Exploiting Biased Models to De-bias Text: A Gender-Fair Rewriting Model

Chantal Amrhein[1]* **Florian Schottmann**[2,3] **Rico Sennrich**[1,4] **Samuel Läubli**[1,2]

[1]University of Zurich, [2]Textshuttle, [3]ETH Zurich, [4]University of Edinburgh

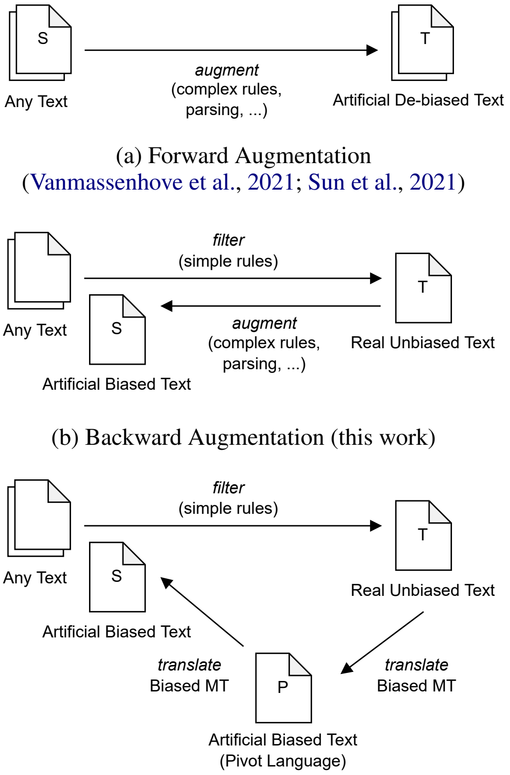
{amrhein,sennrich}@cl.uzh.ch, {schottmann,laeubli}@textshuttle.ai

Abstract

Natural language generation models reproduce and often amplify the biases present in their training data. Previous research explored using sequence-to-sequence rewriting models to transform biased model outputs (or original texts) into more gender-fair language by creating pseudo training data through linguistic rules. However, this approach is not practical for languages with more complex morphology than English. We hypothesise that creating training data in the reverse direction, i.e. starting from gender-fair text, is easier for morphologically complex languages and show that it matches the performance of state-of-the-art rewriting models for English. To eliminate the rule-based nature of data creation, we instead propose using machine translation models to create gender-biased text from real gender-fair text via round-trip translation. Our approach allows us to train a rewriting model for German without the need for elaborate handcrafted rules. The outputs of this model increased gender-fairness as shown in a human evaluation study.[1]

1 Introduction

From facial recognition to job matching and medical diagnosis systems, numerous real-world applications suffer from machine learning models that are discriminative towards minority groups based on characteristics such as race, gender, or sexual orientation (Mehrabi et al., 2021). In natural language processing (NLP), gender bias is a particularly significant issue (Sheng et al., 2021b). While research in psychology, linguistics and education studies demonstrates that inclusive language can increase the visibility of women (Horvath et al., 2016; Tibblin et al., 2022) and encourage young individuals of all genders to pursue stereotypically gendered occupations (Vervecken et al., 2013, 2015) without sacrificing comprehensibility (Friedrich and Heise, 2019), state-of-the-art text generation models still overproduce masculine forms and perpetuate gender stereotypes (Stanovsky et al., 2019; Nadeem et al., 2021; Renduchintala and Williams, 2022).



(a) Forward Augmentation (Vanmassenhove et al., 2021; Sun et al., 2021)

(b) Backward Augmentation (this work)

(c) Round-trip Augmentation (this work)

Figure 1: De-biasing rewriters can be implemented as neural sequence-to-sequence models trained on source (**S**) to target (**T**) text examples. Previous work creates artificial **T** from real **S** through complex augmentation (a). We propose to use real **T** and generate artificial **S** to accommodate morphologically complex languages and avoid target-side noise (b). Furthermore, we show that by leveraging biased off-the-shelf machine translation (MT) models, complex rules can be avoided altogether to generate training data for de-biasing rewriters (c).

There is a variety of work on correcting gender bias in generative models. Some approaches

*Work done during an internship at Textshuttle.

[1] We publicly release our data and code here: https://github.com/textshuttle/exploiting-bias-to-debias

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4486–4506
July 9-14, 2023 ©2023 Association for Computational Linguistics

include curating balanced training data (Saunders et al., 2020), de-biasing models with modifications to the training algorithms (Choubey et al., 2021), and developing better inference procedures (Saunders et al., 2022). The focus of our work lies on so-called rewriting models, yet another line of de-biasing research that revolves around models that map any input text (e.g., the output of a biased generative model) to a gender-fair version of the same text. The main challenge here is training data: due to the lack of large amounts of parallel biased-unbiased text segments, previous work (Section 2) produces the latter through handcrafted rules (Forward Augmentation, Figure 1a).

We identify two key problems with the Forward Augmentation paradigm. First, the rule-based de-biasing of real-world text comes at a risk of introducing target-side noise, which tends to degrade output quality more than source-side noise (Khayrallah and Koehn, 2018; Bogoychev and Sennrich, 2019). Second, while already intricate for English, it is likely even harder to define de-biasing rules for more morphologically complex languages with grammatical gender (Figure 2). An approach proposed by Diesner-Mayer and Seidel (2022), for example, requires morphological, dependency and co-reference analysis, as well as named entity recognition and a word inflexion database.

In this paper, we propose two modifications to the prevalent data augmentation paradigm and we find that biased models can be used to train de-biasing rewriters in a simple yet effective way. Our three main contributions are:

- We reverse the data augmentation direction (Backward Augmentation, Figure 1b). By using human-written unbiased segments filtered from large monolingual corpora as target-side data, we train a neural rewriting model that matches or outperforms the word error rate (WER) of two strong Forward Augmentation baselines in English (Section 3.3).
- We dispose of handcrafted de-biasing rules (Round-trip Augmentation, Figure 1c). By leveraging biased off-the-shelf NLP models, we train a neural rewriting model that outperforms the WER of a heavily engineered rule-based system in German (Section 4.3).
- We test our best model with potential stakeholders. In our human evaluation campaign, participants rated the outputs of our German rewriter model as more gender-fair than the original (biased) input texts (Section 4.4).

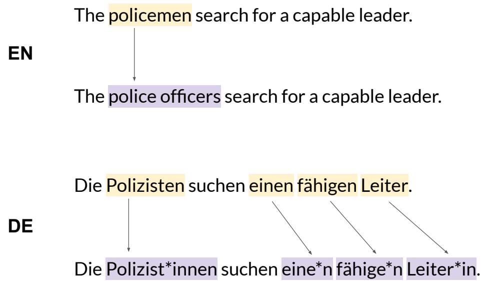

EN
The policemen search for a capable leader.
The police officers search for a capable leader.
DE
Die Polizisten suchen einen fähigen Leiter.
Die Polizist*innen suchen eine*n fähige*n Leiter*in.

Figure 2: Gendered words (yellow) that need to be altered for gender-fairness (purple) in English (EN) and German (DE).

2 Background

Gender-fair[2] rewriting is a conditional text generation problem. It can be approached with conventional sequence-to-sequence models trained on large amounts of parallel data, i.e., biased segments $\mathbf{S}$ and their gender-fair counterparts $\mathbf{T}$. Since such corpora do not exist in practice,[3] Sun et al. (2021) and Vanmassenhove et al. (2021) create artificial gender-fair target segments $\mathbf{T}_{\text{pseudo}}$ from existing source segments $\mathbf{S}$ with a rule-based de-biasing pipeline (Forward Augmentation, Figure 1a). The sequence-to-sequence model trained on the outputs of this pipeline removes the need for computationally expensive toolchains (such as dependency parsing) at runtime and increases robustness towards noisy inputs (Vanmassenhove et al., 2021).

2.1 Rule-based De-biasing for English

Converting $\mathbf{S}$ to $\mathbf{T}_{\text{pseudo}}$ is relatively straightforward for languages like English, which only expresses social gender in pronouns and a small set of occupation nouns. A simple dictionary lookup is often sufficient to produce a gender-fair variant of the biased text, except for two issues:

> **Issue 1: Rewritings that affect other dependencies**, e.g. when rewriting third-person subject pronouns in English, verbs in present tense need to be pluralised (e.g. “she knows” rewritten as “they know”).

[2] We define “gender-fair” as marking all genders explicitly; in contrast to “gender-neutral” where gender is deliberately *not* marked. For ease of reading, we only use the term “gender-fair” in this paper, even though our English models produce gender-neutral text.

[3] One exception: Qian et al. (2022) recently published a 100k segment crowd-sourced parallel corpus for English.

> **Issue 2: Ambiguities that need to be resolved from context**, e.g. English "her" can be a possessive pronoun (rewritten as "their") or a personal pronoun (rewritten as "them").

