms from the terminology, not all of the resulting segments are genuinely gendered in such that they require reformulation. This is due to the limitations of the terminology-based extraction also described in chapter 3. To account for this, I additionally annotated the segments, if they already are neutral (i.e. do not need reformulation).

To assess the quality of the changes made, I computed the percentage of neutral and completely neutral changes relative to the total number of changes that a system performed from the resulting annotations. Furthermore, I computed the percentage of meaning-preserving changes, and grammatical changes relative to all changes performed by a system. Additionally, I computed the percentage of changes that were completely correct, meaning that the segment was completely neutral, meaning-

preserving and grammatical. Lastly, I assessed the models in terms of Precision and Recall, where true positives were changes of gendered segments, false positives changes of already neutral segments, false negatives unchanged gendered segments and true negatives unchanged already neutral segments.

4.2.2.3 Large Scale Automatic Evaluation

With only 300 test segments, the manual evaluation is very limited in size. I therefore additionally create a large-scale test set that contains 10,000 real gendered segments that I subsampled from the extracted segments from OSCAR filtered with OpusFilter (Aulamo et al., 2020) with the following filtering steps:

- LengthFilter: unit=word, min=1, max=150
- LongWordFilter: threshold=40
- AlphabetRatioFilter: threshold=0.5
- LanguageIDFilter fasttext: threshold=0.0

To complement the findings of the manual evaluation, I compute the Terminology Match Accuracy and the Gender-Fair Match Accuracy respectively on the rewritings of the models to evaluate the extent to which the models change gendered source segments. I exclude the Baseline and genderapp from this evaluation as they directly apply the terminology that the metrics rely on.

From the accuracies on each of the frequency buckets, I will analyze if the frequency with which a model saw a term a during training has an effect on the performance on such a term.

4.2.2.4 Copy Evaluation

The trained models should not only reformulate gendered terms, they are also required to leave already neutral or ungendered segments unchanged. To assess whether the models make unwanted changes, I created a copy test set that contains 1,000 gender-neutral and 1,000 ungendered segments that I subsampled from the extracted segments from OSCAR filtered with OpusFilter with the same filtering steps as for the large-scale test set.

I computed WER (with lower results being better) on the rewritings of the models to detect the number of unwanted changes.

5 Results

5.1 Frequency Distribution Evaluation

The results of the frequency distribution evaluation can be seen in plots XY. We see that the gender-fair rewriting model is able to perform reformulations on gendered terms that it has rarely or never seen during training, although performance is better for terms it has seen frequently (i.e. more than 1000 times). Also, it can be observed that the performance is higher for feminine than for masculine terms. The gender-neutral models, in contrast, hardly do replacements for zero-shot and very-low frequency terms. While for the feminine terms the models learned to do some reformulations already after heaving seen a term more than 10 times, masculine terms have to be seen more than 1000 times during training that the models do reformulations.

Comparing the gender-neutral models, it can be observed that the Backward Replacement model does the most reformulations for almost all frequency buckets where reformulations are made. The Round-Trip Translation model and the Subsampled model show the lowest performance, while being almost equal. The Term Based Filtering can make up for parts of the performance loss of the Round-Trip Translation model on the Terminology Match Accuracy but stays below the Backward Replacement except for the 11-100 frequency bucket of the feminine terms.

5.2 Manual Evaluation

Table 8 shows the percentage of changed segments for each of the models under test. We see that especially the Baseline but also Genderapp have high percentage of changes. This can be accounted to the fact, that the segments of the manual evaluation were obtained by term-based extraction using the same terminology as the Baseline and Genderapp. It can be observed, that both round-trip translation and term based filtering increase the number of changed segments. Also, it is noteworthy, that the Bias-to-Debias also changes less than half of the segments.

