Is Gender Reference Gender-specific? Studies in a Polar Domain

Manfred Klenner, Dylan Massey
Department of Computational Linguistics, Andreasstrasse 15, 8050 Zürich
klenner@cl.uzh.ch, massey@ifi.uzh.ch

Abstract

We investigate how gender authorship influences polar, i.e. positive and negative gender reference. Given German-language newspaper texts where the full names of the authors are known and their gender can be inferred from the first names. And given that nouns in the text have gender reference, i.e. are labeled by a gender classifier as female or male denoting nouns. If these nouns carry a polar load, they count towards the gender-specific statistics we are interested in. A polar load is given either via phrase-level sentiment composition, or by a verb-based analysis of the polar role a noun (phrase) plays: is it framed by the verb as a positive or negative actor, or as receiving a positive or negative effect? Also, reported gender-gender relations (in favor, against) might be gender-specific. Statistical hypothesis testing is carried out in order to find out whether significant gender-wise correlations exist. We found that, in fact, gender reference is gender-specific: each gender significantly more often focuses on their own gender than the other one and e.g. positive actorship supremacy is claimed (intra-) gender-wise.

Keywords: gender-tailored text analysis, gender classification, sentiment inference

1. Introduction

Our research question is: Is there a correlation between the gender of the author of a text and the way gender denoting nouns are framed along the positive-negative axis? We infer the gender of the author from the first name given as part of the metadata of our corpus, a medium-sized German-language newspaper corpus. We also infer the gender of a noun in the text with a gender classifier trained on the basis of the grammatical gender of German human-denoting nouns.

Given gender tags for the author of a text and given all gender-denoting nouns in a text, we can investigate whether there is a gender-specific way of gender reference. Since we are dealing with newspapers where political events and their participants are being evaluated, a natural dimension to pursue is the positive and negative polarity of gender reference. In this paper we mainly use (and evaluate and partly improve) four freely available resources in order to identify and quantify the polar load of a reference: a German valence lexicon, a German polarity lexicon comprising 7,580 positive and negative words, a gender classifier for German, and a German sentiment inference system based on a verb resource where polar roles, polar effects, and polar relations between a source and target are specified for each verb.

With sentiment composition at the phrase level and by exploiting the valence lexicon, we determine gender-specific polar attribution like in *die herausragende Schauspielerin* (Eng. the extraordinary actress) and equivalent predicative constructions like *the actress is extraordinary*. Moreover, we take into account polar effects (positive, negative), polar actorship (positive, negative) and polar relations (in favour, against). In *He is cheating on her*, *to cheat* is the polar verb expressing an against relation between the referent of the male pronoun which is being understood as denoting a negative actor (the source) and the female pronoun which identifies the victim (the target) - we could say that a negative effect is cast on the target. Given these scenarios, we try to find out whether a gender-specific way of polar gender reference can be claimed. For instance, whether male authors refer to male positive actors significantly more often than female authors do? Or is gender reference not just gender-specific but even gender-centered, i.e. do genders pay significantly higher and possibly stronger attention to their own gender than cross-wise?

In order to put our claims on a sound statistical basis, we use traditional hypothesis testing: the (unpaired) t-test for independent samples.

The main contribution of this paper is the evaluation, fine-tuning and combination of existing resources for a new task: the investigation of gender-specific gender reference verified on the basis of statistical methods. The insights we gain are empirical (there is a correlation) and methodological (we describe the resources and methods needed). We believe that it is a substantial scientific advance to be able to reveal gender perspectives and make it available for evaluation.

2. Gender: Reference and Identity

In German, every noun has a grammatical gender: neutral, feminine (female) or masculine (male). A noun denoting a human being moreover has a gender reference. The word *Schwester* (Engl. sister) has a female grammatical gender and refers to a

LREC-COLING 2024, pages 9315–9324
20-25 May, 2024.

woman[1]. The same is true for first names: *Peter* is a male first name and has a male gender reference. For a long time in the German language, the male wordform of e.g. profession names was regarded as generic and gender inclusive. The masculine noun *Präsident* (Engl. president) then could be used to refer to all genders. Some decades ago, female wordforms were established and are now being used consistently in newspaper texts: *Präsidentin* is used with *in* as a suffix indicating female reference[2]. This still is a binary distinction. Recently, the gender star (*Präsident*in*) and other indicators of gender-inclusive reference like the colon (:) have been added. Still, traditional newspapers do not use it. As a consequence, we can only find and use binary references. We are aware of this limitation, but we cannot escape it. But certainly, we do not claim that gender is a binary category.