These issues are tractable for English because they only happen in a limited number of cases that can be covered with a limited number of rules. Sun et al. (2021) solve *Issue 1* based on part-of-speech, morphological and dependency information, and *Issue 2* by scoring different variants with a language model. Vanmassenhove et al. (2021) solve *Issue 1* using a grammatical error correction tool and *Issue 2* using part-of-speech, morphological and dependency information. Both Sun et al. (2021) and Vanmassenhove et al. (2021) train Transformer models (Vaswani et al., 2017) using the original texts as the source and the de-biased augmentations as target data (Forward Augmentation), and achieve WER below 1% on several test sets.

2.2 Rule-based De-biasing for Other Languages

In languages with more complex morphology, *Issue 1* is much more prevalent than in English but *Issue 2* is even more challenging because it requires animacy prediction: A direct application of Vanmassenhove et al.'s (2021) "first rule-based, then neural" approach to Spanish (Jain et al., 2021) results in a model that does not distinguish between human referents and objects. Similarly, Alhafni et al. (2022)[4] train an end-to-end rewriting system for Arabic but their pipeline for creating training data requires labelled data to train an additional classification model to identify gendered words. Both of these works only focus on a binary interpretation of gender. Non-sequence-to-sequence approaches have also been explored (Zmigrod et al., 2019; Diesner-Mayer and Seidel, 2022) but required extensive linguistic tools such as morphological, dependency and co-reference analysis, named entity recognition and word inflexion databases.

2.3 Round-trip Translation

Previous work employed round-trip translations to create pseudo data for automatic post-editing (Junczys-Dowmunt and Grundkiewicz, 2016; Freitag et al., 2019; Voita et al., 2019), grammatical error correction (Madnani et al., 2012; Lichtarge et al., 2019) or paraphrasing (Mallinson et al., 2017; Iyyer et al., 2018; Fabbri et al., 2021; Cideron et al., 2022). While such uses of round-trip translations exploit the fact that machine translations can be diverse and can contain accuracy and fluency errors, we are the first to exploit them for their social biases.

3 Backward Augmentation

We hypothesise that the data augmentation direction for gender-fair rewriters can be reversed (Figure 1b) without a negative impact on quality. Our motivation is rooted in work on data augmentation for MT (Sennrich et al., 2016b), where back-translation of monolingual target text tends to result in better quality than forward-translation of original source text (Khayrallah and Koehn, 2018; Bogoychev and Sennrich, 2019). We use Backward Augmentation to train a gender-fair rewriter model for English and compare its performance to the Forward Augmentation approach proposed by Vanmassenhove et al. (2021) and Sun et al. (2021).

3.1 Method

We propose to filter large monolingual corpora for gender-fair text $\mathbf{T}$ and use a rule-based pipeline to derive artificially biased source text $\mathbf{S}_{\text{pseudo}}$ from $\mathbf{T}$. Based on this data, we can train a sequence-to-sequence model which maximises $p(\mathbf{T}|\mathbf{S}_{\text{pseudo}}, \boldsymbol{\theta})$, rather than $p(\mathbf{T}_{\text{pseudo}}|\mathbf{S}, \boldsymbol{\theta})$ as in previous work (Section 2).

3.2 Experimental Setup

Data We extract English training data from OSCAR (Abadji et al., 2022), a large multilingual web corpus. For Forward Augmentation, we select segments that contain at least one biased word as $\mathbf{S}$, following Vanmassenhove et al.'s (2021) and Sun et al.'s (2021) lookup tables (Appendix E). For Backward Augmentation, we filter for segments that contain at least one of the corresponding gender-fair words in the lookup tables as $\mathbf{T}$. We filter out duplicates and noisy segments with `OpusFilter` (Aulamo et al., 2020) and then randomly subselect 5M segments each. For both models and as in previous work, we extend the training data by creating complementary source versions with only masculine forms, only feminine forms and copies of gender-fair targets and by adding additional non-gendered segments where no rewriting is necessary (amounting to 30% of the total data). A full overview of the training data can be found in Appendix A.

[4] The latest continuation of a series of work for gender-fair rewriting in Arabic (Habash et al., 2019; Alhafni et al., 2020).

	Sun et al.		Vanmassenhove et al.		
	non-gendered	**gendered**	**OpenSubtitles**	**Reddit**	**WinoBias+**
Source (no rewriting)	**0.00**	10.72	14.03	10.85	8.70
Forward Augmentation:					
(a) Vanmassenhove et al. (2021)	-	-	0.43	0.75	0.09
(b) Sun et al. (2021)	**0.00**	**0.57**	-	-	-
(c) Reimplementation (a + b)	**0.00**	**0.42**	**0.30**	**0.46**	**0.05**
Backward Augmentation (this work)	**0.00**	**0.43**	**0.24**	**0.40**	**0.04**

Table 1: Tokenised WER (lower is better) of different rewriting approaches for English. Best systems (no other statistically significantly better) marked in bold; Backward Augmentation matches Forward Augmentation.

Rule-based Processing To be able to compare directly to previous work, we first reproduce the rule-based Forward Augmentation approach proposed by Sun et al. (2021) and Vanmassenhove et al. (2021) to create $\mathbf{T}_{\text{pseudo}}$ from $\mathbf{S}$. We combine their lookup tables (Appendix E) and re-implement their rules based on part-of-speech, morphological and dependency information via `spaCy`[5] (Honnibal et al., 2020), both for resolving ambiguities and producing the correct number for verbs. We decide to follow Sun et al. (2021) and use "themself" and not "themselves" as a gender-fair form of "herself" and "himself". Taking this implementation as a basis, we derive a Backward Augmentation pipeline by reversing the lookup tables and rules to map from $\mathbf{T}$ to $\mathbf{S}_{\text{pseudo}}$.

Model Architecture Following Sun et al. (2021) and Vanmassenhove et al. (2021), we train 6-layer encoder, 6-layer decoder Transformers (Vaswani et al., 2017) with 4 attention heads, an embedding and hidden state dimension of 512 and a feed-forward dimension of 1024. For optimization, we use Adam (Kingma and Ba, 2015) with standard hyperparameters and a learning rate of $5e-4$. We follow the Transformer learning schedule in Vaswani et al. (2017) with a linear warmup over 4,000 steps. The only differences to Sun et al. (2021) and Vanmassenhove et al. (2021) are that we train our models with `sockeye 3` (Hieber et al., 2022) and use a smaller joint byte-pair vocabulary (Sennrich et al., 2016c) of size 8k computed with `SentencePiece` (Kudo and Richardson, 2018).

3.3 Automatic Evaluation

Test Sets We benchmark our models with the test sets published in conjunction with our baselines:

- **Sun et al. (2021)**: Two test sets (gendered/non-gendered) with 500 sentence pairs each, from five different domains: Twitter, Reddit, news articles, movie quotes and jokes. For the gendered version, there are balanced numbers of sentences with feminine and masculine pronouns for each domain. The non-gendered source texts do not contain any forms that need to be rewritten and should not be changed.

- **Vanmassenhove et al. (2021)**: Three test sets from three different domains: OpenSubtitles (Lison and Tiedemann, 2016, 500 sentence pairs), Reddit (Baumgartner et al., 2020, 500 sentence pairs), and WinoBias+ (Zhao et al., 2018, 3,167 sentence pairs). Each test set has a balanced amount of gender-fair pronoun types.

We manually double-check the target side of all test sets from previous work and if necessary correct sporadic human annotation errors. The test sets used by Vanmassenhove et al. (2021) also cover grammatical error corrections outside the scope of gender-fair rewriting. To restrict evaluation to the phenomenon of interest, we produce a target side version that only covers gender-fair rewriting. Note that this means that the model outputs by Vanmassenhove et al. (2021) will perform slightly worse on this version of the test set than reported in their paper because this model also makes such additional grammatical corrections. We revert tokenization and change "themselves" to "themself" in the model outputs of Vanmassenhove et al. (2021) to be able to compare them against our models' outputs and our references.

Method We evaluate our English model outputs and compare them to previous work with tokenised[6] WER based on the Python package

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

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

jiwer[7]. We compute statistical significance $p < 0.05$ with paired bootstrap resampling (Koehn, 2004), sampling 1,000 times with replacement.

Results Results are shown in Table 1. Backward Augmentation matches the low WER of the original as well as our combined reproduction of the Forward Augmentation models by Sun et al. (2021) and Vanmassenhove et al. (2021), and performs slightly better than previous work on OpenSubtitles and Reddit and WinoBias+.

4 Round-trip Augmentation

Artificially biasing gender-fair target segments rather than de-biasing gender-biased source segments is especially useful for languages with grammatical gender and more complex morphology than English. Taking German as a running example, we would need some form of animacy prediction in Forward Augmentation to transform ambiguous nouns such as "Leiter" only if they refer to a person ("leader") and not to an object ("ladder"). In Backward Augmentation, we do not need an animacy prediction model since this information is implicitly encoded in the gender-fair forms, as seen in Figure 2. Nevertheless, defining rules for mapping gender-fair segments to gender-biased segments or vice versa requires expert knowledge and will likely never completely cover morphologically complex languages.