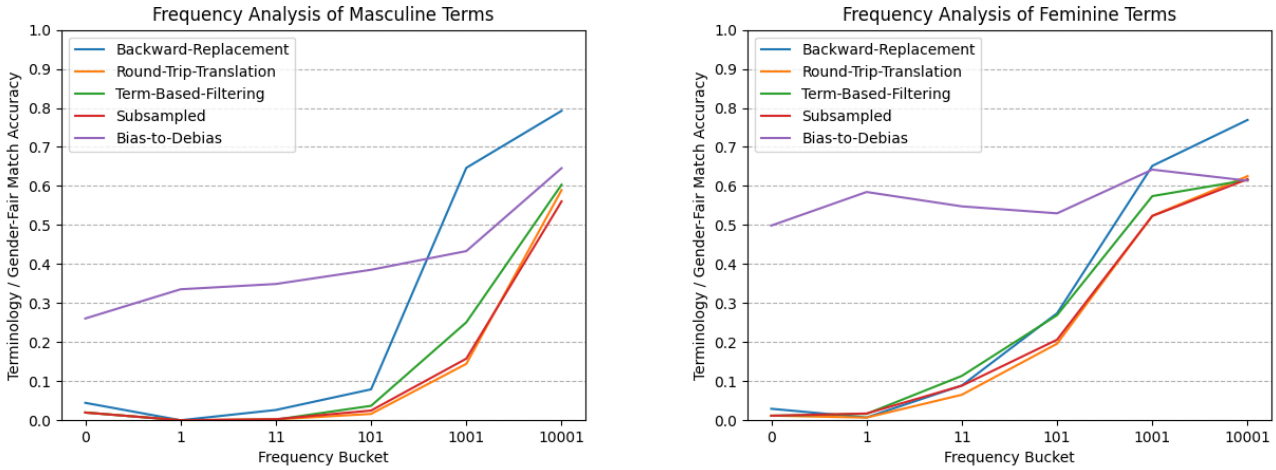

Figure 1: Terminology Match Accuracy and Gender-Fair Match Accuracy respectively of neural rewriting models for segments with feminine and masculine terms of different frequency ranges.

Method	
Bias-to-Debias	48.00%
Backward Replacement	37.33%
Round-Trip Translation	42.33%
Term Based Filtering	52.67%
Baseline	98.33%
Genderapp	78.33%

Table 8: Percentage of changed segments relative to the total number of segments

The overall results of the manual evaluation can be seen in table 9. We see that the gender-neutral models can not match the performance of the gender-fair Bias-to-Debias model in almost all regards. Proportionally more of the changes that the gender-fair model performs lead to completely or partially gender-fair segments that are grammatically correct and preserve the original meaning. The only regard in which the best gender-neutral model is on par with the Bias-to-Debias model is the number of changes it performs. However, the gender-neutral rewriters are more error-prone in what they change: They reformulate already neutral segments more often and at the same time miss more of the gendered segments that would require reformulation.

It can also be observed, that while the simplistic replacements performed for the

Method	Completely Correct	Neutral	Partially Neutral	Meaning preserving	Grammatical
Bias-to-Debias	47.92%	76.39%	23.61%	96.53%	70.14%
Backward Replacement	15.18%	75.89%	22.32%	51.79%	50.00%
Round-Trip Translation	28.35%	74.80%	11.02%	62.20%	55.91%
Term Based Filtering	34.18%	68.99%	8.86%	75.32%	60.76%
Baseline	10.17%	74.92%	16.95%	61.36%	17.97%
Genderapp	42.18%	82.55%	11.91%	82.55%	55.74%

Table 9: Percentage of evaluated change types relative to the total number of changes

Baseline perform very poorly, the sophisticated rule-based and statistical replacement approach of Genderapp sets a strong baseline which in most regards outperforms the neural gender-neutral rewriters presented in this work. Especially noteworthy is the good performance of Genderapp in changing actually gendered segments but leaving already neutral or ungendered segments unchanged, as compared to all the neural approaches — including the Bias-to-Debias model —, which only reformulate conservatively, missing 30% needed changes or more.

Furthermore, it can be observed, that while the gender-neutral models are not able to match the performance of the gender-fair Bias-to-Debias model, each additional step in the data creation pipeline brings improvements in the resulting model both in terms of grammaticality and meaning preservation. Notably, the model resulting from the last data preparation step (the term-based filtering) even surpasses Genderapp in the proportion of performed reformulations that are grammatical.

Method	Precision	Recall
Bias-to-Debias	94.44%	59.65%
Backward Replacement	92.92%	46.05%
Round-Trip Translation	85.04%	47.37%
Term Based Filtering	77.22%	53.51%
Baseline	76.95%	99.56%
Genderapp	93.19%	97.33%

Table 10: Precision and Recall

5.3 Large Scale Evaluation

Model	Terminology Match Accuracy	Gender-Fair Match Accuracy
Bias-to-Debias	-	0.31
Backward Replacement	0.30	-
Round-Trip Translation	0.18	-
Term Based Filtering	0.21	-
Subsampled	0.17	-

Table 11: Terminology Match Accuracy and the Gender-Fair Match Accuracy respectively

