Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)

Do gender neutral affixes naturally reduce gender bias in static word embeddings?

Jonas Wagner and **Sina Zarrieß**
Bielefeld University
Faculty for Linguistics and Literary Studies
{jonas.wagner,sina.zarriess}@uni-bielefeld.de

Abstract

In German, substituting gendered role nouns with gender neutral versions, known as *gendergerechte Sprache*, has rapidly been gaining ground, with the primary aim being the inclusion of non-male people. Its effectiveness, however, has not been conclusively demonstrated. Previously, word embeddings have been shown to contain gender biases similar to natural language. They thus can be used to measure whether this practice impacts gender association of role nouns. Methods of debiasing pretrained word embeddings have been devised, but their effectiveness in German, especially compared to *gendergerechte Sprache*, has not been tested. In this paper, we systematically compare two methods of gender neutral affixation to a base corpus to examine the effect on gender bias of role nouns. We also compare the gender biases of analogy resolutions generated with embeddings trained on the base corpus, on the base corpus after undergoing an established post-hoc debiasing method, and the corpus after introduction of gender neutral affixation. Our results show a mixed picture: affixation leads to increased gender bias of role nouns, but decreased gender bias of generated analogy resolutions, even outperforming post-hoc debiasing methods.

1 Introduction

Gender bias in word embeddings and its reduction have received significant attention from computational linguists and NLP researchers over the past years, and a substantial body of research around the topic has accumulated (Bolukbasi et al. 2016; Caliskan et al. 2017; Ethayarajh et al. 2019; Kaneko and Bollegala 2019 among others). Given the wide use of word embeddings and the resulting danger of perpetuating and reinforcing gender stereotypes (Hansen et al., 2015; Musto et al., 2015; Dastin, 2018; Schnitzer et al., 2019), this is a pressing concern. But existing research has failed to address two aspects of the issue: firstly, as is a common problem in NLP, it mostly investigates English (but see Sahlgren and Olsson 2019 and Katsarou et al. 2022 for investigations of Swedish, Chávez Mulsa and Spanakis 2020 for Dutch, and Basta et al. 2020 for Spanish), which, in contrast to German, does not have regular gender marking on nouns.

Secondly, and possibly as a result of this, it ignores societal efforts to mitigate gender bias in natural language. Blodgett et al. (2020, p. 5458) criticise this detachment from such societal processes, instead calling for researchers to "[e]xamine language use in practice by engaging with the lived experiences of members of communities affected by NLP systems". One way in which language users are addressing gender biases in their languages is by changing these gender markings, such as the *-e* suffix in Spanish or the addition of the female suffix *-in* to German role nouns,[1] which is the subject of the present study. Instead, research has focused on post-hoc debiasing of pre-trained word embeddings rather than the impact of these societal processes.

The practice of adding the female role noun affix *-in* to male role nouns in German is known is *gendergerechte Sprache* (henceforth *GGS*). *GGS* has become a controversial topic in Germany (Stöber, 2021), which may (at least partially) be rooted in the fact that a quantitative investigation into its effectiveness has not yet been conducted. While this paper sets out to begin an investigation into quantifiable gender bias reduction through *GGS*, due to the complex nature of the subject and its ideological components, the question whether it measurably reduces gender bias may not be answerable, especially in the short term. Nevertheless, given the tools supplied with word embeddings, an initial

[1] For the purposes of this work, "role noun" refers to nouns that denote someone's activity or occupation, such as *runner*, *teacher*, or *listener*.

investigation is warranted and valuable. We thus investigate two research questions:

RQ1: Does gender neutral language in German lead to a reduction in gender association of role nouns' embeddings?

RQ2: Is altering corpora on which embeddings are trained so as to make their language more gender neutral as effective as post-hoc debiasing of word embeddings?

To answer these questions, we conduct two experiments. First, we train word embeddings on a corpus of German language texts and measure the gender association of role nouns in the text before and after altering them to conform to *GGS* (Section 3.4). Second, we compare the reduction of gender association of this to hard-debias (Bolukbasi et al. 2016; see also Section 2.2) to gauge its effectiveness (Section 3.5). Although it is not its focus, this research will also contribute to the growing body of research of gender bias in word embeddings in non-English languages.