3. Newspaper Corpus

We have downloaded German-language newspaper texts from Swissdox[3], which is a media repository open for research purposes. We kept those texts where the metadata specified the full name (incl. the first name) of the author[4], so the gender reference of the author of each text is known. We looked into 4 newspapers (n_1 to n_4), altogether 2,993,094 articles, 2,200,389 written by male authors, and 792,705 by female authors.

One newspaper is a boulevard product with an unclear political orientation (n_2), one newspaper is left-leaning (n_4), one conservative (n_3) and one in-between (n_1). The data set comprises texts from the years 2019 to 2022. Table 1 shows the distribution of articles with a particular gender reference (i.e. authorship (AS)) and the percentage of unique names with a particular gender reference per medium (i.e. editorial membership (EM)).

	AS♂	AS♀	EM♂	EM♀
n1	75.10	24.90	70.99	29.01
n2	75.13	24.87	70.34	29.66
n3	72.60	27.40	61.14	38.86
n4	70.63	29.37	66.39	33.61

Table 1: Distribution of authorship (AS) and editorial membership (EM)

The authorship (AS) columns quantify how many articles are written by female (AS♀) or male (AS♂) authors while editorial membership (EM) refers to the percentage of unique names in media articles. The third row n_3 e.g. reveals that 72.60% of the articles have been written by male authors although only 61.14% of the authors of the newspaper n_3 are males. That is, male authors are producing more articles. This is true for all newspapers[5]. The overall editorial membership ratios are 64.71% (male) and 35.29% (female). The overall authorship ratios are 73.67% (male) and 26.33% (female). Altogether the corpus comprises 15,630 different authors.

We base our experiments on these 4 newspapers in order to see whether there are differences due to the political orientation of writers and whether local trends (at the level of a single newspaper) and global trends (all data points) converge or diverge.

4. German Valence Resources

If the polarity of a word is known, its valence can be used as polarity strength or polar load. The notion of valence can be traced back to the work of Osgood et al. (1957). Valence is related to the positive-negative connotation of a word. Low (or negative) values mean negative (e.g. evil), and high (positive) values mean positive connotations (e.g. good). Since the early work of Osgood and colleagues, a number of resources have been generated not only for English (Mohammad, 2018) but for other languages like German (Köper and Schulte im Walde, 2016) as well. Especially in psychology, such ratings have been created in a controlled way by human raters (Võ et al., 2009). Also crowd-sourcing has been used (Mohammad, 2018; Warriner et al., 2013). Manually created resources - as often - are small, e.g. Schmidtke et al. (2014) (1,000 words) and Võ et al. (2009) (2,900 words) which is the German version of the often cited English ANEW resource (Bradley and Lang, 1999). They are of limited direct usage, however they can be used to evaluate automatically generated versions of valence lexicons, e.g. by measuring the correlation.

Researchers starting with Turney et al. (2003) have tried to automatically infer lexicons on the basis of small seed lists. The very idea of Turney et al. (2003) was to use the seed list of known strong positive and negative words and to determine the strength values of new words on the basis of a similarity measure (they used PMI).

Köper and Schulte im Walde (2016) refined Turney et al. (2003) by using word embeddings instead of PMI and induced a lexicon comprising 351,617 German lemmas. They used existing manually created German resources as seed lexicons and parts of an automatically translated English resource

[1]There are only very few exceptions where the grammatical gender of a noun does not indicate sex, for instance: the neutral noun *Mädchen* (Engl. girl).

[2]Not all nouns with a female reference end with *in* and not all words with suffix *in* have a female reference.

[3]see: https://www.liri.uzh.ch/en/services/swissdox.html

[4]Only texts with a single author are kept.

[5]A possible explanation is that in Switzerland women are said to be more often part-time workers than men.

(Brysbaert et al., 2013). They trained a word2vec (Mikolov et al., 2013) German model for similarity determination and evaluated their approach with Pearson's correlation metric and achieved a result of 0.798.

Lüdtke and Hugentobler (2022) automatically created a large lexicon (933,814 inflected German wordforms) by applying the algorithm described in Turney et al. (2003). They used BAWL-R (Võ et al., 2009) to evaluate it and found a Pearson's correlation with the BAWL-R human ratings of 0.78.