Table 11 shows the Terminology Match Accuracy and the Gender-Fair Match Accuracy respectively for the gender-neutral models presented in this work and the Bias-to-Debias model presented by Amrhein et al. (2023) on a large-scale test set containing 10,000 real gendered segments. While we can not directly compare the result on the Gender-Fair Match Accuracy to the results on the Terminology Match Accuracy of the gender-neutral models as described in section 4.2.1.4, we still see, that the Bias-to-Debias model does reformulations to a comparable extent as the best performing gender-neutral model. The results suggest that round-trip translation of the data created using backward replacement results in the model producing fewer reformulations. The Term Based Filtering model can make up for part of the performance loss on the metric, however, does not reach the performance of the Backward Augmentation model. However, it has to be kept in mind, that this does not reflect the quality of the changed sentences in terms of grammaticality and meaning preservation.

5.4 Copy Evaluation

The results for the copy evaluation can be seen in table 12. We see that the gender-fair Bias-to-Debias model has significantly higher WER, meaning that it more often reformulates segments that actually should be left unchanged. No significant difference between the very low WER values for the gender-neutral rewriting models presented in this work can be found.

Model	WER
Bias-to-Debias	2.27
Backward Replacement	**0.2**
Round-Trip Translation	**0.21**
Term Based Filtering	**0.19**
Subsampled	**0.2**

Table 12: WER of the neural models on the copy test set. Numbers in bold mean, that none other is statistically more significant.

6 Discussion

With the four models that were trained for this work, an investigation was conducted to determine the applicability of the data creation pipeline proposed by Amrhein et al. (2023) for gender-neutral reformulation, as opposed to gender-fair reformulation. The results of the manual evaluation reveal that round-trip translation of artificial source segments after the insertion of gendered terms leads to improvements in quality. Specifically, the Round-Trip Translation model outperforms the Backward Replacement model in terms of grammaticality and preservation of meaning. Adopting a soft version of the merging approach used by Amrhein et al. (2023) and filtering the round-trip translated segments for the presence of gendered terms further helps to make the learning more targeted. This additional step in data preparation improved the production of fully neutral segments while maintaining grammatical correctness and preserving meaning. However, both round-trip translations and filtering resulted in an increase in the total number of changes made by the models as compared to simple backward replacement, leading to reduced reliability: As we see in table 10, both steps consistently reduce the precision with which changes are made. Notably, however, the ability to leave ungendered or already neutral segments unchanged is the only regard, in which the gender-neutral models are able to outperform the gender-fair Bias-to-Debias model, as indicated by the significantly lower Word Error Rate on the copy test set for the gender-neutral models. Nevertheless, reliability remains a general drawback of neural models, including the Bias-to-Debias model, when compared to the rule-based and statistical approach employed by Genderapp, which rarely misses gendered segments but also leaves already neutral segments unchanged.

It has to be pointed out, that during the annotation of the manual evaluation it was observed, that the neural gender-neutral models hardly do reformulations apart from replacements of gendered terms such as reformulating complete sentences or phrases. I hypothesize, that the machine translation models used for the round-trip translation stick close to the source when translating, therefore, the artificial source segments resulting from the round-trip translations are similar to the original target segments in terms of sentence structure. Possibly, training data where the sentence

structure is more varied between source and target segment could be beneficial to encourage the model to do more flexible reformulations.

The findings from the large-scale evaluation using the accuracy metrics proposed in this work indicate that the gender-fair Bias-to-Debias model tends to perform a greater number of reformulations, with the Backward Replacement model being the only one comparable in the extent of replacements made. Notably, the disparity between the gender-neutral models is striking: the Round-Trip Translation model exhibits an accuracy more than 10 percentage points lower than that of the Backward Replacement model. This outcome is particularly striking given that the results of the manual evaluation do not suggest such substantial differences among the models. While it is true that the Backward Replacement model generates more "completely neutral" segments, potentially resulting in a higher overall count of reformulated gendered terms compared to the other neutral models, it is unlikely that this factor alone can account for such a considerable discrepancy. Considering the limitations of the metric discussed in Section 4.2.1.4, it remains questionable how accurately the proposed accuracy metrics can reflect the actual performance of gender-neutral rewriting models when used on uncontrolled real test data. Consequently, a more thorough investigation into the capabilities of the metric should be conducted in the future.