2 Background

2.1 *Gendergerechte Sprache*: gender neutral language in German

German, like many Indo-Eurpean languages, has grammatical gender with a regular derivational pattern for role noun generation. For example, *Programmierer* means "male programmer", while *Programmiererin* means "female programmer"; *-er* serves as a derivational morpheme with which male role nouns can be generated from verbs (*programmieren*, "to program"), and *-in* changes male to female role nouns. Gender neutral alternatives to these gendered suffixes do not exist.

Masculine generics have, therefore, been used to refer to not only male individuals in occupations, but all individuals – *Programmierer* could refer to male as well as female and non-binary programmers, despite being morphologically masculine. Criticism of this practice goes back several decades (see Braun et al., 2005 and Kotthoff, 2020 for an overview), but has been mounting in recent years. This has led to the establishment of more formalised ways of explicitly including none-male people in generic role nouns (Kotthoff, 2020). These largely add the female suffix *-in*, separated by a typographic symbol such as * (see Table 1 for an example).

2.2 Gender bias in word embeddings

Given the wide use of word embeddings in downstream tasks, the mitigation of gender biases present in them has been of interest to researchers. This necessitates a method to measure gender bias in word embeddings first, which Ethayarajh et al. (2019) provide with the *Relational Inner Product Association* (RIPA). This method identifies the vector $\overrightarrow{b}$, which captures the subspace of the embedding space that denotes gender. This is done by first creating a set (*S*) of pairs of words that define the gender association. The two words in each pair only differ by gender, but the relationship between the pairs can be arbitrary. An example for *S* would be ($\{woman, man\}$, $\{queen, king\}$, $\{girl, boy\}$). Of these pairs, the difference vectors ($\overrightarrow{woman}$–$\overrightarrow{man}$, $\overrightarrow{queen}$–$\overrightarrow{king}$, etc.) are taken, and the first principal component of all difference vectors is computed. This first principal component is $\overrightarrow{b}$, and a word's gender association is simply the dot product of its embedding and $\overrightarrow{b}$. RIPA is highly interpretable: if, as in the example, the first word in each pair in the set *S* is the female word, positive RIPA scores show female association and negative scores show male association. The strength of the association is reflected by the absolute value of the score.[2]

Once a gender subspace is captured, debiasing can proceed. Bolukbasi et al. (2016) establish several methods, of which only hard-debias will be discussed here. To hard-debias an embedding, it is re-embedded with the following formula:

$$\vec{w} := \frac{\vec{w} - \vec{w}_B}{||\vec{w} - \vec{w}_B||}$$

Where $\overrightarrow{w}$ is the word's embedding and $\overrightarrow{w}_B$ is the embedding's projection on the gender subspace - in our case, this subspace is $\overrightarrow{b}$ as introduced above. Vectors enclosed in || denote the vectors' norms.

Investigations of gender bias in contextualised embeddings are emerging, but still less well-researched than static embeddings. However, it has been shown that despite their sensitivity to context, gender bias is still present in contextualised embeddings, especially for occupations (Basta et al., 2020), though less pronounced than in static embeddings (Sahlgren and Olsson, 2019). Established debiasing methods may not mitigate gender bias in contextualised embeddings well (Sahlgren and Ols-

[2] For a more in-depth discussion of RIPA and other bias measurements, see Ethayarajh et al., 2019 and Caliskan et al., 2017.

Original sentence (male role noun)	*Der*	*Programmierer*		*schläft*
Original sentence (female role noun)	*Die*	*Programmiererin*		*schläft*
New sentence after affixation	<[ART]>	*Programmierer*in*		*schläft*
New sentence after inserting the *in-token	<[ART]>	*Programmierx*	<*in>	*schläft*
Translation	The	*programmer*		sleeps

Table 1: Example of a sentence that was changed with both methods.

son, 2019). Translingual research has also revealed that in Swedish, occupations are less gender biased than in English (Katsarou et al., 2022).

Post-hoc debiasing methods like hard-debias have the advantage of being employable on large pre-trained models, thus circumventing the need to gather large corpora of gender neutral language to train new embeddings. However, they rely on several assumptions. Most importantly, for post-hoc debiasing to be at all effective, it is crucial that the gender subspace with regards to which words are debiased accurately captures gender. But, as Ethayarajh et al. (2019) point out, the selection of words that define the gender subspace is arbitrary and subject to beliefs and biases of those who conduct the debiasing, even with the more robust RIPA. Additionally, they crucially ignore the contextual and societal aspects of language. Language users are already implementing their idea of gender neutral language, but this type of language, which is desired by its users, may not be reflected in the corpora that word embeddings are trained on. Post-hoc debiased word embeddings therefore do not reflect natural gender neutral language, but a computationally altered version of gender biased language. Given that gender bias is, at its core, a societal and cultural phenomenon, this is a serious shortcoming which the present study aims to investigate. For the purposes of this study, we will refer to these natural-language-like debiasing methods as *corpus debiasing*, and to post-hoc debiasing methods like hard-debias as *embedding debiasing*.