With Köper and Schulte im Walde (2016) and Lüdtke and Hugentobler (2022) two large lexicons are available for German. Both have been evaluated wrt. human-labeled data. The results are very close. In section 6 we propose an additional evaluation possibility based on the German wordnet GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010) and a polarity lexicon.

Finally, there is a German version of Mohammad (2018) with about 20,000 entries (automatically translated). We do not consider it because the aforementioned ones are magnitudes larger.

Note that valence lexicons are not polarity lexicons. Whereas a polarity lexicon enumerates polar (positive and negative) words or word senses, a valence lexicon specifies valence strength values for neutral words as well. As a consequence, we only considered those words from the valence lexicon that at the same time are polar words, i.e. are in the polarity lexicon we used. The polarity of a word, thus, comes from the polarity lexicon, its polar load is determined from the valence lexicon.

5. Polarity Lexicon and Composition

For German, a couple of polarity lexicons are available. For an exhausting overview and evaluation see Fehle et al. (2021). For our experiments, we choose the updated version of our own lexicon Clematide and Klenner (2010)[6]. The lexicon was manually created, it comprises 7,580 entries, especially 4,150 adjectives. Each word is annotated for polarity (positive, negative) and its appraisal category (judgment, appreciation, emotion), see Martin and White (2005).

This lexicon forms the basis of the simple phrase-level sentiment composition we carried out. We take the majority vote on the basis of the word-level polarity. If a phrase is negated, we switch the polarity. This metric is sufficient since 97% of the phrases referring to humans just have a single polar word (either the noun is polar like in *the thief* or an adjective is as in *the cheating minister*).

[6]Download at: https://sites.google.com/site/iggsahome/downloads

6. Choosing the Best German Valence Lexicon

The lexicon performance in terms of correlation of Köper and Schulte im Walde (2016) and Lüdtke and Hugentobler (2022) are very close. How to choose among them? We could use the lexicon with the highest overlap with the polarity lexicon. However, it turned out that the overlap with both lexicons is comparable in size (with different subsets, though). One property of a sound valence lexicon might be that similar words do have similar valence values. Word embeddings could be used to measure this, but we can even think of a stronger version of the similarity criterion, namely synonymy. While the embedding space still is noisy, synsets taken from wordnets are not. Synonyms should have very close valence values to each other. Thus, the similarity of values within a synset seems to be an indicator of the goodness of the lexicon.

We used the German wordnet GermaNet (Hamp and Feldweg, 1997). Firstly, we generated 100 synsets for positive and 100 for negative words from the polarity lexicon. Then, we manually inspected the GermaNet synsets and evaluated whether the synset companions of the polar words preserve the polar load. For instance, the adjective *doof*'s (Eng. dumb) synset is *doof, blöd, dämlich* (Engl. dumb, stupid, silly). This clearly preserves the polar load. We manually inspected 200 synsets for this kind of consistency. Our evaluation showed a high preservation rate of 96%. Our criterion for choosing a lexicon, thus, seems to be valid.

As a statistical indicator of closeness, we took the standard deviation of the valence values within the synset. The mean synset lengths are 2.55 for Köper and Schulte im Walde (2016) and 2.6 for Lüdtke and Hugentobler (2022), so the variances actually must come from the value differences, not from differences in the number of values.

For each word pertaining to a GermaNet synset, we determined the standard deviation of the valence values of all synset members. We summed it up and took the mean. The smaller this mean standard deviation the better. For Köper and Schulte im Walde (2016) is was 0.038, for Lüdtke and Hugentobler (2022) is was 0.025. The latter is the better.

7. Verb-based Sentiment Inference

The intersection of the polarity and valence lexicon is meant for phrase-level analysis to answer the question: how (and how strongly) are gender denoting nouns referred to - in a positive or negative way. Besides such direct polar qualification and quantification, a gender denoting noun can also be cast or framed in a particular way as a positive or negative actor or as receiving a positive or negative

effect if it occupies a particular argument position of a polar verb.