Based on the aforementioned results, it can be concluded that the data creation pipeline proposed by Amrhein et al. (2023), involving backward augmentation, round-trip translation, and merging, can - in an adapted form - successfully be applied for gender-neutral reformulation after real source data extraction from a large web corpus with a terminology-based approach instead of a pattern-based approach. Each step of the pipeline brings the expected quality improvements. However, the resulting models still perform substantially worse than the gender-fair model presented by Amrhein et al. (2023). Furthermore, in most aspects, these models fail to surpass the robust rule-based and statistical baseline established by Genderapp. Nevertheless, it should be noted that the most effective neural gender-neutral model does manage to outperform Genderapp in terms of grammaticality. This finding highlights the potential of neural models to generate grammatically correct sentences in a morphologically complex language like German without relying on intricate rules.

The Frequency Distribution analysis makes clear, where the gender-neutral models presented in this work fail in comparison to the gender-fair model of Amrhein et al. (2023). While the Bias-to-Debias model learned to generalize well with the synthetic training data and has stable performance even for zero-shot terms it has never

seen during training, the gender-neutral models only start to consistently change gendered terms to gender-neutral formulations after they have seen a term 1000 times or more. This can be well explained by the fact, that gender-fair reformulations follow clear and simple patterns that mainly operate on the surface level that can easily be generalized. We also see from the Figures in 1, that the performance of the gender-fair model is lower on masculine than on feminine terms, which shows that the task is more difficult if there are no surface-level cues to decide whether a token is animate and therefore needs to be reformulated. The patterns to reformulate in a gender-neutral way are much more complex and often do not operate on the surface level but require the model to have learned complex semantic connections. From the fact, that just as for the gender-fair model, the gender-neutral models' performance on feminine terms was learned easier, we see even more strongly, that the models have mainly learned to focus on surface-level patterns.

6.1 Future Work

The gender-neutral models lag behind the gender-fair model in several regards. On the one hand, they are too conservative in changing source segments and consequently leave too many gendered segments unchanged. Also, grammaticality and meaning preservation need to be further improved to reach a level that can be useful in production. Furthermore, it was observed that the gender-neutral models only perform replacements of gendered terms, while more holistic phrase- or even sentence-level reformulations would be desirable.

One possibility that could be explored is to further improve the filtering of the synthetic parallel data to improve grammaticality and meaning preservation. For this, language models could be used to compute the perplexities of artificial source and original target segments. Then, the parallel segments could be filtered by their perplexity difference. We saw that the replacement of gender-neutral terms with gendered terms can result in unnatural artificial source sentences, as often the gender-neutral match is actually inanimate. We would expect that having a gendered term in an unnatural context (i.e. in a context that would not require an animate referent) would result in a higher perplexity as compared to the original target segment and therefore the pair could be discarded because of a high perplexity difference. Additionally, this approach would filter out segment pairs where the original target is not a fluent sentence and the round-trip translation produced is (more) grammatical. From filtering out such segments, we would expect gains in grammaticality since having segments with noise on the target side can result in reduced output

quality (Sennrich et al., 2016b).

Furthermore, other data creation methods could be explored, to encourage the learning of phrase- or sentence-level reformulations that go beyond the mere replacement of gendered terms. While the concept of leveraging the gender bias inherent to neural models is promising, the outputs generated by machine translation models are possibly not varied enough for the use case of gender-neutral rewriting. Employing different model types such as paraphrasing models that potentially create more varied output could be employed for artificial data creation. Additionally, the recent improvements in large language models open interesting avenues to data creation by language model prompting, as can be seen in an example with ChatGPT in Table 13. Anecdotal examples such as Example 2 in Table 13 suggest, that also with large language models, a backward augmentation approach might be preferable, as gender-neutral formulations tend to be more often grammatically wrong.

Prompt	***Generiere einen Satz in einer gegenderten Version und danach denselben Satz in gender-neutraler Version:***
Example 1	*Gegendert: Der Lehrer korrigiert die Tests der Schüler.*
	Gender-neutral: Die Lehrperson korrigiert die Tests der Lernenden.
Example 2	*Gegenderte Version: Der Arzt behandelte den Patienten und verschrieb ihm Medikamente, während die Krankenschwester ihm bei der Genesung half.*
	Gender-neutrale Version: Die medizinische Fachkraft behandelte die Person und verschrieb ihnen Medikamente, während eine Pflegekraft ihnen bei der Genesung half.

Table 13: Example of how a large language model such as ChatGPT could be prompted to create artificial training data.