3 Experiments

3.1 Data

We use the *Gebrauchsliteratur* subset of the German-language fiction corpus (henceforth *DTA-Gebrauchsliteratur*; available at `https://www.deutschestextarchiv.de/download`) on works from 1750 onwards, totalling some 120 books. Using the CBOW implementation in *Word2Vec* from *gensim* (Řehůřek and Sojka, 2010, version 4.1.2), we train embeddings on this corpus with a vector size of 50 (due to the comparatively small size of the corpus) and a window size of 10.

3.2 Role Nouns

We extract role nouns from the corpus by filtering out capitalised words (as all German nouns are capitalised) that end in *-er* or *-erin* (see Section 2 for information on the morphology of German role nouns). Using *spaCy* (Honnibal and Johnson, 2015), we then filter this list of nouns twice: the first step removes all plural nouns, as *-er* is also a standard plural morpheme for German nouns – not just role nouns – resulting in many false positives. This is filtered again, allowing only entries that were clearly derived from verbs. For this, we remove the role noun suffixes (*-er* and *-erin*) and replace them with *-en*, the default ending for German non-finite verbs. Only if *spaCy* recognises this as a verb is the noun retained in the list. We then manually investigate this final list and remove any false positives. False negatives, however, cannot be added back in. In total, the list includes 764 role nouns: 636 male and 128 female, with 71 of them occurring in both the male and female forms.

3.3 Affixation patterns

Then, we alter the role nouns in the corpus in two ways:

Affixation: Substituting each role noun with a version of itself with the role noun endings removed and *-er*in* appended (both *Programmierer* and *Programmiererin* become *Programmierer*in*)

Inserting an *in-token: Substituting each role noun with a version of itself with the role noun endings removed and *-x* appended (both *Programmierer* and *Programmiererin* become *Programmierx*) and inserting **in* as an additional token **after** every role noun.

In both cases, we replace any determiner preceding the role noun with the token *[ART]* (from German *Artikel*, "determiner"). See Table 1 for an example. We then train embeddings from scratch on both altered versions of the corpus with the same hyperparameters as above.

The reason for inserting **in* as a token after the role nouns is that simple affixation (i.e. exchanging all instances of role nouns with gender neutral versions of themselves) should necessarily lead to a reduction in gender association, provided the male and female versions have different gender associations. If, in a hypothetical corpus, the words *Programmierer* and *Programmiererin* are of equal frequency and the former is male-associated while the latter is female-associated, the new version (which would substitute both in the entire text) would have the mean gender assocation of the two, i.e. it would lie somewhere in between them. This would reduce measurable gender bias, but would likely not work in cases where the two versions' frequencies are unequal or one does not occur at all. Introducing the new token **in* while also changing all role nouns to a gender neutral version allows the gender neutralising effect that *GGS* has on role nouns where both versions occur to carry over to those of which only either the male or the female version occurs – though potentially not as strong – as all role nouns now occur in the vicinity of the **in*-token. The validity of this approach will be tested in this experiment.

Note that if sub-word embeddings had been learned (using e.g. fastText, Bojanowski et al., 2017), this approach may not have been necessary in cases where the role noun would be recognised as consisting of a verb (e.g. *programmier-*) and the derivational affixes (*er* and **in*, respectively). However, gender association and bias are much more well-researched in word embeddings generated with *Word2Vec*, making it the preferred approach here.

3.4 Experiment 1: Impact of *gendergerechte Sprache* on gender association

We calculate the RIPA score (Ethayarajh et al., 2019) of the role nouns we extracted in the base corpus and of the altered role nouns in the corpus-debiased corpora (see Sections 3.2 and 3.3). For the gender defining set S we use kinship terms (see Table 2).

Shapiro tests from *scipy.stats* (version 1.6.2, Virtanen et al., 2020) show that RIPA scores are not normally distributed. Thus, we use two-sided Wilcoxon tests (from the same package) for significance testing. We run separate tests for each gender. We use the *median* function from *statistics* to calculate medians, and create boxplots with *pyplot* from *matplotlib* (Hunter, 2007, version 3.3.4). Since multiple tests were run, we Bonferroni adjust p-values with *multipletests* from *statsmodels* (Seabold and Perktold, 2010, version 0.12.2).