For this, we used the output of our rule-based system described in Klenner et al. (2017). The system carries out sentiment inference, it assigns verb roles like positive or negative actors but also positive and negative relations. The system is rule-based using a verb lexicon with about 1,000 polar verb frames.For each verb the relation expressed between the source (most of the time the agent) and the target (patient, theme, or recipient) is specified. For instance, the verb *loben* (Eng. to honor) expresses a positive relation between the source and the target. Moreover, it is good to be honoured, so a positive perspectivation (effect) is expressed. Some verbs assign to the source a positive or negative actorship, e.g. the actor of *ermorden* (Eng. to murder) is negative. We use the output of our system inference system in our investigations.

8. Gender Classification for German

An essential part of our empirical investigation is gender classification. We need to know for each noun and (ideally) pronoun its gender reference. To the best of our knowledge, for German, our approach to gender classification is the only one, see Klenner (2023). The gold data[7] comprises lists of 5,885 female (*Schwester, Nonne, Professorin*, Eng. sister, nun, female professor), 5,020 male (*Bauer, Minister, Fußballer*, Engl. farmer, minister, soccer player) and 5,831 non-animacy denoting nouns (*Milch, Straße, Kaugummi*, Engl. milk, street, chewing gum). In Klenner (2023), we have used fastText embeddings (Joulin et al., 2017) to train a logistic regression classifier. The accuracy of a 75/25 split was 96%, F1 of the class *female* was 97.1%, and 94% for *male*. Although this seemed to provide a good basis for our experiments, when we applied the classifier to real texts, the accuracy dropped dramatically (from 96% to 71.5% determined on a sample of 1,000 nouns). The reason probably is that the majority of nouns in texts are non-actor denoting nouns but the distribution of the classes in the gold data is (almost) balanced.

In order to approach a more realistic distribution, we retrained our model by using GermaNet (Hamp and Feldweg, 1997) noun classes. There are 23 basic noun classes (e.g. artefact, location - separate files are given), from which we excluded the obvious human denoting noun candidates *Gruppe* (Engl. group) and *Mensch* (Engl. human). The rest formed the start of our new non-animacy list. Due to ambiguity, some words from the list of female and male denoting nouns also are expected to be in our initial non-animacy list. To give an example: *Reiseführer* (Engl. travelling guide) is a profession, but also a book (noun class *artefact*). We removed such words (and their synonyms) from the initial non-animacy list.

Table 2 shows the performance of the new classifier based on the final non-animacy list.

	non-animate	♀	♂
precision	99.01	95.44	94.74
recall	99.64	92.59	83.67
F1	99.32	93.99	88.86

Table 2: Performance of the gender classifier

The accuracy is 98.71% which is 2.7% better than our original classifier. However, the performance wrt. gender classes dropped (from 97,1% to 93.99% for female, from 94.7% to 88.86% for male - recall is the problem here). Though this seems to be a substantial quality loss for gender classification, applied to the (above-mentioned) 1,000 real text samples performance increased from 71.5% accuracy (original classifier) to 91.5% (retrained classifier). Our attempts to make gender classification more robust have been successful.

9. Coreference Set Gender Labeling

After we have downloaded the Swiss newspaper texts from Swissdox, we (dependency) parsed them with ParZu (Sennrich et al., 2009) and normalized passive voice. Spacy (Honnibal et al., 2020) (version 3.5.4) was used to do named-entity recognition and coreference resolution. Each noun of a text then was classified as female, male denoting, or non-animate. Next, all coreference sets were created from the (pairwise) output of coreferee (Spacy's coreference resolution approach), and the sets were labeled as female or male, where possible. Sets without a gender noun are omitted. A set gets a gender label if at least one noun of it has a gender reference. In case of conflicts (misclassifications) a majority vote was taken, in case of parity, the set was labeled male (the majority class).

Labeling coreference sets is beneficial since in German pronouns are not (in general) indicative of the gender of their referents. For instance, the pronoun *sie* (Eng. she) as plural can be used to refer to human referents independent of their gender, it can used to refer to female referents in singular, but also to non-animate objects with a female grammatical gender like *die Brücke .. Sie* (Eng. the bridge ... *she). By assigning gender to coreference sets, we make all pronouns of the coreference set available for inference.

[7] Download at: https://www.cl.uzh.ch/en/texttechnologies/research/opinionmining/sentiment-inference.html

10. Empirical Results

Now that we have everything at hand, medium-sized preprocessed data, lexical resources for quantifying polar reference, and a well-performing classifier for gender identification, the next step is to find out whether gender reference is gender-specific. We have defined three subtasks and we state our claims on the basis of a traditional statistical test, the t-test.