Finally, when pushing forward in developing such systems, we have to understand better, what the needs and preferences of potential users of such a system are. For example, as for now, we hardly know whether potential users prefer a rewriter to produce a reformulation for every gendered formulation in the source even though the result might be ungrammatical, or if they prefer to have a grammatically and meaning-wise correct output that might in turn not be completely gender-neutral. Also, it should be explored how gender-fair and gender-neutral reformulations could be combined to complement each other and if this is an approach that is desired by potential users. Lastly, the topic of inclusive language is an ongoing debate in the German-speaking society and broadly accepted strategies are only just emerging. This will have to be kept in mind and requirements for such a system will frequently need to be reconsidered.

7 Conclusion

As the current public discourse suggests that forms of gender-inclusive language which avoid gendered terms altogether are favored over marked gender-inclusive suffixes, a rewriter with the ability to reformulate text gender-neutrally is desirable. Especially in the context of corporate language, it is a smooth way to avoid the politically charged debate while still moving towards more inclusive communication. With this motivation in mind, I have investigated whether the approach of Amrhein et al. (2023) to gender-inclusive rewriting, which reformulates gendered terms with inclusive special characters, can be applied to a different German gender-inclusive reformulation strategy that circumvents the use of gendered terms.

I replicated the data creation pipeline proposed by Amrhein et al. (2023) while adapting it to the gender-neutral case. Instead of utilizing pattern-based extraction, I employed a crowd-sourced German terminology to extract segments containing gender-neutral terms from a large web corpus. Through this approach, I aimed to investigate whether the findings reported by Amrhein et al. (2023) regarding their data augmentation steps would hold true in the gender-neutral context. The results confirmed that the data augmentation steps employed by Amrhein et al. (2023) also had a positive impact on the quality of the artificial training data in the gender-neutral case. Specifically, backward augmentation, round-trip translation, and term-based filtering (serving as a softened version of their merging approach) all contributed to improving the quality of the artificial training data, thereby enhancing the output quality of the resulting models. However, the overall performance of the gender-neutral models trained in this thesis remains substantially lower than the performance of the Bias-to-Debias model presented by Amrhein et al. (2023) that focuses on marked gender-inclusive suffixes rather than gender-neutral terms as in this thesis. An analysis of the performance on gendered terms with different frequencies in the training data of the models reveals that while the Bias-to-Debias model performs well even for terms it has rarely or never seen during training, no generalization was learned by the gender-neutral models, which are only able to reformulate terms they have seen 1000 times or more with a certain consistency. This can be explained by the fact, that the patterns to produce marked gender-fair

suffixes mainly operate on the surface level, while complex semantic relations have to be learned to produce gender-neutral reformulations.

Furthermore, it is noteworthy that although this was not the main research question, it could be observed that the neural approach to gender-neutral rewriting stayed far below the performance of the rule-based system of Genderapp. However, the best neural gender-neutral model was still able to outperform the heavily engineered rule-based approach in terms of grammaticality, highlighting the potential that the neural approach nevertheless has. Moreover, while the models trained with the data creation approach followed in this thesis did not prove to produce phrase- or even sentence-level reformulation to avoid gendered terms, neural models are in principle still more flexible to perform reformulations that go beyond the replacement and reinflection of a static terminology than what a rule-based system is capable of. To encourage such behavior in a neural rewriter, different approaches to synthetic data creation such as prompting large language models that lead to more holistic reformulations could be explored in the future.

In conclusion, the data creation approach that succeeded in gender-inclusive rewriting using marked gender-inclusive suffixes did not achieve comparable results in gender-neutral rewriting. However, the fundamental concept of utilizing the gender bias of neural language models remains promising for data creation. This opens up new avenues for exploration of data augmentation techniques to enhance the effectiveness of neural gender-neutral rewriting models.

References

J. Abadji, P. Ortiz Suarez, L. Romary, and B. Sagot. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.463.

M. M. i. Alam, A. Anastasopoulos, L. Besacier, J. Cross, M. Gallé, P. Koehn, and V. Nikoulina. On the Evaluation of Machine Translation for Terminology Consistency, June 2021. URL http://arxiv.org/abs/2106.11891. arXiv:2106.11891 [cs].

B. Alhafni, N. Habash, and H. Bouamor. Gender-Aware Reinflection using Linguistically Enhanced Neural Models. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 139–150, Barcelona, Spain (Online), Dec. 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.gebnlp-1.12.