As an alternative to handcrafting rules, we propose to exploit the fact that current MT models generate inherently biased text: we create pseudo source segments via round-trip translation through a pivot language that (mostly) does not mark gender (Figure 1c). We use this method to train a gender-fair rewriter model for German, and benchmark it against a highly engineered fully rule-based baseline (Diesner-Mayer and Seidel, 2022).

4.1 Method

We propose to filter large monolingual corpora for gender-fair text $\mathbf{T}$ and use off-the-shelve MT to first translate $\mathbf{T}$ into a pivot language as $\mathbf{P}_{\text{pseudo}}$, and then translate $\mathbf{P}_{\text{pseudo}}$ back into the original language as $\mathbf{S}_{\text{pseudo}}$. As in Backward Augmentation, we use the resulting data to train a sequence-to-sequence model that maximises $p(\mathbf{T}|\mathbf{S}_{\text{pseudo}}, \boldsymbol{\theta})$. We enrich this framework with several extensions as detailed in the next section and evaluated separately in Section 4.3.

4.2 Experimental Setup

Data We filter OSCAR (Abadji et al., 2022) for German sentences that contain at least one gender-fair form with simple regular expressions that match gender-fair patterns (Appendix F). After creating pseudo sources with round-trip translation (see next paragraph) and removing duplicates and noisy segments with OpusFilter (Aulamo et al., 2020), we obtain 8.8M parallel sentences. As in our Backward Augmentation experiment (Section 3.2), we complement the augmented training data with copies of the gender-fair segments on the source side and non-gendered segments where no rewriting is necessary (amounting to 30% of total data).

Roundtrip Translation English is a natural choice for the pivot language since it does not express gender in most animate words, meaning that this information is often lost when translating from a language with grammatical gender to English. Indeed, we find that gender-fair forms are translated to generic masculine forms in about 90% of the cases when we translate them to English and back to German.[8] We make use of this bias to create pseudo source segments, without the need for any hand-crafted rules, by leveraging Facebook's WMT 2019 models (Ng et al., 2019) for German-to-English[9] and English-to-German[10]. To avoid training on other translation differences aside from gender-fair forms, we identify the counterparts of gender-fair words in the round-trip translation and merge those into the original gender-fair segment to form the pseudo source. We explain our merging algorithm in detail in Appendix C.

LM Prompting One potential issue we discovered with our training data is that gender-fair plural noun forms are much more frequent than gender-fair singular noun forms. To boost singular forms, we generate additional gender-fair training data by prompting GerPT2-large[11] (Minixhofer, 2020) – a large German language model – using a seed list of gender-fair animate nouns.[12] Since we do not want to bias the model towards segments that start with a prompt, we sentence split the language model outputs and only keep singular-form segments that either do not start with a prompt or that

[7] https://github.com/jitsi/jiwer

[8] As manually evaluated on a random set of 100 sentences.

[9] https://huggingface.co/facebook/wmt19-de-en

[10] https://huggingface.co/facebook/wmt19-en-de

[11] https://huggingface.co/benjamin/gerpt2-large

[12] See an example prompt and verification that prompting is also possible for other languages in Appendix D.

contain at least one other gender-fair form.

Gender Control As the majority of the nouns in the German round-trip outputs are masculine forms, we create additional training data by finetuning the English-to-German MT model on data marked with sentence-level gender tags (Appendix G), similar to a previous approach for controlling politeness in MT outputs (Sennrich et al., 2016a). We leverage the original training data[13] for the WMT 2019 shared task and finetune the original `wmt19-en-de` checkpoint for 50,000 steps with a batch size of 30 on a single GPU, following the official Hugging Face (Wolf et al., 2020) translation finetuning script. The resulting model does not always translate according to the given tag but produces much more balanced translations overall: with the feminine tag, only 36% of the produced forms are masculine as compared to 90% with the original checkpoint and 94% with the masculine tag.[14]

Model Architecture We train a Transformer model using the same hyperparameters and training procedure as in our Backward Augmentation experiment described in Section 3.2.

4.3 Automatic Evaluation

We compare the performance of our model to the rule-based rewriter by Diesner-Mayer and Seidel (2022); we are not aware of any neural rewriter for German.

Test Set Diesner-Mayer and Seidel (2022) evaluate their system on a subset of the TIGER Treebank (Brants et al., 2004) with German news articles from the 1990s. Since masculine forms are prevalent in these articles, we create a random set of 1,200 TIGER sentences with gendered forms which we balance for masculine, feminine, singular and plural forms, and a random set of 300 non-gendered sentences. For singular forms, we decide to create our test set with a balanced mix of forms referring to unspecific people as well as real persons. There are two reasons for this design choice: First, we want to closely mirror the setup in English, where e.g. any occurrence of "she" or "he" is rewritten to "they", irrespective of whether it refers to a specific person or not. Second, there are several cases where we cannot assume that an input text referring to a specific person uses their desired pronouns. One example is machine translation output from a language that does not mark gender on pronouns. We believe that a rewriter that produces gender-fair pronouns (i.e. mentioning all genders) is less biased than one that assumes all gender mentions in the input text are correct (while those could be actual instances of misgendering).

	TIGER	
	non-gendered	**gendered**
Source (no rewriting)	**0.00**	20.56
Diesner-Mayer and Seidel (2022)	0.17	15.01
Round-trip Augmentation (this work)	0.13	17.95
+ merged	0.29	16.36
+ merged + LM prompting	0.27	15.37
+ merged + gender control	0.17	14.02
+ all	0.17	**13.18**

Table 2: Tokenised WER (lower is better) of different rewriting approaches for German. Best systems (no other statistically significantly better) marked in bold.

Method We compute tokenised WER and statistical significance as in Section 3.3.

Results Results are shown in Table 2. Our simplest model (Round-trip Augmentation), which is only trained on filtered data round-tripped with the original Facebook model, already reduces the WER compared to the biased inputs from the test set (no rewriting). Avoiding differences in roundtrip translations aside from gender-fair forms (+ merged) further reduces the WER, as do additional training examples obtained through a language model (+ LM prompting) and an MT system finetuned for gender control (+ gender control).

Combining all of these extensions into a single model (+ all) results in the best WER. It also performs surprisingly well compared to Diesner-Mayer and Seidel (2022), a rule-based system that uses an abundance of language-specific NLP tools for gender-fair rewriting (Section 2.2).

4.4 Human Evaluation

While making biased text more gender-fair according to our automatic evaluation, our best-performing model still produces numerous errors: 13.18% of the words in the model's output differ from the gender-fair reference texts in our test set (Table 2). We conduct a human quality rating experiment to assess whether the imperfect outputs of this model are perceived as more gender-fair and

[13] https://huggingface.co/datasets/wmt19

[14] As manually evaluated on a random set of 100 sentences.

whether erroneous gender-fair forms in general are preferred over unaltered gender-biased text.

Participants To refrain from convenience sampling and to focus on potential beneficiaries of our model (i.e., people who may be offended by gender-biased text), we recruit 294 volunteers (141 female, 82 non-binary, 55 male, 16 other) through unpaid posts in newsletters and social media channels of special interest groups against gender discrimination (Appendix I). Participation is anonymous and voluntary.

Materials We select one paragraph each from six unrelated German texts (P1–6, Appendix H). Each paragraph contains at least one gender-biased word and is about an unspecific person whose gender is unknown, a specific non-binary person, and/or a group of people whose gender is unknown. We use three transformations of each paragraph in the experiment: unaltered (Original), automatically de-biased by our best model (Rewriter), and manually de-biased by one of the authors (Gold).

Task and Procedure We use a 6 (paragraphs) x 3 (transformations) mixed factorial design, implemented as a questionnaire in three versions (A–C) to which participants are assigned at random. Participants see all factor levels but not all combinations: to avoid repetition priming, each questionnaire includes all six paragraphs in the same order but in different transformations. For example, P1 is presented as Original in questionnaire A and as Rewriter in questionnaire B, etc.; participants are not informed if and how the paragraphs were transformed for gender-fairness.

After completing a pre-experiment survey with demographic questions, participants are shown a single paragraph at a time and asked if that paragraph is gender-fair, which they answer using a 5-point Likert scale (1: strongly disagree, 5: strongly agree). A short post-experiment survey on general opinions regarding gender-fair writing concludes the experiment.

Results The distribution of Likert-scale ratings by transformation is shown in Figure 3. Albeit not as good as the human references (Gold, mean=3.98), our model outputs (Rewriter, mean=2.93) are rated better than the unaltered gender-biased paragraphs (Original, mean=1.79) overall. This finding equally holds for all individual paragraphs (Table 3): Rewriter consistently outperforms Original in terms of perceived gender-fairness and, in some cases, comes close to Gold (P5, P6).