- Task 1 is concerned with polar gender reference in phrases (and predicative sentences) and whether a) some gender significantly more often is referred to by a particular gender and whether this reference is b) significantly stronger/weaker in the mean.
- Task 2 is about the roles gender referents take in the context of polar verbs (denoting polar events). Do authors of some gender assign a particular role significantly more often to referents of their own gender, e.g. that of a positive or negative actor?
- Task 3 focuses on polar relations among gender pairs. Is there a statistically significant difference in the way one gender writes about the positive (in favor of) or negative (against) relationship between gender pairs? For instance, do male authors significantly more often report about male-female oppositions?

To properly verify trends in the data, we carry out (unpaired) t-tests for independent samples. The samples are independent since the authors write their articles usually independently of each other. According to Ross and Willson (2017), a prerequisite for the unpaired t-test is that the standard deviations of the samples are equal. This is (reasonably) true according to the authors, if the ratio of the larger standard deviation to the smaller standard deviation is less than 2. We verified that this holds for our data. Notation: we use $w^{\male}$ and $w^{\female}$ to refer to male and female authors (w for writer), respectively.

10.1. Task 1: Polar Gender Reference

In this task, the gender-specific positive and negative references of phrases (e.g. the genius actress) and predicative sentences (e.g. the actress is genius) were quantified: we counted the frequencies of each gender-gender constellation for both, positive and negative reference. To get a mean value, we normalized per author gender. We can interpret this as conditional probabilities. For instance $p(\female|w^{\female})$, the probability of a positive reference to a female given a female author. Let $f^{+}\female\female$ be the number of cases female authors refer positively to female referents. Let $f^{+}\female$ be the number of references made by female authors. The mean, i.e. conditional probability, then is given by $p(\female|w^{\female}) = f^{+}\female\female / f^{+}\female$.

In order to see whether female reference is gender-specific, we compared this with the mean of male authors $w^{\male}$ referring to female referents in a positive way: $p(\female|w^{\male})$. The two-sided null hypothesis is $h_0 : p(\female|w^{\female}) = p(\female|w^{\male})$. If h_0 is rejected and $p(\female|w^{\female}) > p(\female|w^{\male})$ then $w^{\female}$ reference to females is regarded as significantly higher than $w^{\male}$ reference to females. We could have used directed h_0 versions, but the undirected cases are even stronger since we have to take as a significance level $\alpha/2$.

We did it media-wise (4 newspapers: n_i for $i \in [1..4]$) for positive (n_i+) and negative (n_i-) phrases separately. Table 3[8] shows the results[9]. The significance level is indicated at the end of each pair: * means $\alpha = 0.01$, # is $\alpha = 0.025$.

	$p(\female\|w^{\female})$	$p(\female\|w^{\male})$	$p(\male\|w^{\female})$	$p(\male\|w^{\male})$
n_1+	0.28	0.20 *	0.72	0.80 *
n_1-	0.21	0.19 #	0.79	0.81 #
n_2+	0.31	0.21 *	0.69	0.79 *
n_2-	0.18	0.18	0.82	0.82
n_3+	0.24	0.18 *	0.76	0.82 *
n_3-	0.21	0.15 *	0.79	0.85 *
n_4+	0.32	0.19 *	0.68	0.81 *
n_4-	0.23	0.16 *	0.77	0.84 *

Table 3: Phrasal polar gender reference

We can see that for positive reference (n_i+) in all media $p(\female|w^{\female}) > p(\female|w^{\male})$. We might conclude from this that female authors significantly more often refer positively to female referents in their texts than male authors. For positive reference to males the opposite holds: $p(\male|w^{\female}) < p(\male|w^{\male})$. Male authors refer significantly more often to male referents in a positive way than female authors do.

This pattern, namely that each gender refers to its own gender statistically more frequently than to the other one, holds in every newspaper. The only exception is the boulevard newspaper (n_2) where negative reference is not gender-specific.

Table 4 shows the results, if we do it for all media at once, both, for positive (+) and negative ($-$)

[8] Note that for all tables except table 5 the numerical differences between the means of fields of each row are identical (e.g. 1st row, field 0.28 & 0.20 and 0.72 & 0.80). We show both for convenience. If one pair is significant, then the other as well, since the t-value of the unpaired t-test depends on the variance (which is identical) and the differences between the means (again identical).