B. Alhafni, N. Habash, and H. Bouamor. User-Centric Gender Rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 618–631, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.46. URL https://aclanthology.org/2022.naacl-main.46.

C. Amrhein, F. Schottmann, R. Sennrich, and S. Läubli. Exploiting biased models to de-bias text: A gender-fair rewriting model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada, 2023. Association for Computational Linguistics.

M. Aulamo, S. Virpioja, and J. Tiedemann. OpusFilter: A Configurable Parallel Corpus Filtering Toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156, Online, July 2020. Association for Computational Linguistics. doi:

10.18653/v1/2020.acl-demos.20. URL https://aclanthology.org/2020.acl-demos.20.

S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485.

N. Bogoychev and R. Sennrich. Domain, Translationese and Noise in Synthetic Data for Neural Machine Translation, Oct. 2020. URL http://arxiv.org/abs/1911.03362. arXiv:1911.03362 [cs, stat].

T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of Thirtieth Conference on Neural Information Processing Systems (NIPS)*, pages 4349–4357, Barcelona, Spain, 2016.

D. Dahlmeier and H. T. Ng. Better Evaluation for Grammatical Error Correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://aclanthology.org/N12-1067.

T. Diesner-Mayer and N. Seidel. Supporting Gender-Neutral Writing in German. In *Proceedings of Mensch und Computer 2022*, MuC '22, pages 509–512, New York, NY, USA, Sept. 2022. Association for Computing Machinery. ISBN 978-1-4503-9690-5. doi: 10.1145/3543758.3547566. URL https://dl.acm.org/doi/10.1145/3543758.3547566.

J. Escudé Font and M. R. Costa-jussà. Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3821. URL https://aclanthology.org/W19-3821.

A. Fabbri, S. Han, H. Li, H. Li, M. Ghazvininejad, S. Joty, D. Radev, and Y. Mehdad. Improving Zero and Few-Shot Abstractive Summarization with Intermediate Fine-tuning and Data Augmentation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 704–717, Online, June 2021.

Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.57. URL https://aclanthology.org/2021.naacl-main.57.

M. Freitag, I. Caswell, and S. Roy. APE at Scale and Its Implications on MT Evaluation Biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5204. URL https://aclanthology.org/W19-5204.

N. Habash, H. Bouamor, and C. Chung. Automatic Gender Identification and Reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3822. URL https://aclanthology.org/W19-3822.

F. Hieber, M. Denkowski, T. Domhan, B. D. Barros, C. D. Ye, X. Niu, C. Hoang, K. Tran, B. Hsu, M. Nadejde, S. Lakew, P. Mathur, A. Currey, and M. Federico. Sockeye 3: Fast Neural Machine Translation with PyTorch, Aug. 2022. URL http://arxiv.org/abs/2207.05851. arXiv:2207.05851 [cs].

N. Jain, M. Popović, D. Groves, and E. Vanmassenhove. Generating Gender Augmented Data for NLP. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 93–102, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.gebnlp-1.11. URL https://aclanthology.org/2021.gebnlp-1.11.

M. Junczys-Dowmunt and R. Grundkiewicz. Log-linear Combinations of Monolingual and Bilingual Neural Machine Translation Models for Automatic Post-Editing. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 751–758, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2378. URL https://aclanthology.org/W16-2378.

H. Khayrallah and P. Koehn. On the Impact of Various Types of Noise on Neural Machine Translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2709. URL https://aclanthology.org/W18-2709.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning*

Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL http://arxiv.org/abs/1412.6980.

H. Kotthoff. Gender-Sternchen, Binnen-I oder generisches Maskulinum, ... (Akademische) Textstile der Personenreferenz als Registrierungen? *Linguistik Online*, 103(3):105–127, Oct. 2020. ISSN 1615-3014. doi: 10.13092/lo.103.7181. URL https://bop.unibe.ch/linguistik-online/article/view/7181. Number: 3.

H. Kotthoff and D. Nübling. *Genderlinguistik Eine Einführung in Sprache, Gespräch und Geschlecht*. Narr Francke Attempto Verlag GmbH + Co. KG, 2018. ISBN 9783823301523. URL https://elibrary.narr.digital/book/99.125005/9783823379133.

T. Kudo and J. Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012.

J. Lichtarge, C. Alberti, S. Kumar, N. Shazeer, N. Parmar, and S. Tong. Corpora Generation for Grammatical Error Correction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3291–3301, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1333. URL https://aclanthology.org/N19-1333.