3.5 Experiment 2: Analogy resolution

We debias the role nouns' embeddings from the base corpus (Sections 3.1 and 3.2) using hard-debias (Bolukbasi et al. 2016; see Section 2.2). It is not possible to evaluate the resulting gender associations with RIPA, since hard-debias reduces gender association w.r.t. RIPA – that is, the RIPA scores of words after undergoing hard-debias are necessarily minimal.

Instead, we alter the methodology used by Bolukbasi et al. (2016), who generate analogy resolutions for each investigated word, e.g. "he is to doctor as she is to X", with the analogy being solved for X. In Bolukbasi et al. (2016), crowd-workers then rate whether the analogy resolution is biased (e.g. *nurse*) or not (e.g. *physician*). This works well for their research, but is expensive and time-consuming. There are also other reasons why it would not work for our experiment:

- **Ambiguity of German nouns and pronouns**. *Sie* is the third person singular female pronoun, but also the gender neutral third person plural pronoun, and, if capitalised, the second person honorific pronoun. *Frau* ("woman") also is a honorific for women ("Mrs"), so they do not differ only by gender. This means that the analogy "er verhält sich zu Arzt wie sie zu X" ("he is to doctor as she is to X") would not necessarily have a gendered resolution, as *sie* does not refer strictly to female individuals. The analogy therefore cannot be constructed using pronouns nor words for *man* and *woman*
- **Loss of natural language gender bias**. The corpus-based gender bias reduction methods introduced in Section 3.3 lead to analogies like "man is to male or female doctor as woman is to male or female nurse". Human raters would rate these as gender neutral, as they employ the gender neutral suffixes that they are used to from natural language

To solve the first issue, we calculate the mean embeddings of the male and female words in S (see

German kinship terms	English translation
Frau, *Mann*	woman, man
Schwester, *Bruder*	sister, brother
Tante, *Onkel*	aunt, uncle
Tochter, *Sohn*	daughter, son
weiblich, *männlich*	female, male
Cousine, *Cousin*	female cousin, male cousin
Nichte, *Neffe*	niece, nephew
Enkelin, *Enkel*	granddaughter, grandson
Schwägerin, *Schwager*	sister-in-law, brother-in-law

Table 2: Set S that defines the gender association.

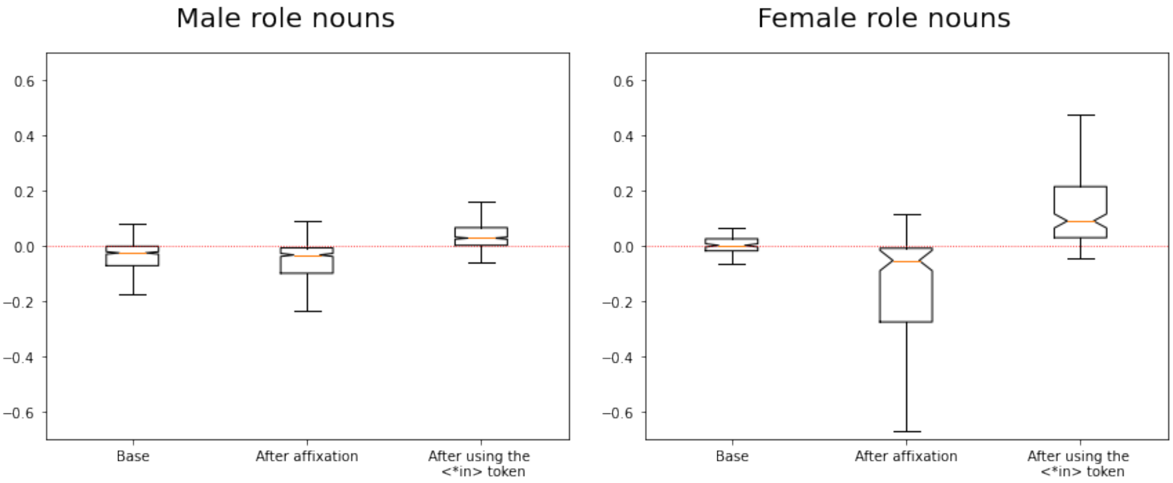

Figure 1: Boxplot of gender associations of role nouns: RIPA scores of unaltered role nouns and after undergoing gender association reduction. Outliers omitted. Whiskers end at 1.5*IQD. Positive scores indicate female, negative scores male association.