[9] We used scipy to determine the p-values.

reference. We can see that now all tendencies are significant at $\alpha = 0.01$.

	$p(♀\|w^{♀})$	$p(♀\|w^{♂})$	$p(♂\|w^{♀})$	$p(♂\|w^{♂})$
+	0.27	0.19 *	0.73	0.81 *
-	0.21	0.16 *	0.79	0.84 *

Table 4: All media collapsed: positive and negative reference

To sum up task 1a): female authors are significantly more interested in female referents and less in male referents. For male authors, this is the other way round. This is a binary dimension: interested versus not interested. Since we have a valence lexicon with scores per word, we could also determine the mean *strength* of positive or negative reference, what we introduced as task 1b). Is it for both genders identical? Or do, for instance, female authors refer stronger (positively or negatively) to female referents in the mean as male authors do?

Let $\bar{s}_{♀♀}$ be the mean strength valence values of $w^{♀}$ wrt. to female referents. Let further be $\bar{s}_{♂♀}$ the mean strength valence values of $w^{♂}$ wrt. to female referents. The two-sided null hypothesis is h_0: $\bar{s}_{♀♀} = \bar{s}_{♂♀}$. See table 5 for the results of the gender-specific mean valence patterns.

	$\bar{s}_{♀♀}$	$\bar{s}_{♂♀}$	$\bar{s}_{♀♂}$	$\bar{s}_{♂♂}$
n_1+	1.04	0.99	0.94	0.94
n_1-	-0.53	-0.59 #	-0.54	-0.56
n_2+	1.15	1.16	0.99	1.01
n_2-	-0.59	-0.56	-0.54	-0.51
n_3+	0.96	0.89 *	0.95	0.88 *
n_3-	-0.55	-0.55	-0.55	-0.54
n_4+	1.02	1.00	0.94	0.92
n_4-	-0.53	-0.58 #	-0.56	-0.57

Table 5: Media-wise valence means

Since the valence values range from -4 (extremely negative) to 4 (very positive), we have negative mean values for negative reference. We do not see a huge difference, but in two cases (n_1- and n_4-), female authors do refer significantly less negatively to females (at the 5% level) than male authors do and there is one case ($n3^{+}$) where they do refer significantly more positively to both genders than male do.

The mean differences are small, 0.07 being the highest one, see n3+. Can we really speak of a stronger polar reference of female authors? The usual way to measure the impact of statistically significant results is to use a metric for effect size, e.g. Cohen's d (Cohen, 1988), defined here as $d = (\bar{s}_{♀♀} - \bar{s}_{♀♂})/\sigma$ where σ is the pooled standard deviation (Kotz, 1982). It turned out that only in one case (namely n_3) the d-value was above 0.2 which is the lower threshold for a small effect. The two other significant cases are near, but still below 0.2 which is considered as having only little or even no impact. We, thus, lean to reject that female (male) authors in any respect refer stronger to their own gender than their counterpart gender.

Using valence for quantifying the polar load was not discriminative, thus. This, however, does not disqualify the idea of using valence for polar strength. It just means that we have not found a gender-specific stronger or weaker kind of polar reference, in the mean. We have found statistically significant cases that are gender-specific (task 1a), but the intensity of a single of these references (represented by the mean) is not gender-specific.

Please note that effect size for the other settings we discuss (the following cases but also task 1a) is not needed. In these cases we are looking at binary dimensions: a particular polar reference was made to female (1) or male (0). Here we cannot find a strong or weak effect, because nothing increases or decreases like in, for instance, the comparison of grades or diseases given different "treatments", or - as we did in task 1b) the strength of word valences. We, thus, ignore effect size in the rest of the experiments.

10.2. Task 2: Polar Gender Roles

Whereas in task 1 the polar load was determined based on the valence of polar adjectives and nouns from the valence lexicon, in task 2, gender reference is neutral (most of the time), but a polar verb frames gender reference in a polar way: the semantic roles are qualified as bearing a polar load. The agent (source role) of a verb can be a positive or negative actor. The patient (target role) can receive a positive or negative effect. We try to find out for each polar role whether each gender assigns it significantly more often to its own gender than to the other one.

Table 6 shows the results of the media-wise independent t-test. An $a+$ means positive, $a-$ negative actor, $e+$ means positive, $e-$ negative effect. Again * and # at the end of each pair denote the significance level, 0.01 and 0.025 respectively.