K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta. Gender Bias in Neural Natural Language Processing. In V. Nigam, T. Ban Kirigin, C. Talcott, J. Guttman, S. Kuznetsov, B. Thau Loo, and M. Okada, editors, *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, Lecture Notes in Computer Science, pages 189–202. Springer International Publishing, Cham, 2020. ISBN 978-3-030-62077-6. doi: 10.1007/978-3-030-62077-6_14. URL https://doi.org/10.1007/978-3-030-62077-6_14.

J. Mallinson, R. Sennrich, and M. Lapata. Paraphrasing Revisited with Neural Machine Translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long*

Papers, pages 881–893, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1083.

N. Ng, K. Yee, A. Baevski, M. Ott, M. Auli, and S. Edunov. Facebook FAIR's WMT19 News Translation Task Submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5333. URL https://aclanthology.org/W19-5333.

A. Schach. Gendergerechte Sprache. In A. Schach, editor, *Diversity & Inclusion in Strategie und Kommunikation: Vielfalt in Konzeption, Kultur und Sprache im Unternehmen*, pages 265–306. Springer Fachmedien, Wiesbaden, 2023. ISBN 978-3-658-40153-5. doi: 10.1007/978-3-658-40153-5_6. URL https://doi.org/10.1007/978-3-658-40153-5_6.

R. Sennrich, B. Haddow, and A. Birch. Controlling Politeness in Neural Machine Translation via Side Constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California, June 2016a. Association for Computational Linguistics. doi: 10.18653/v1/N16-1005. URL https://aclanthology.org/N16-1005.

R. Sennrich, B. Haddow, and A. Birch. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, Aug. 2016b. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009.

R. Sennrich, B. Haddow, and A. Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, Aug. 2016c. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162.

E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng. Societal Biases in Language Generation: Progress and Challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online, Aug. 2021. Association for Computational Linguistics.

doi: 10.18653/v1/2021.acl-long.330. URL https://aclanthology.org/2021.acl-long.330.

T. Sun, K. Webster, A. Shah, W. Y. Wang, and M. Johnson. They, Them, Theirs: Rewriting with Gender-Neutral English, Feb. 2021. URL http://arxiv.org/abs/2102.06788. arXiv:2102.06788 [cs].

E. Vanmassenhove and C. Hardmeier. Europarl Datasets with Demographic Speaker Information. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, page 391, Alicante, Spain, May 2018. URL https://aclanthology.org/2018.eamt-main.59.

E. Vanmassenhove, C. Hardmeier, and A. Way. Getting Gender Right in Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1334. URL https://aclanthology.org/D18-1334.

E. Vanmassenhove, C. Emmery, and D. Shterionov. NeuTral Rewriter: A Rule-Based and Neural Approach to Automatic Rewriting into Gender Neutral Alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.704. URL https://aclanthology.org/2021.emnlp-main.704.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

A Tables

Gender/Number	Template
F SG	Man muß keine {term} sein, um das langgezogene Schauerstück mit Unruhe zu verfolgen.
	Die {term} muß nur unterschreiben, um für den Erhalt ihres Schnäppchenparadieses zu stimmen.
	Bislang hat aber noch keine {term} Geld zurückzahlen müssen.
	Ernster jedenfalls als die Euphorie von einer {term}, die nur die Sonnenseite der Bonanza wahrnehmen will.
	Eine {term}, die zur Polizei kommt, hat ein Problem, ist unsicher, manchmal hilflos.
	Eine verschwundene {term} - kein Einzelfall.
	Wie sieht denn Ihr Idealbild von einer {term} aus?
	Das Vorgehen sei das ”seit Jahren mieseste und fieseste Manöver von einer {term}”.
	Eine Aufforderung, die schon so manche {term} erbleichen ließ.
	Und dies sogar in dem Rechtssystem, in dem diese ”{term}” schon so lange zu Hause ist?
	Die Ausbildung zur {term} in Deutschland soll in den nächsten zehn Jahren radikal umgestellt und vor allem kürzer werden.
	Natürlich ist die Kür von einer {term} keine einfache Sache.
	es sind Bilder des Grauens, die die {term} vor sich sieht.
	Keine einzige {term} beantwortete das Schreiben fristgerecht.
	Sie fragen danach - nach wie vor -, ob man überhaupt mit einer {term} reden dürfe.
	Wenn zum Beispiel eine {term} auf dem Gehsteig hinsegelt, sich verletzt und nachweislich das schlüpfrige Laub dafür verantwortlich ist, darf sie Schadenersatz fordern.
	Seuchen scheinen der {term} ein Gespenst aus dunkler Vergangenheit.
	Und sogar für einen ausgesprochenen Problemfall fand sich eine {term}:
	”Das Problem ist, den Gestus der {term} loszuwerden.”
	Man kann als {term} gegen Pollock nicht gewinnen.