Table 2) for each corpus and insert them into their respective embedding spaces, thus getting a better measure of gender than using only a pronoun. Then, we generate ten analogy resolutions per role noun and compute the mean RIPA score for them. The hard-debiased data, however, still poses a problem here. Since, in a good model, role nouns should be generated for the analogy resolutions, we would encounter the same problem as above: the analogy might still be solved as e.g. "man is to doctor as woman is to nurse", only that both *doctor* and *nurse* would have been debiased w.r.t. RIPA. Thus, the model could generate a clearly biased resolution that would still have a low RIPA score.

We generate the analogy resolution in the hard-debiased embedding space and then take the RIPA score of the generated resolutions in the base, non-debiased space. The analogy ($\mu_{male_{HD}}$ *is to male doctor*$_{HD}$ *as* $\mu_{female_{HD}}$ *is to* X_{HD}), where *HD* denotes the hard-debiased embedding space and μ_{male} and μ_{female} are the mean male and female embeddings described above, is solved for X_{HD}. Then, we compute the RIPA scores not of X_{HD}, but of X_{base} in the base corpus. This means that analogy resolutions that would still be perceived as biased by human raters ("he is to doctor as she is to nurse") will be recognised as such. This is not possible for the corpus-debiased embeddings, as role nouns generated in those models have no gender markings on them, meaning it would be impossible to decide whether to calculate the RIPA score of the male or female version in the base corpus. Their RIPA scores are thus computed in their own embedding spaces. We also calculate how many nouns and role nouns are generated for each analogy as an indicator of the quality of the resolutions.

Comparison	RIPA scores 1st median	 2nd median	 median change (abs.)	p	p_{adj}	signif.
Male role nouns						
base vs affixation	-0.0249	-0.0321	-0.0072	1.24E-16	2.23E-15	···
base vs **in*-token	-0.0249	0.0286	-0.0037	2.57E-91	4.63E-90	···
affixation vs **in*-token	-0.0321	0.0286	0.0035	1.95E-96	3.51E-95	···
Female role nouns						
base vs affixation	0.0023	-0.0524	-0.0501	2.27E-16	4.08E-15	···
base vs **in*-token	0.0023	0.0907	-0.0885	1.34E-17	2.41E-16	···
affixation vs **in*-token	-0.0524	0.0907	-0.0384	1.53E-22	2.74E-21	···

Table 3: Gender association of role nouns before and after corpus debiasing, separated by gender. Significance codes: · (p <.05), ·· (p <.01), · · · (p <.001); codes also apply to other tables.

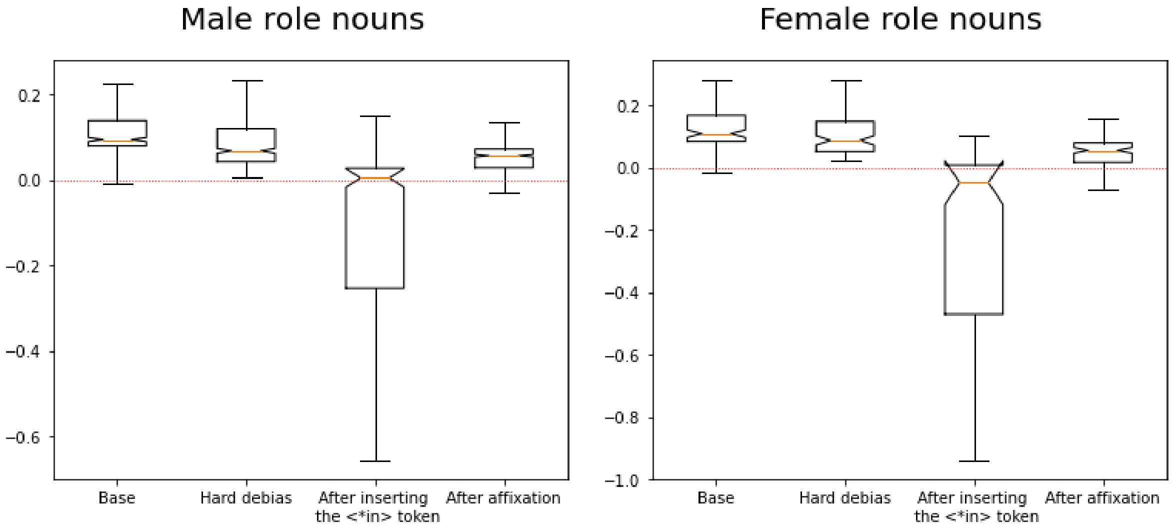

Figure 2: Boxplot of gender associations of role nouns: RIPA scores of unaltered role nouns and after undergoing gender association reduction. Outliers omitted. Whiskers end at 1.5*IQD. Positive scores indicate female, negative scores male association.