In all newspapers $p(♀|w^{♀})$ is significantly higher than $p(♀|w^{♂})$ wrt. positive actor attribution (n_ia+). For reference to males, this is the inverse: male authors identify significantly more often positive male actors than female authors do. For negative actorship, there are only two significant cases at n_2a-. Negative actorship attribution (n_ia-) in general is not gender-specific, thus. For positive and negative effects (n_ie+, n_ie-), these gender-specific patterns are significant: more own-gender positive and negative reference than cross-gender reference. Table 7 shows the results for the whole dataset. There is

	$p(♀\|w^{♀})$	$p(♀\|w^{♂})$	$p(♂\|w^{♀})$	$p(♂\|w^{♂})$
n_1a+	0.34	0.23 *	0.66	0.77 *
n_1a-	0.20	0.17	0.80	0.83
n_1e+	0.28	0.20 *	0.72	0.81 *
n_1e-	0.27	0.18 *	0.73	0.82 *
n_2a+	0.35	0.21 *	0.65	0.79 *
n_2a-	0.23	0.17 #	0.77	0.83 #
n_2e+	0.33	0.22 *	0.67	0.78 *
n_2e-	0.28	0.24 *	0.72	0.76 *
n_3a+	0.28	0.21 *	0.72	0.79 *
n_3a-	0.17	0.17	0.83	0.83
n_3e+	0.23	0.19 *	0.77	0.81 *
n_3e-	0.24	0.19 *	0.76	0.81 *
n_4a+	0.36	0.24 *	0.64	0.76 *
n_4a-	0.17	0.17	0.83	0.83
n_4e+	0.30	0.22 *	0.70	0.78 *
n_4e-	0.27	0.22 *	0.73	0.78 *

Table 6: Polar role perspectives per newspaper

no significant result, no gender-specific pattern for negative actorship (a-). The rest of cases is in line with patterns described media-wise (from table 6).

	$p(♀\|w^{♀})$	$p(♀\|w^{♂})$	$p(♂\|w^{♀})$	$p(♂\|w^{♂})$
a+	0.33	0.22 *	0.67	0.78 *
a-	0.19	0.17	0.81	0.83
e+	0.28	0.21 *	0.72	0.79 *
e-	0.27	0.20 *	0.73	0.80 *

Table 7: Overall polar role perspectives

Female and male authors see their own gender more positively acting, but also targeted by both, positive and negative effects more than the other one. This is an unexpected tendency. There is no cross-gender variation.

10.3. Task 3: Gender-Gender Relations

As a final task, we checked whether positive and negative gender-gender (in-favour, against) relations are gender-specific given $w^{♀}$ or $w^{♂}$. Again we determine the mean gender-wise. Take for instance ♀ $\rightarrow$ ♂ (in-favour) and female writer $w^{♀}$. For $w^{♀}$, let $f^{+}_{w^{♀}♀}$ be all cases where a female source (the agent) is in a positive relation ($+$) towards either male or female targets (themes). $f^{+}_{♂}$ is the number of cases of male targets out of $f^{+}_{w^{♀}♀}$. The mean of ♀ $\rightarrow$ ♂ is given by $f^{+}_{♂}/f^{+}_{w^{♀}♀}$. For the given example (row 2, table 8) this is 0.42. This is the $w^{♀}$-specific conditional probability of a male target given a female referent as a source, $p(♂|♀, w^{♀})$. The complementary case is the $w^{♀}$-specific conditional probability of a female target given a female referent as a source, which is 0.58 (row 1). Both add up to 1. Table 8 shows the results (blue arcs denote *in favour*, red *against* relations). The colored cells are discussed below.

		α	$w^{♀}$	$w^{♂}$
1	♀ $\rightarrow$ ♀	-	0.580	0.547
2	♀ $\rightarrow$ ♂	-	0.420	0.453
3	♂ $\rightarrow$ ♀	1%	0.165	0.279
4	♂ $\rightarrow$ ♂	1%	0.835	0.721
5	♀ $\rightsquigarrow$ ♀	1%	0.461	0.341
6	♀ $\rightsquigarrow$ ♂	1%	0.538	0.658
7	♂ $\rightsquigarrow$ ♀	1%	0.245	0.189
8	♂ $\rightsquigarrow$ ♂	1%	0.754	0.810

Table 8: Gender-gender relations: in favour $\rightarrow$ and against $\rightsquigarrow$ for $w^{♀}$ and $w^{♂}$.