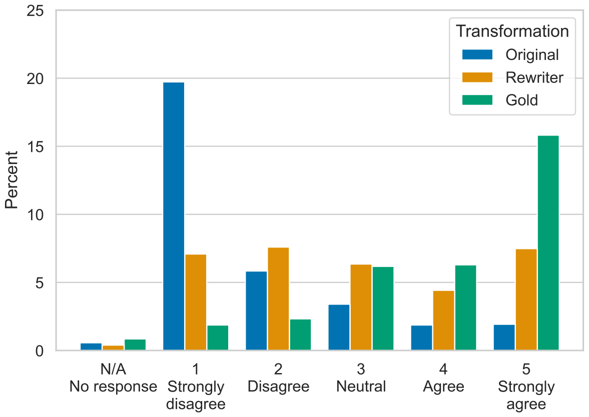


Figure 3: Distribution of responses (N=1,764) to the statement “This text is gender-fair” over all paragraphs.

	P1	P2	P3	P4	P5	P6	All
Original	2.25	1.82	1.27	1.67	1.57	2.18	1.79
Rewriter	2.97	3.27	1.86	2.21	3.56	3.71	2.93
Gold	4.21	3.91	4.03	3.67	4.05	4.01	3.98

Table 3: Mean rating on gender-fairness by paragraph (P1–6) and overall (All).

While these findings confirm – as in our automatic evaluation in Section 3.3 – that our model outputs are more gender-fair than original input texts, it leaves the most relevant question for use in practice unanswered: if given the choice, would users choose potentially erroneous Rewriter outputs over error-free but gender-biased original texts? We include this question in the post-experiment survey independently of any specific text, and compare participants’ responses with their average rating for Original and Rewriter outputs in the main part of the experiment. Out of the 294 participants, 201 (68.37%) disagreed or strongly disagreed that “If a text has errors in gender-fair wording, I prefer an error-free non-gender-fair version (e.g., generic masculine) instead.”, and 198 (98.51%) of these participants rated Rewriter more gender-fair than Original in the experiment. In comparison, 35 (68.63%) of the 51 participants who agreed or strongly agreed with that statement still gave higher gender-fairness scores to Rewriter in the experiment (where, recall from above, the type of transformation was not revealed).

5 Discussion

Our experimental findings yield insights for future research and design implications for real-world NLP applications.

Biased models are useful for de-biasing. At least in the subfield of gender-fair rewriting, debiasing research has focussed extensively on human annotation (Qian et al., 2022) and rule-based processing and training data creation (e.g., Sun et al., 2021; Vanmassenhove et al., 2021; Jain et al., 2021; Alhafni et al., 2022; Diesner-Mayer and Seidel, 2022). Conversely, our work demonstrates that robust de-biasing rewriters can be implemented by leveraging inherently biased NLP models. Our Round-trip Augmentation experiment covers a single language, but we note that both the training data (Abadji et al., 2022) and MT models (Ng et al., 2019) we leverage are readily available for many other languages. We also assume that (biased) MT models – the only requirement for gender-fair rewriters based on Round-trip Augmentation apart from simple data filters – are available for more languages or are easier to create than the different NLP tools used in typical rule-based augmentation pipelines, such as robust models for lemmatisation, morphological analysis, dependency parsing, named entity recognition, and co-reference resolution.

Rule-based de-biasing lacks robustness. Handwritten rules are limited by design. For example, a breakdown of the results shown in Table 2 (see Appendix B) reveals that Diesner-Mayer and Seidel's (2022) rewriter handles masculine forms better than our model (with a WER as low as 7.81). However, their rewriter performs poorly with feminine forms (with a WER as high as 22.22, which is worse than the WER of the biased input texts) likely because these forms are not covered in its rule set. Additionally, we find that while Diesner-Mayer and Seidel's (2022) approach features a solution for compounds, e.g. it can correctly de-bias "Strassenbauarbeiter" (road construction worker), this only applies to compounds where the gendered word appears at the end, e.g. it does not de-bias "Arbeitergesetz" (employee legislation). Furthermore, their approach uses a word database to identify gendered nouns which does not generalise to unknown gendered words. Finally, there is always a risk of error propagation with the NLP tools used in rule-based approaches. We conclude that language-specific rule sets will likely never cover all relevant phenomena for gender-fair rewriting. As shown by previous work, seq2seq models provide a model-based alternative that boosts generalisation (Vanmassenhove et al., 2021) which is why they should be used in as many languages as possible. We believe that Round-trip Augmentation provides an easy way to create parallel data to train gender-fair rewriting models for new languages without the need for in-depth linguistic knowledge of the language.

Users prefer errors over bias. Potential beneficiaries of gender-fair rewriters – the 294 participants of our human evaluation campaign – rated the outputs of our German rewriter as more gender-fair than the biased original texts and explicitly (in the post-experiment survey) stated they prefer (potentially erroneous) gender-fair forms over error-free non-gender-fair forms. This is an important finding because our model is far from perfect, as evidenced by a high error rate compared to English and manual inspection of the outputs used in the evaluation campaign (Appendix H). Previous work has found that non-binary people consider NLP as harmful, particularly due to the risk of being misgendered by MT outputs (Dev et al., 2021). Vanmassenhove et al. (2021) caution that gender-fair rewriters may not be applicable to languages other than English because "few languages have a crystallized approach when it comes to gender-neutral pronouns and gender-neutral word endings." While there is an active debate (and no established standard) about the form that gender-fair German should take (e.g., Burtscher et al., 2022), our evaluation campaign makes a strong case for using NLP technology – even if not flawless – to offer *one* form of de-biased text in real-world applications. Since rewriters are relatively lightweight models that operate independently from any input provider, be it a human author or a biased MT model, they would seem suitable for integration into a wide range of systems with reasonable effort.

6 Conclusions and Future Work

Despite an impressive performance in a wide range of tasks and applications, state-of-the-art NLP models contain numerous biases (Stanovsky et al., 2019; Nadeem et al., 2021; Renduchintala and Williams, 2022). Our work shows that knowledge of a bias can be used to correct that bias with the biased models themselves. In the case of gender-fair rewriting, we demonstrated that reversing the data augmentation direction and using round-trip translations from biased MT models can substitute the prevalent rewriting paradigm that relies on handcrafted and often complex rules on top of morphological analysers, dependency parsers, and many other NLP tools. In our case study for German, our model surpasses the performance of a strong baseline in terms of WER and produces outputs that were perceived as more gender-fair than unaltered biased text in a human evaluation campaign with 294 potential beneficiaries.

While our approach enables the application of gender-fair rewriting to any language for which translation models exist, we believe there are several other uses cases where biased models can be leveraged for de-biasing, including dialect rewriting (Sun et al., 2022), subjective bias neutralisation (Pryzant et al., 2020), and avoiding discrimination in dialogue systems (Sheng et al., 2021a).

Limitations

While we consider our approach more easily applicable to new languages than rule-based Forward Augmentation, it relies on the existence of sufficient original gender-fair text in the language of interest and it is currently unclear what the minimum amount of parallel data is to learn a gender-fair rewriting model. Additionally, our survey only targets affinity groups which limits the generalisability of our results to all German speakers. Since people who choose to not use gender-fair language can simply not use a rewriting system, we do not think that this lack of generalisability is a problem in this case. Another limitation is that we use a specific form of gender-fair German in our survey. We made participants aware of this in a disclaimer at the beginning of the survey. It should be stated that there are many different acceptable gender-fair forms in German (see Section F). While using a different gender-fair form could affect the individual ratings in our survey, we do not expect that it would change our finding that Rewriter outputs are rated more gender-fair than the Original texts.

Ethics Statement

Participation in our study was voluntary and fully anonymous. We did not collect any personal data that would allow us to identify people and did not exclude any participants unless they specifically requested their participation be ignored in the last open commentary field of our survey or they stated technical difficulties. Concerning our rewriting models, we did not filter the publicly available data to exclude harmful content. However, since our models mainly learn to copy text, we do not believe they will hallucinate such text of their own accord.

Acknowledgements

We thank Marcos Cramer for their valuable inputs and support with our human evaluation and all contacts who shared our survey on their mailing lists and social media accounts (Appendix I) as well as all survey participants. Many thanks also go to the OSCAR project who shared their data with us. Further, we are grateful to Alex Flückiger, Anne Göhring, Nikita Moghe and the anonymous reviewers for their helpful feedback. Chantal Amrhein and Rico Sennrich received funding from the Swiss National Science Foundation (project MUTAMUR; no. 176727).

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2020. 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). Association for Computational Linguistics.

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022. 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. Association for Computational Linguistics.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. 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. Association for Computational Linguistics.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):830–839.

Nikolay Bogoychev and Rico Sennrich. 2019. Domain, translationese and noise in synthetic data for neural machine translation.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. Tiger: Linguistic interpretation of a german corpus. *Research on Language and Computation*, 2(4):597–620.