Table 14: Female Singular Templates for Term Frequency Distribution Evaluation

Gender/Number	Template
M SG	Man muß kein {term} sein, um das langgezogene Schauerstück mit Unruhe zu verfolgen.
	Der {term} muß nur unterschreiben, um für den Erhalt seines Schnäppchenparadieses zu stimmen.
	Bislang hat aber noch kein {term} Geld zurückzahlen müssen.
	Ernster jedenfalls als die Euphorie eines {term}, der nur die Sonnenseite der Bonanza wahrnehmen will.
	Ein {term}, die zur Polizei kommt, hat ein Problem, ist unsicher, manchmal hilflos.
	Ein verschwundener {term} - kein Einzelfall.
	Wie sieht denn Ihr Idealbild von einem {term} aus?
	Das Vorgehen sei das ”seit Jahren mieseste und fieseste Manöver von einem {term}”.
	Eine Aufforderung, die schon so manchen {term} erbleichen ließ.
	Und dies sogar in dem Rechtssystem, in dem dieser ”{term}” schon so lange zu Hause ist?
	Die Ausbildung zum {term} in Deutschland soll in den nächsten zehn Jahren radikal umgestellt und vor allem kürzer werden.
	Natürlich ist die Kür von einem {term} keine einfache Sache.
	es sind Bilder des Grauens, die der {term} vor sich sieht.
	Kein einziger {term} beantwortete das Schreiben fristgerecht.
	Sie fragen danach - nach wie vor -, ob man überhaupt mit einem {term} reden dürfe.
	Wenn zum Beispiel ein {term} auf dem Gehsteig hinsegelt, sich verletzt und nachweislich das schlüpfrige Laub dafür verantwortlich ist, darf er Schadenersatz fordern.
	Seuchen scheinen dem {term} ein Gespenst aus dunkler Vergangenheit.
	Und sogar für einen ausgesprochenen Problemfall fand sich ein {term}:
	”Das Problem ist, den Gestus von dem {term} loszuwerden.”
	Man kann als {term} gegen Pollock nicht gewinnen.

Table 15: Male Singular Templates for Term Frequency Distribution Evaluation

Gender/Number	Template
PL	Die auf die Probe gestellte Wahrnehmung der {term} ist wüstenklar und doch leicht verschleiert.
	Hinzu kommt ein eindeutiger Trend, der besagt, daß die Leute auf die {term} setzen:
	Mit Reis kochen {term} ihr eigenes Süppchen
	Die {term} glauben nicht so recht, daß sich Unternehmen durch die BUND-Tips zu Öko-Firmen wandeln.
	Ich denke, es gibt viele {term}, die das noch nicht begriffen haben.
	{term} der Provinz sitzen der Regierung in Paris im Nacken
	Die {term} sollen von dieser Entwicklung profitieren.
	Das schockiert vielleicht die {term}, weil es so gewalttätig ist, aber für mich ist es vor allem ein Bild des Bösen und der Versuchung.
	Hierzulande schufen die {term} 1148 neue Jobs.
	Es gebe auch noch mehrere {term}, sagte Burns.
	Am aktuellen Rand, wie die {term} sagen, sieht das Bild indessen wesentlich weniger rosig aus.
	In den französischen Alpen sind zwei US-amerikanische {term} in einen Schneesturm geraten und nach Behördenangaben erfroren.
	Europäische Kommission regt {term} auf
	Doch die {term} erschienen nicht zum vereinbarten Treffen.
	Dort blieben am Wochenende zahllose Schnellzüge stecken, von denen nicht wenige gerade als Entlastungszüge für die von der Straße vertriebenen {term} eingesetzt worden waren.
	Aber die {term} lachen nur:
	Mit Gewinnspielen machen immer mehr Organisationen nicht das Glück der {term}, sondern ihr eigenes.
	Das alles aber hatte den {term} nicht genügt.
	Das sehen auch die {term} so.
	Mannheimer {term} hatten in den letzten Jahren wiederholt für negative Schlagzeilen gesorgt.

Table 16: Plural Templates for Term Frequency Distribution Evaluation