4 Results

In Tables 3 and 4, positive RIPA scores show female association and negative RIPA scores show male association. The absolute value of the scores shows the strength of the association. The first median refers to the median RIPA score of the first part of the comparison (e.g. role nouns in the base corpus in the comparison *base vs affixation*), the second median to the second one (e.g. role nouns after undergoing affixation in that same comparison). The higher the absolute value of the RIPA score, the stronger the association. Negative median changes indicate that gender association is stronger in the second part of the comparison, positive ones indicate that it is weaker.

4.1 Experiment 1

The mean absolute RIPA scores for male role nouns in the base corpus (-0.0249, male biased) are lower than after affixation (-0.0321; p_{adj}<0.001) and inserting the **in*-token (0.0286, female biased; p_{adj}<0.001; see Table 3). The difference between both debiasing methods is significant (p_{adj}<0.001).

For female role nouns, mean absolute RIPA scores are lower in the base corpus (0.0023) than after affixation (-0.0524; p_{adj}<0.001) and inserting the **in*-token (0.0907; p_{adj}<0.001 see Table 3). The difference between both debiasing methods is significant (p_{adj}<0.001).

4.2 Experiment 2

For male role nouns, mean absolute RIPA scores of analogy resolutions in the base corpus (0.0935, fe-

		RIPA scores				
Comparison	1st median	2nd median	median change (abs.)	p	p_{adj}	signif.
Male role nouns						
base vs hard-debias	0.0935	0.0670	0.0264	6.55E-10	1.18E-08	···
base vs affix	0.0935	0.0555	0.0380	5.10E-34	9.19E-33	···
base vs **in*-token	0.0935	0.0037	0.0898	3.86E-65	6.95E-64	···
hard-debias vs affix	0.0670	0.0555	0.0116	3.16E-12	5.69E-11	···
hard-debias vs **in*-token	0.0670	0.0037	0.0633	1.67E-61	3.01E-60	···
**in*-token vs affix	0.0580	0.0572	0.0008	1.14E-13	5.36E-12	···
Female role nouns						
base vs hard-debias	0.1078	0.0866	0.0212	1.63E-06	2.94E-05	···
base vs affix	0.1078	0.0545	0.0533	3.37E-17	6.08E-16	···
base vs **in*-token	0.1078	-0.0486	0.0592	2.95E-20	5.32E-19	···
hard-debias vs affix	0.0866	0.0545	0.0321	2.39E-09	4.30E-08	···
hard-debias vs **in*-token	0.0866	-0.0486	0.0380	1.06E-19	1.90E-18	···
**in*-token vs affix	0.2254	0.0584	0.1670	3.16E-12	1.49E-10	···

Table 4: Gender association of role noun: base, after hard-debias, after affixation, after adding the **in*-token, separated by gender. Significance codes: · (p <.05), ·· (p <.01), ··· (p <.001).

Model	Word	Generated analogy resolutions
Base	Erzieherin	*Erzieherin*, *Lehrerin*, zieherin, *Gesellschafterin*, *Dichterin*
Hard-debiased	Erzieherin	*Erzieherin*, *Dichterin*, Zahl, Cuvier'schen, Aufbewahrung
Affixation	Erzieher*in	*Erzieher*in*, *Lehrer*in*, *Beamter*, *Gesellschafterin*, *Buchhalter*
**in*-token	Erzieherx	*Erziehx*, *Pflegx*, *Leitx*, *Schülx*, *Verwaltx*
Base	Maler	*Maler*, Tieck, verwandt, *Kaufmann*, Nadelbäume
Hard-debiased	Maler	Freundschaft, Censoriade, vollkommnen, Freundin, geneigt
Affixation	Maler*in	*Maler*in*, *Sieger*in*, *Dichter*in*, entschiedener, *Musiker*
**in*-token	Malx	*Malx*, *Beschreibx*, *Kellnx*, *Porträtmalx*, *Nothelfx*

Table 5: Sample analogy resolutions from each model. Role nouns in resolutions in *italics*. First five resolutions per word.

male biased) are significantly higher than after hard-debias (0.0670; p_{adj}<0.001), affixation (0.0555; p_{adj}<0.001), or inserting the **in*-token (0.0037; p_{adj}<0.001). Analogies generated after affixation have significantly (p_{adj}<0.001) weaker gender association than those generated after hard-debias or inserting the **in*-token (see Table 4, Figure 2).