Sabrina Burtscher, Katta Spiel, Lukas Daniel Klausner, Manuel Lardelli, and Dagmar Gromann. 2022. "es geht um respekt, nicht um technologie": Erkenntnisse aus einem interessensgruppen-Übergreifenden workshop zu genderfairer sprache und sprachtechnologie. In *Proceedings of Mensch Und Computer 2022*, page 106–118, New York, NY, USA. Association for Computing Machinery.

Prafulla Kumar Choubey, Anna Currey, Prashant Mathur, and Georgiana Dinu. 2021. GFST: Gender-filtered self-training for more accurate gender in translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1640–1654, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Geoffrey Cideron, Sertan Girgin, Anton Raichuk, Olivier Pietquin, Olivier Bachem, and Léonard Hussenot. 2022. vec2text with round-trip translations.

Ander Corral and Xabier Saralegi. 2022. Gender bias mitigation for nmt involving genderless languages. In *Seventh Conference on Machine Translation (WMT22)*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marta R. Costa-jussà and Adrià de Jorge. 2020. Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online). Association for Computational Linguistics.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Theodor Diesner-Mayer and Niels Seidel. 2022. Supporting gender-neutral writing in german. In *Proceedings of Mensch Und Computer 2022*, MuC '22, page 509–512, New York, NY, USA. Association for Computing Machinery.

Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2021. 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. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. 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. Association for Computational Linguistics.

Marcus C. G. Friedrich and Elke Heise. 2019. Does the use of gender-fair language influence the comprehensibility of texts? *Swiss Journal of Psychology*, 78(1-2):51–60.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. 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. Association for Computational Linguistics.

Felix Hieber, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, Surafel Lakew, Prashant Mathur, Anna Currey, and Marcello Federico. 2022. Sockeye 3: Fast neural machine translation with pytorch.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Lisa K. Horvath, Elisa F. Merkel, Anne Maass, and Sabine Sczesny. 2016. Does gender-fair language pay off? the social perception of professions from a cross-linguistic perspective. *Frontiers in Psychology*, 6.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Nishtha Jain, Maja Popović, Declan Groves, and Eva Vanmassenhove. 2021. Generating gender augmented data for NLP. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 93–102, Online. Association for Computational Linguistics.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. 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. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. 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. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. 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. Association for Computational Linguistics.

Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. 2019. 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. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Nitin Madnani, Joel Tetreault, and Martin Chodorow. 2012. Exploring grammatical error correction with not-so-crummy machine translation. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 44–53, Montréal, Canada. Association for Computational Linguistics.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. 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. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Benjamin Minixhofer. 2020. GerPT2: German large and small versions of GPT2.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. 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 5356–5371, Online. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. 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. Association for Computational Linguistics.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):480–489.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Williams Adina. 2022. Perturbation augmentation for fairer nlp. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Online and Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Adithya Renduchintala and Adina Williams. 2022. Investigating failures of automatic translationin the case of unambiguous gender. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3454–3469, Dublin, Ireland. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online). Association for Computational Linguistics.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2022. First the worst: Finding better gender translations during beam search. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3814–3823, Dublin, Ireland. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. 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. Association for Computational Linguistics.

Emily Sheng, Josh Arnold, Zhou Yu, Kai-Wei Chang, and Nanyun Peng. 2021a. Revealing persona biases in dialogue systems.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021b. 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. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2022. Dialect-robust evaluation of generated text.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral english.

Julia Tibblin, Joost van de Weijer, Jonas Granfeldt, and Pascal Gygax. 2022. There are more women in joggeur·euses than in joggeurs: On the effects of gender-fair forms on perceived gender ratios in french role nouns. *Journal of French Language Studies*, page 1–24.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. 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. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Dries Vervecken, Pascal Gygax, Ute Gabriel, Matthias Guillod, and Bettina Hannover. 2015. Warm-hearted businessmen, competitive housewives? effects of gender-fair language on adolescents' perceptions of occupations. *Frontiers in Psychology*, 6.

Dries Vervecken, Bettina Hannover, and Ilka Wolter. 2013. Changing (s)expectations: How gender fair job descriptions impact children's perceptions and interest regarding traditionally male occupations. *Journal of Vocational Behavior*, 82(3):208–220.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 877–886, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Additional Data Details

When filtering our data with `OpusFilter` (Aulamo et al., 2020), we define noisy segments as segments that do not pass the following filters:

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

The individual dataset sizes after deduplication and filtering can be seen in Table 4. Note that for English, we restricted the total of gendered data to 15 million parallel segments to be comparable to Sun et al. (2021) by only considering a subset of the English portion of the OSCAR corpus (Abadji et al., 2022).

For German, we use Round-trip Augmentation to produce the gendered pseudo sources. The finetuned MT model we use to produce feminine pseudo sources sometimes produces round-trip translations that are identical to the gender-fair segments or the round-trip translations with the original checkpoint. Consequently, the Table has fewer parallel segments for this category because fewer unseen segments are added to the training data overall for + gender control models in Table 2.

	# SRC segments		
	masculine	feminine	gender-fair
English			
OSCAR SRC	∼ 5.0M	∼ 5.0M	∼ 5.0M
OSCAR TRG	∼ 5.0M	∼ 5.0M	∼ 5.0M
German			
OSCAR TRG	∼ 8.8M	∼ 8.7M	∼ 8.8M
LM TRG	∼ 3.4M	∼ 2.0M	∼ 3.4M

Table 4: Data statistics for all rewriter models. Gender-fair sources are copies of the target segments before normalisation. For English, masculine and feminine are created with a rule-based approach. For German, masculine refers to the round-trip translations from the original MT model checkpoint, and feminine refers to the round-trip translations from the finetuned checkpoint with the feminine tag present.

Our English models in Table 1 are trained on the following dataset combinations:

- **Forward Augmentation Reimplementation (a+b)**: OSCAR SRC masculine + feminine + gender-fair
- **Backward Augmentation**: OSCAR TRG masculine + feminine + gender-fair

Our German models in Table 2 are trained on the following dataset combinations:

- **Round-trip Augmentation (+ merged)**: OSCAR TRG masculine + gender-fair
- **+ LM prompting**: OSCAR TRG masculine + gender-fair and LM TRG masculine + gender-fair
- **+ gender control**: OSCAR TRG masculine + feminine + gender-fair
- **+ all**: OSCAR TRG masculine + feminine + gender-fair and LM TRG masculine + feminine + gender-fair

Note that for every combination of data sets, we also add non-gendered data such that the non-gendered data from OSCAR makes up 30% of the total parallel data.

B Detailed Results for German

We provide a more detailed evaluation of our results for German in Table 5. The first two columns show the results isolated for generic feminine and generic masculine forms in the input that should be rewritten as gender-fair forms. The third and fourth columns show the results grouped by plural and singular gender-fair forms, respectively. This evaluation highlights two points. First, we can see our strategies introduced in Section 4.1 are effective for the cases they were designed for: Using a gender-aware machine translation model for round-trip translation (+ gender control) is particularly helpful on the feminine test set and adding language model generated singular-form training data (+ LM prompting) reduces the WER on the singular test set significantly. Second, as discussed in Section 5, the results show that the rule-based approach by Diesner-Mayer and Seidel (2022) is limited by design. While it performs well on generic masculine forms, it does not cover generic feminine forms at all which results in a higher WER than the source where no rewriting is performed.

C Merging Algorithm

One issue with round-trip translations is that they are likely to contain edits unrelated to the

	TIGER			
	feminine	**masculine**	**plural**	**singular**
Source (no rewriting)	22.05	19.07	15.98	24.06
Diesner-Mayer and Seidel (2022)	22.22	**7.81**	10.39	**18.55**
Round-trip Augmentation (this work)	19.76	16.13	11.04	23.22
+ merged	18.99	13.72	8.29	22.52
+ merged + LM prompting	17.51	13.24	8.05	20.97
+ merged + gender control	14.00	14.04	**4.12**	21.58
+ all	**13.03**	13.33	4.67	19.69

Table 5: Tokenised WER (lower is better) of different rewriting approaches for German evaluated separately for different test cases. Best systems (no other statistically significantly better) marked in bold.

gender-fair words in the target:

original (de) Denn jede Begegnung mit einem*r Schüler*in ist anders, auch die Familien sind verschieden.

interm. (en) Because every encounter with a student is different, even the families are different.

round-trip (de) **Weil** jede Begegnung mit **einem Schüler anders ist, sind** auch die Familien **anders**.

This round-trip translation not only contains the desired generic masculine form for student (marked in orange) but also several other deviations from the original sentence (marked in bold) which even changes the meaning slightly.

Ideally, we would want to identify the generic masculine form and allow no other changes in the pseudo source:

merged (de) Denn jede Begegnung mit **einem Schüler** ist anders, auch die Familien sind verschieden.

To this end, we develop a merging algorithm that aims to insert the generic forms in the round-trip translation into the context of the original gender-fair segment. For all gender-fair words in the target segment, we check if there are any close matches in the round-trip translation using the `difflib` Python library[15] with a cutoff of 0.6. If yes, we replace the gender-fair word with its closest match. This process can be seen in Algorithm 1. If not all gender-fair words can be matched in the round-trip translation, we keep the round-trip translation as our pseudo source and do not merge with the gender-fair target. Note that this merging algorithm can potentially introduce case and other grammatical errors in the pseudo source. This is, however, not a serious problem as these potentially non-grammatical forms will only occur on the source side, meaning our model does not learn to produce such forms on the target side. The merging algorithm could be improved in the future to also consider grammatical acceptability, e.g. by scoring the merged pseudo source against the gender-fair target with a language model.

Algorithm 1 Merging Round-Trip Translations

```
Input: list of tokens in gender-fair target $t$, list of
    tokens in RT translation $r$
Output: merged pseudo source $s$
for token $w$ at index $i$ in $t$ do
    if $w$ is gender-fair form then
        $m$ = get_close_matches($w$, $r$, 0.6)
        if len($m$) > 0 then $t[i] = m[0]$
        end if
    end if
end for
$s$ = detokenise($t$)
```

D Additional LM Prompting Details

Our prompts consist of a gender-fair determiner and noun from a seed list, following the pattern “Ein*e NOUN*in”. An example can be seen here with the prompt in bold and gender-fair forms generated by the language model in orange:

German: **Ein*e Leiter*in** für unser Team „Soziales“. Als Ansprechpartner*in in Sachen Finanzen oder für Veranstaltungen, mit einem Blick für’s Wesentliche sind Sie für uns eine Bereicherung Wir suchen eine*n freundliche*n, zuverlässige*n und flexible*n Mitarbeiter*in für eine geringfügige Beschäftigung. Sie haben Freude am Umgang mit Kolleg*innen und eine hohe Einsatzbereitschaft. Auch

[15] https://docs.python.org/3/library/difflib.html#difflib.get_close_matches

wenn Sie noch keine Erfahrung mitbringen lohnt sich dies schon sehr. Eine abgeschlossene Ausbildung als Köch*in oder eine vergleichbare Qualifikation sind von Vorteil. Es erwartet Sie ein spannendes Arbeitsumfeld und ein gutes Betriebsklima sowie verschiedene Weiterbildungs- und Fortbildungsmöglichkeiten.

Prompting large language models for gender-fair text also works with other language models and for languages other than German. Here, we show example gender-fair prompts (marked in bold) and the text generated by a multilingual language model[16]. Many forms in the generated text are also gender-fair (marked in orange) even over long distances. Not all generated forms include non-binary people (e.g. French "il/elle") but these can easily be normalised.

French: **Un·e citoyen·ne** parmi tant d'autres, qui vit sur la commune et la fréquente dans toute sa diversité. L'un·e qui a envie de vous donner de son temps pour la ville dans laquelle il/elle vit. C'est avant tout une adhésion à notre vision d'une ville plus durable, plus juste, plus proche de vous. En devenant membre, vous agissez directement en faveur de la démocratie locale, parce que les citoyen·nes comme vous peuvent exprimer leur avis, ou défendre les valeurs qu'ils et elles partagent et portent avec nous. Et vous pouvez le faire par vos votes aux assemblées générales, ou dans le cadre des groupes thématiques dans lesquels vous êtes invité·e à vous engager.

Spanish: **Unx companerx** presx. Esa son nuestras armas, pensadlo. Vayamos al Estado con la fuerza de todas nuestras rebeliones, hacernos fuertes en todo y donde todxs podamos participar, unificándonos como un movimiento, concentrando la rabia y lanzándola contra quienes nos mantienen en la miseria, que eso es lo que hacen estos gobiernos a base de recortes, no trabajan para el pueblo, se han posicionado en contra de lxs más desprotegidxs, de lxs desempleadxs, de lxs jóvenes, de lxs de mediana edad. Contra todxs nosotrxs, contra lxs que resistimos al poder y luchamos en sus filas de igual a igual, contra lxs de arriba y abajo, contra el capitalismo y el Estado y su falsa democracia.

E English Lookup Tables

For completeness, we list the lookup tables used in our reproduction of Vanmassenhove et al. (2021) and Sun et al. (2021) and in the other direction for our Backward Augmentation.

E.1 Pronouns

he, she	↔	they
his, her	↔	their
him, her	↔	them
his, hers	↔	theirs
himself, herself	↔	themself

E.2 Nouns

Gender-neutral alternatives for gender-marked job titles

chairman, chairwoman	↔	chairperson
chairmen, chairwomen	↔	chairpeople
anchorman, anchorwoman	↔	anchor
anchormen, anchorwomen	↔	anchors
congressman, congresswoman	↔	member of congress
congressmen, congresswomen	↔	members of congress
policeman, policewoman	↔	police officer
policemen, policewomen	↔	police officers
spokesman, spokeswoman	↔	spokesperson
spokesmen, spokeswomen	↔	spokespeople
steward, stewardess	↔	flight attendant
stewards, stewardesses	↔	flight attendants
headmaster, headmistress	↔	principal
headmasters, headmistresses	↔	principals
businessman, businesswoman	↔	business person
businessmen, businesswomen	↔	business persons
postman, postwoman	↔	mail carrier
postmen, postwomen	↔	mail carriers
salesman, saleswoman	↔	salesperson
salesmen, saleswomen	↔	salespersons
fireman, firewoman	↔	firefighter
firemen, firewomen	↔	firefighters
barman, barwoman	↔	bartender
barmen, barwomen	↔	bartenders
cleaning man, cleaning lady	↔	cleaner
cleaning men, cleaning ladies	↔	cleaners
foreman, forewoman	↔	supervisor
foremen, forewomen	↔	supervisors

Gender-neutral alternatives for generic 'man'

average man	↔	average person
average men	↔	average people
best man for the job	↔	best person for the job
best men for the job	↔	best people for the job
layman	↔	layperson
laymen	↔	laypeople
man and wife	↔	husband and wife
mankind	↔	humankind
man-made	↔	human-made
workmanlike	↔	skillful
freshman	↔	first-year student

Gender-neutral alternatives for unnecessary feminine forms

actress	↔	actor
actresses	↔	actors
heroine	↔	hero
heroines	↔	heroes
comedienne	↔	comedian
comediennes	↔	comedians
executrix	↔	executor
executrices	↔	executors
poetess	↔	poet
poetesses	↔	poets
usherette	↔	usher
usherettes	↔	ushers
authoress	↔	author
authoresses	↔	authors
boss lady	↔	boss
boss ladies	↔	bosses
waitress	↔	waiter
waitresses	↔	waiters

F German Gender-fair Patterns

For German, we cannot use a lookup approach to identify gender-fair noun forms but rather work with several gender-fair patterns. Here, we describe the different gender-fair forms we consider and, for each, show a plural form example and its corresponding pattern:

Pair forms are forms that explicitly state the feminine and masculine form connected with a coordinating conjunction like "and" or "or". The

[16] https://huggingface.co/bigscience/bloom

order of the feminine and masculine forms can be variable. This form assumes binary gender and does not include non-binary people.

Example: Studentinnen und Studenten

Pattern: (\S{2,})innen und -?\1(?!innen)(en|e|n)?

Binnen-I forms are forms that take the feminine form but with a capitalised "I" at the beginning of the feminine suffix "innen" (plural) or "in" (singular). This form also assumes binary gender.

Example: StudentInnen

Pattern: \w+Innen

Gender slash forms are forms that take the feminine form but with a slash ("/") separating the feminine suffix "innen" (plural) or "in" (singular) from the stem. This form also assumes binary gender.

Example: Student/innen

Pattern: \w+\ ?/\ ?innen

Gender gap forms are forms that take the feminine form but with an underscore ("_") separating the feminine suffix "innen" (plural) or "in" (singular) from the stem. This form includes non-binary people.

Example: Student_innen

Pattern: \w+_innen

Gender colon forms are forms that take the feminine form but with a colon (":") separating the feminine suffix "innen" (plural) or "in" (singular) from the stem. This form also includes non-binary people.

Example: Student:innen

Pattern: \w+:innen

Gender star forms are forms that take the feminine form but with an asterisk ("*") separating the feminine suffix "innen" (plural) or "in" (singular) from the stem. This form also includes non-binary people.

Example: Student*innen

Pattern: \w+*innen

Alternatively, and out of the scope of this work, gender-fair text can also use present participles as gender-neutral nouns (only gender-neutral in plural forms, e.g. Studierende - "those who are studying"), synonymous gender-neutral nouns or it can completely avoid gendered words and express content with structures where no gender-fair forms are needed (e.g. "bei den Dorfbewohner*innen" - "among the villagers" could also be expressed as "im Dorf" - "in the village"). We believe that our proposed approach can be extended to those cases in the future.