For female role nouns, mean absolute RIPA scores of analogy resolutions in the base corpus (0.1078; female biased) are significantly higher than after hard-debias (0.0866; p_{adj}<0.01), after affixation (0.0545; p_{adj}>0.99), or after inserting the **in*-token (-0.0486, male-biased; p_{adj}>0.99). Affixation leads to significantly (p_{adj}>0.99) lower gender association than hard-debias or inserting the **in*-token (see Table 4, Figure 2).

The base model generates a mean of 0.71 nouns and 0.27 role nouns per analogy resolution, the hard-debiased model 1.20 nouns and 0.85 role nouns, the model that we debiased by inserting the **in*-token 7.54 nouns and role nouns, and the model that we debiased by affixation generates a mean of 2.07 nouns and 1.86 role nouns per analogy resolution (see Table 5 for examples).

5 Interpretation

The experiments conducted in this research have demonstrated that *GGS*, i.e. the practice of substituting role nouns with gender neutral versions (e.g. turning *Programmierer* ("male programmer") and *Programmiererin* ("female programmer") into *Programmierer*in*) does not lead to a significant reduction in gender association of these words' embeddings. As can be seen in Figure 1, the two methods of implementing *GGS* in the corpus (see Section 3.4) lead to different results: affixation leads to an overall shift to male associations, while adding the *in*-token shifts gender association to female values. Affixation also leads to a greater spread of gender associations compared to adding the *in*-token, especially for female role nouns. While adding the *in*-token leads to overall greater absolute gender association for female role nouns, they are shifted towards female association, while they were already almost completely gender neutral in the base corpus. This indicates that *GGS* may simply shift gender associations towards the female end overall.

The second experiment (see Section 3.5) shows that hard-debias (see Bolukbasi et al., 2016) does work on languages with grammatical gender such as German, though it consistently performed worse than corpus debiasing. The results after adding the *in*-token also had a far greater spread than any other condition (see also Figure 2), while affixation had the smallest spread.

The corpus-debiased models also generate a far greater proportion of role nouns per analogy resolution, with the model that was debiased by inserting the *in*-token performing best in this regard. While this underlines their strong performance, it must be noted that this is not a fair comparison. We compute RIPA scores from the base model for the hard-debiased role nouns, while for the other two we compute them in their respective models. Role nouns, in the two altered corpora, occur in more similar environments, as they (and only they) are often preceded by the token *[ART]*, and in the case of the corpus that we debiased by inserting the *in*-token, all role nouns are always followed by the the *in*-token (see Table 1). This leads to role nouns' embeddings being more similar compared to other models, impacting the results of the experiment overall. This is a limitation of the methodology that we could not circumvent. However, the fact that all methods, but especially corpus debiasing, outperformed the base model by such a large margin is interesting, as it suggests that debiasing may lead to better analogy resolution performance, at least when it comes to role model analogies.

This may also be the reason for the seemingly contradictory results from both experiments: *GGS* leads to increased gender association in role nouns' embeddings, but reduced bias in analogy resolution. It appears that the analogy resolutions from the base corpus are fewer role nouns, but that those resolutions have stronger gender association than role nouns. The samples in Table 5 also (subjectively) appear qualitatively better to us: for example, *Maler*in* ("painter") is analogous to *Dichter*in* ("poet") and *Musiker*in* ("musician"), and *Malx* ("painter") is analogous to *Porträtmalx* ("portrait painter") after corpus debiasing, but not after other methods. In the base corpus, *Erzieherin* ("governess") is analogous to *Lehrerin* ("female teacher"), but also to *Dichterin* ("female poet") – the latter is not a good analogy, since the only relation appears to be gender.

The poor performance of the base model in analogy resolution, where it only managed to generate a noun in its top ten resolutions in 71% of cases, suggests that there may be issues with the data used, and a larger corpus (or one more tailored to role noun usage) may be necessary.