G Gender-Tagged Data With Pair Forms

To make the English-to-German machine translation model that we use for round-trip translations gender-aware, we finetune on artificial data with sentence-level gender labels. Previous work presented several approaches how parallel data can be filtered or created to specifically contain masculine or feminine forms (Costa-jussà and de Jorge, 2020; Saunders and Byrne, 2020; Choubey et al., 2021; Corral and Saralegi, 2022). In our work, we create such data by making use of pair forms (see Appendix F) in existing parallel data that consist of the feminine and the masculine form of the same noun:

SRC: **Students** from many nations learn together here.

TRG: **Schülerinnen und Schüler** aus vielen Nationen lernen hier gemeinsam.

Using a simple replace operation, we can use this data to create two contrasting targets in German and tag them with a corresponding tag that indicates whether the translation should contain feminine or masculine noun forms:

SRC: **<f> Students** from many nations learn together here.

TRG: **Schülerinnen** aus vielen Nationen lernen hier gemeinsam.

SRC: **<m> Students** from many nations learn together here.

TRG: **Schüler** aus vielen Nationen lernen hier gemeinsam.

This is possible for plural nouns because all German plural nouns share inflexion across genders which means that no rewriting is necessary for pronouns, adjectives or determiners that refer to them. For singular forms, we cannot easily construct contrasting versions because sometimes additional modifications to pronouns, adjectives or determiners are necessary to preserve the grammatical agreement. Instead, we create feminine examples for singular pair forms if the first form in the pair form is feminine and we create masculine examples if the first form in the pair form is masculine.

Pair forms are not only specific to German but are also common in many other languages. For example, the United Nations advise the use of pair forms in all their official languages with grammatical gender - Arabic, French, Russian and Spanish - as well as in English and Chinese for added emphasis when gender is relevant for communication. Our approach to generating finetuning data for gender-aware machine translation models based on pair forms is not limited to German but is applicable to other languages as well.

H Overview of Survey Texts

The six text excerpts used in our survey and the corresponding three versions can be seen in Table 6.

P1 is from a website informing about study regulations for becoming a forester and uses generic masculine forms and gender slash forms (that do not include non-binary people) in plural and singular. P2 is from an online privacy policy of an insurance company and uses generic masculine forms in plural and singular. P3 is from game instructions and uses generic masculine in singular. P4 is from a news article about Ezra Miller who uses "they/them" pronouns in English and the text uses generic masculine forms in singular. P5 is from a company blog and uses generic masculine forms in plural and singular. P6 is from a package insert of a birth control pill and uses generic masculine forms in singular and generic feminine forms in plural. (Websites last accessed on 27.12.2022)

I List of Survey Contacts

Our survey was kindly shared on mailing lists or other platforms by the networks below.

Austria:

- Venib - Verein Nicht-Binär

Germany:

- MinaS - Verein für Menschen im nichtbinären und agender Spektrum
- Verein für geschlechtsneutrales Deutsch e.V.

Switzerland:

- Gender Campus
- nonbinary.ch
- Queerstudents Bern
- Queers usem Kaff
- romanescos mailing list
- Transgender Network Switzerland
- Verein Geschlechtergerechter
- Zurich Pride

Each contact was provided with a link to the survey and the following accompanying text (English translation below):

German original:

> "Das Institut für Computerlinguistik (Uni Zürich) und Textshuttle arbeiten an Textgenerierungssystemen, die genderfair(er)e Sprache ausgeben und möchten in einer Umfrage untersuchen, wie unterschiedliche Texte in Bezug auf genderfaire Sprache wahrgenommen werden.
>
> Die Umfrage dauert 10–15 Minuten und ist anonym. Es geht darum, verschiedene Textausschnitte zu lesen und diese zu ihrer Verständlichkeit und Inklusivität zu bewerten. Über untenstehenden Link geht es zur Google Form der Umfrage:
>
> LINK TO SURVEY"

English translation:

> "The Department of Computational Linguistics (University of Zurich) and Textshuttle work on text generation systems that output (more) gender-fair language and want to explore in a survey how different texts are perceived with respect to gender-fair language.
>
> The survey takes 10-15 minutes and is anonymous. The goal is to read different text excerpts and to rate them according to their understandability and their inclusivity. Following the link below you can reach the Google Form for the survey:
>
> LINK TO SURVEY"

J Intended Use

- The models presented in this paper are intended to rewrite biased text with possible gender-fair forms; they are *not* intended to identify a person's gender nor to prescribe a particular gender-fair form.

	Original Text	Rewriter Output	Human Reference
P1	Zugelassen zur Försterausbildung werden Kandidaten mit einem eidg. Fähigkeitszeugnis als Forstwart/in (oder einer gleichwertigen Ausbildung). Erforderlich sind zudem 12 Monate Praxis in einem Forstbetrieb oder -unternehmen. Zudem müssen Interessenten die nachfolgenden Grundlagenmodule und die Eignungsprüfung absolviert und bestanden haben. Details dazu sind erhältlich bei den Anbietern: BZW Lyss und ibW BZW Maienfeld. Angehende Forstwart-Vorarbeiter/innen und Förster/innen besuchen die gleichen Grundlagenmodule.	Zugelassen zur Försterausbildung werden Kandidaten mit einem eidg. Fähigkeitszeugnis als Forstwart*in (oder einer gleichwertigen Ausbildung). Erforderlich sind zudem 12 Monate Praxis in einem Forstbetrieb oder -unternehmen. Zudem müssen Interessent*innen die nachfolgenden Grundlagenmodule und die Eignungsprüfung absolviert und bestanden haben. Details dazu sind erhältlich bei den Anbietern: BZW Lyss und ibW BZW Maienfeld. Angehende Forstwart-Vorarbeiter*innen und Förster*innen besuchen die gleichen Grundlagenmodule.	Zugelassen zur Förster*innenausbildung werden Kandidat*innen mit einem eidg. Fähigkeitszeugnis als Forstwart*in (oder einer gleichwertigen Ausbildung). Erforderlich sind zudem 12 Monate Praxis in einem Forstbetrieb oder -unternehmen. Zudem müssen Interessent*innen die nachfolgenden Grundlagenmodule und die Eignungsprüfung absolviert und bestanden haben. Details dazu sind erhältlich bei den Anbieter*innen: BZW Lyss und ibW BZW Maienfeld. Angehende Forstwart*in-Vorarbeiter*innen und Förster*innen besuchen die gleichen Grundlagenmodule.
P2	1. Einführung Durch die Technik des Internets und der elektronischen Datenverarbeitung kann der Einzelne das Gefühl bekommen, den Überblick darüber zu verlieren, wo und zu welchem Zweck seine Daten gespeichert werden. Gerade im finanziellen Bereich ist das Vertrauen in die sorgfältige und sichere Behandlung von Kundendaten besonders wichtig. Deshalb möchten wir Ihnen als Besucher unserer Web-Seiten erläutern, wie die Unternehmen der R+V Versicherungsgruppe die Vertraulichkeit Ihrer personenbezogenen Daten sicherstellt und die Persönlichkeitsrechte respektiert.	1. Einführung Durch die Technik des Internets und der elektronischen Datenverarbeitung kann der Einzelne das Gefühl bekommen, den Überblick darüber zu verlieren, wo und zu welchem Zweck seine Daten gespeichert werden. Gerade im finanziellen Bereich ist das Vertrauen in die sorgfältige und sichere Behandlung von Kund*innendaten besonders wichtig. Deshalb möchten wir Ihnen als Besucher*in unserer Web-Seiten erläutern, wie die Unternehmen der R+V Versicherungsgruppe die Vertraulichkeit Ihrer personenbezogenen Daten sicherstellt und die Persönlichkeitsrechte respektiert.	1. Einführung Durch die Technik des Internets und der elektronischen Datenverarbeitung kann der*die Einzelne das Gefühl bekommen, den Überblick darüber zu verlieren, wo und zu welchem Zweck seine*ihre Daten gespeichert werden. Gerade im finanziellen Bereich ist das Vertrauen in die sorgfältige und sichere Behandlung von Kund*innendaten besonders wichtig. Deshalb möchten wir Ihnen als Besucher*in unserer Web-Seiten erläutern, wie die Unternehmen der R+V Versicherungsgruppe die Vertraulichkeit Ihrer personenbezogenen Daten sicherstellt und die Persönlichkeitsrechte respektiert.
P3	Hat er keine passende Karte ist der nächste Spieler an der Reihe. Wer die vorletzte Karte ablegt, muss „UNO!“ (das bedeutet „Eins“) rufen und signalisiert damit, dass er nur noch eine Karte auf der Hand hat. Vergisst ein Spieler das und ein anderer bekommt es rechtzeitig mit (bevor der nächste Spieler eine Karte gezogen oder abgeworfen hat) so muss er 2 Strafkarten ziehen. Die Runde gewinnt derjenige, welcher die letzte Karte abgelegt hat. Die Punkte werden addiert und eine neue Runde wird gespielt.	Hat er keine passende Karte ist der nächste Spieler*in an der Reihe. Wer die vorletzte Karte ablegt, muss „UNO!“ (das bedeutet „Eins“) rufen und signalisiert damit, dass er nur noch eine Karte auf der Hand hat. Vergisst ein Spieler das und ein anderer bekommt es rechtzeitig mit (bevor der nächste Spieler*innen eine Karte gezogen oder abgeworfen hat) so muss er 2 Strafkarten ziehen. Die Runde gewinnt derjenige, welcher die letzte Karte abgelegt hat. Die Punkte werden addiert und eine neue Runde wird gespielt.	Hat er*sie keine passende Karte ist der*die nächste Spieler*in an der Reihe. Wer die vorletzte Karte ablegt, muss „UNO!“ (das bedeutet „Eins“) rufen und signalisiert damit, dass er*sie nur noch eine Karte auf der Hand hat. Vergisst ein*e Spieler*in das und ein*e andere*r bekommt es rechtzeitig mit (bevor der*die nächste Spieler*in eine Karte gezogen oder abgeworfen hat) so muss er*sie 2 Strafkarten ziehen. Die Runde gewinnt der*diejenige, welche*r die letzte Karte abgelegt hat. Die Punkte werden addiert und eine neue Runde wird gespielt.
P4	Erst Ende März war Miller bereits negativ aufgefallen. Ebenfalls auf Hawaii randalierte er in einer Karaoke-Bar in Honolulu derartig, dass die Polizei einschreiten musste und den Star wegen Ruhestörung und Belästigung festnahm. Er habe Obszönitäten geschrien und versucht, einer 23-jährigen Besucherin das Mikrofon aus der Hand zu reißen. Dies stellte laut Polizei eine Ordnungswidrigkeit dar. Später griff er einen 32 Jahre alten Mann an, der Darts spielte. Dies erfülle den Tatbestand der Belästigung. Seit 2016 verkörpert Miller, der sich als non-binäre Person identifiziert, im DC Extended Universe den blitzschnellen Superhelden The Flash.	Erst Ende März war Miller bereits negativ aufgefallen. Ebenfalls auf Hawaii randalierte er in einer Karaoke-Bar in Honolulu derartig, dass die Polizei einschreiten musste und den Star wegen Ruhestörung und Belästigung festnahm. Er habe Obszönitäten geschrien und versucht, einer 23-jährigen Besucher*in das Mikrofon aus der Hand zu reißen. Dies stellte laut Polizei eine Ordnungswidrigkeit dar. Später griff er*sie einen 32 Jahre alten Mann an, der Darts spielte. Dies erfülle den Tatbestand der Belästigung. Seit 2016 verkörpert Miller, der sich als non-binäre Person identifiziert, im DC Extended Universe den blitzschnellen Superhelden The Flash.	Erst Ende März war Miller bereits negativ aufgefallen. Ebenfalls auf Hawaii randalierte er*sie in einer Karaoke-Bar in Honolulu derartig, dass die Polizei einschreiten musste und den Star wegen Ruhestörung und Belästigung festnahm. Er*sie habe Obszönitäten geschrien und versucht, einem*r 23-jährigen Besucher*in das Mikrofon aus der Hand zu reißen. Dies stellte laut Polizei eine Ordnungswidrigkeit dar. Später griff er*sie einen 32 Jahre alten Mann an, der Darts spielte. Dies erfülle den Tatbestand der Belästigung. Seit 2016 verkörpert Miller, der*die sich als non-binäre Person identifiziert, im DC Extended Universe den*die blitzschnelle*n Superheld*in The Flash.
P5	Klassische Maßnahmen zur Kundenbindung 1. Kundenclub Durch eine Kundenkarte bekommen Käufer zum Beispiel Prozente oder andere Vorzüge, während Sie als Unternehmer die Daten erhalten. Dadurch sind Sie in der Lage, Kontakt mit ihm aufzunehmen. Es gibt kostenlose Kundenkarten und kostenpflichtige, die einen höheren Rabatt geben.	Klassische Maßnahmen zur Kundenbindung 1. Kundenclub Durch eine Kundenkarte bekommen Käufer*innen zum Beispiel Prozente oder andere Vorzüge, während Sie als Unternehmer*in die Daten erhalten. Dadurch sind Sie in der Lage, Kontakt mit ihm*ihr aufzunehmen. Es gibt kostenlose Kundenkarten und kostenpflichtige, die einen höheren Rabatt geben.	Klassische Maßnahmen zur Kund*innenbindung 1. Kund*innenclub Durch eine Kund*innenkarte bekommen Käufer*innen zum Beispiel Prozente oder andere Vorzüge, während Sie als Unternehmer*in die Daten erhalten. Dadurch sind Sie in der Lage, Kontakt mit ihm*ihr aufzunehmen. Es gibt kostenlose Kund*innenkarten und kostenpflichtige, die einen höheren Rabatt geben.
P6	Es ist wichtig, regelmäßig Ihre Brüste untersuchen zu lassen, und Sie sollten Ihren Arzt aufsuchen, wenn Sie einen Knoten fühlen. In seltenen Fällen wurden gutartige Lebertumore und noch seltener bösartige Lebertumore bei Anwenderinnen von KOKs berichtet. Suchen Sie Ihren Arzt auf, wenn Sie ungewöhnlich starke Bauchschmerzen haben. Gebärmutterhalskrebs wurde bei Langzeitanwenderinnen beobachtet; aber es ist nicht geklärt, in wie weit unterschiedliches Sexualverhalten oder andere Faktoren wie das humane Papilloma-Virus (HPV) dazu beitragen.	Es ist wichtig, regelmäßig Ihre Brüste untersuchen zu lassen, und Sie sollten Ihren Arzt aufsuchen, wenn Sie einen Knoten fühlen. In seltenen Fällen wurden gutartige Lebertumore und noch seltener bösartige Lebertumore bei Anwender*innen von KOKs berichtet. Suchen Sie Ihren Arzt auf, wenn Sie ungewöhnlich starke Bauchschmerzen haben. Gebärmutterhalskrebs wurde bei Langzeitanwender*innen beobachtet; aber es ist nicht geklärt, in wie weit unterschiedliches Sexualverhalten oder andere Faktoren wie das humane Papilloma-Virus (HPV) dazu beitragen.	Es ist wichtig, regelmäßig Ihre Brüste untersuchen zu lassen, und Sie sollten Ihre*n Ärzt*in aufsuchen, wenn Sie einen Knoten fühlen. In seltenen Fällen wurden gutartige Lebertumore und noch seltener bösartige Lebertumore bei Anwender*innen von KOKs berichtet. Suchen Sie Ihre*n Ärzt*in auf, wenn Sie ungewöhnlich starke Bauchschmerzen haben. Gebärmutterhalskrebs wurde bei Langzeitanwender*innen beobachtet; aber es ist nicht geklärt, in wie weit unterschiedliches Sexualverhalten oder andere Faktoren wie das humane Papilloma-Virus (HPV) dazu beitragen.

Table 6: Different versions of six text excerpts used in our survey. Changes compared to the original text are marked in orange.

- The models are primarily trained for research purposes, showing that gender-fair rewriting models can be trained without language-specific handwritten rules.
- While our survey with potential beneficiaries highlights that gender-fair rewriting models – even though not error-free – may also be beneficial in real-world applications, we caution that they should be thoroughly tested by potential users before being deployed outside of research contexts.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 4.4, Section 5, Limitations

☑ A2. Did you discuss any potential risks of your work?
Section 5, Limitations, Ethics Section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract, Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3, Section 4, Appendix A, Appendix B, Appendix C, Appendix D, Appendix F

☑ B1. Did you cite the creators of artifacts you used?
Section 3.2, Section 3.3, Section 4.2, Section 4.3, Appendix A, Appendix B, Appendix C, Appendix D, Appendix G

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Supplementary materials

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 5, Appendix I

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The data collected in our survey is completely anonymous (Section 4.4) and we did not ask for information that would allow identifying individual participants. See the discussion in the Ethics Section regarding filtering offensive content in the publicly available data that was used to train the rewriting models.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3.2, Section 3.3, Section 4.2, Section 4.3, Section 4.4, Limitations

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3.2, Section 4.2, Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Section 3, Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3.2, Section 4.2

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3.2, Section 4.2, we did not run hyperparamter search

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Table 1, Table 2, Table 3, Figure 3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3.2, Section 3.3, Section 4.2, Section 4.3, Appendix A, Appendix B

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 4.4

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 4.4, Limitations, Ethics Statement, Appendix G, Appendix H

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 4.4, Limitations, Appendix H

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 4.4, Limitations, Appendix H

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Self-assessment with the University

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 4.4