6 Conclusion and outlook

While *GGS* significantly increases gender association of role nouns, this does not necessarily invalidate the practice. Other than the ideological and philosophical questions that cannot be answered here, initial research on a smaller subset of this corpus yielded different results, where *GGS* significantly reduced gender association for male role nouns, but not female ones. This, once more, points to a weakness of this research: the results seem to depend on the corpus, and the corpus we use is rather small and may be too general, limiting the number of occurrences of role nouns even more. Further research with better suited corpora is necessary. Job postings, as one of the chief domains of *GGS*, would be particularly interesting data, but such corpora were not available. Use of larger or better suited corpora may also address the poor performance of the base model in analogy resolution and yield more informative data. Future research may also investigate if substituting gendered role nouns with gender neutral versions leads to the same results as balancing the occurrences of male and female versions of the role nouns.[3]

The research presented here has also demonstrated an additional weakness of existing debiasing methods, namely that their evaluation is very time consuming and usually involves crowdsourcing (such as in Bolukbasi et al., 2016). For German, no pre-made evaluation methods for hard-debias were available, and the method used here is far from perfect, as it evaluates analogies generated

[3] We would like to thank an anonymous reviewer for this suggestion.

from hard-debiased role nouns in the non-debiased model (to circumvent the problem where effectively, the same metrics to debias the role nouns are used to then measure their remaining bias), but for corpus-debiased embeddings, analogies are generated in their own models. This means that in reality, hard-debias may perform much better than the results in this research indicate. Nevertheless, it must be considered that debiasing methods that alter the data that embeddings are trained on perform comparably to hard-debiasing in this research.

Despite some weaknesses, this research demonstrates that natural language debiasing strategies are fundamentally different from post-hoc debiasing of pre-trained embeddings, and thus, the latter must be viewed with caution. It may still be used for practical purposes, but users must be aware that it is not analogous to societal efforts to reduce gender bias. These results are in line with Blodgett et al. (2020), who encourage researchers to more strongly relate their work to the experiences of real-world members of affected communities. While we do not directly engage with members of such affected communities, our findings that post-hoc debiasing is not equivalent to real-world natural language debiasing strategies lend further weight to their calls.

Lastly, we also partially address the question posed in (Bolukbasi et al., 2016) regarding the use of post-hoc word embeddings debiasing methods for language with grammatical gender. For the limited set of words investigated here, it does indeed lower gender association in analogy resolution. Future research with more sophisticated evaluation methodologies will shed more light on this area.

References

Christine Basta, Marta R. Costa-Jussà, and Noe Casas. 2020. Extensive study on the underlying gender bias in contextualized word embeddings. *Neural Computing and Applications*, 33(8):3371–3384.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. 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. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Friederike Braun, Sabine Sczesny, and Dagmar Stahlberg. 2005. Cognitive effects of masculine generics in German: An overview of empirical findings. *Communications: The European Journal of Communication Research*, 30(1):1–21.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Rodrigo Alejandro Chávez Mulsa and Gerasimos Spanakis. 2020. Evaluating bias in Dutch word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 56–71, Barcelona, Spain (Online). Association for Computational Linguistics.

Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

C Hansen, M Tosik, G Goossen, C Li, L Bayeva, F Berbain, and M Rotaru. 2015. How to get the best word vectors for resume parsing. In *SNN Adaptive Intelligence/Symposium: Machine Learning*.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Masahrio Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy. Association for Computational Linguistics.

Styliani Katsarou, Borja Rodríguez-Gálvez, and Jesse Shanahan. 2022. Measuring gender bias in contextualized embeddings. *Computer Sciences & Mathematics Forum*, 3(1).

Helga Kotthoff. 2020. Gender-Sternchen, Binnen-I oder generisches Maskulinum: (Akademische) Textstile der Personenreferenz als Registrierungen? *Linguistik Online*, 103(3):105–127.

Cataldo Musto, Giovanni Semeraro, Marco de Gemmis, and Pasquale Lops. 2015. Word embedding techniques for content-based recommender systems: An empirical evaluation. In *RecSys Posters*, volume 1441 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Magnus Sahlgren and Fredrik Olsson. 2019. Gender bias in pretrained Swedish embeddings. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 35–43, Turku, Finland. Linköping University Electronic Press.

Steffen Schnitzer, Dominik Reis, Wael Alkhatib, Christoph Rensing, and Ralf Steinmetz. 2019. Preselection of documents for personalized recommendations of job postings based on word embeddings. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, pages 1683–1686, New York, NY, USA. Association for Computing Machinery.

Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*.

Robert Stöber. 2021. Genderstern und Binnen-I. *Publizistik*, 66(1):11–20.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.