Again we used the two-sided version of the t-test for independent samples. For $\rightarrow$ two cases are significant: $w^{♂}$ see significantly more often positive relations between male sources and female targets than female authors (row 3). $w^{♀}$ claim significantly more *in favour* relations among male referents than male authors do (row 4).

For $\rightsquigarrow$ every pattern is significant. Both, female and male authors see significantly more intra-gender oppositions of their own gender than for the opposite one (rows 5 and 8): female authors cast a high number of female-female oppositions (row 5), male authors a high number of male-male oppositions (row 8). $w^{♀}$ see significantly more cross-gender oppositions between male sources and female targets (row 7) than $w^{♂}$. For the inverse cross-gender case, this is the other way round (row 6): female sources are significantly more often in opposition to male targets according to $w^{♂}$ than $w^{♀}$ claim.

So far we have determined significance between genders, whether they significantly more often claim a particular gender-gender relation. In terms of the table, this is a row-wise comparison (the blue neighborhood cells establish a single example). But we could as well do an intra-gender comparison, whether e.g. $w^{♀}$ significantly more often report on male-female oppositions than on male-male ones. This would combine the cells column-wise like for the red entries of rows 5 and 6 under column $w_{♀}$. If the means are close to 0.5, the significance is unclear, higher differences seem to support a significant imbalance. We determined it for all cases with the t-test. It turned out that all intra-gender cases are significant, even the red ones (with a mean close to 0.5) We found, for instance, that both $w_{♂}$ and $w_{♀}$ claim significantly more often male-male oppositions than male-female ones (rows 7 and 8).

Looking at all pairs, it turned out that both genders are in line. If one gender shows a particular intra-gender imbalance, the other one shows the

same one. For instance, in rows 3 and 4, both, $w_{\male}$ and $w_{\female}$ significantly more often report male-male oppositions than female-male oppositions. They agree. Since they do so for all cases, we do not have any gender-specific perspectives here.

11. Error and Limitations Discussion

Our preprocessing components are not perfect, parsing, gender classification, and sentiment inferences make mistakes. In order to quantify this, we inspected altogether 1,500 assignments: 500 polar noun phrases, and 250 cases for each polar role (positive/negative actor/effect). For noun phrases, we counted noun misclassifications, where the gender classifier failed (15 cases): the error rate is 3% (15/500). For polar actors the error rates are 7% (positive), 8.5% (negative), for effects we have 4.3% (positive) and 7.2% (negative). Error rates for negative polar facts (8.5% and 7.2%) are not neglectable - they might skew the empirical results towards one gender. We measured the gender-wise distributions, i.e. percentage of female/male author statistics affected by the errors. Negative error rates are slightly skewed but not totally imbalanced - male author statistics has a 1.9% higher error rate. There are only a few cases, where the significance claims reported in tables 3,6 and 7 could be affected (if we'd use the error rate difference directely to reduce the means).

Our manual explorations of the corpus material raised the suspicion that in German plural human denoting nouns in coordinations are gendered not as consequently as singular nouns. Multiple plural human referents like *Ermittler und Forensiker* (Eng. investigators and forensic experts) seem to rarely include female wordforms like in *Ermittler, Ermittlerinnen, Forensiker und Forensikerinnen* were the *innen* suffix indicates female plural form. However, we have no means to find out the reason, i.e. whether the groups are in fact single-gendered male or not. If it is the case that plural human-denoting nouns in coordinations are still often used generically, then we make counting errors, since we regard *Ermittler* (Eng. investigators) as male reference given that its grammatical gender is male.

Although we have set up a carefully targeted approach where syntactic and selectional restrictions in combination with a well-performing gender classifier reduce the risk of unwarranted analysis, a certain degree of noise is still present.

12. Related Work

The literature wrt. to valence lexicons has been discussed in section 4, that of gender classification for German in section 8, the sentiment inference approach is discussed in section 7. We are not aware of any approach directly comparable to ours. Certainly not for German. The closest work is Klenner (2023). There the gender classifier that we newly trained (and improved) is used for gender profiling. The goal there is to find out whether particular verb roles are predominantly occupied by particular gender identities. Also, statistical tests are applied, especially on the basis of the binomial distribution. The gender of the author of a text is not crucial there. This is the main focus of this paper. We also looked into phrase polarity, which is not part of the work of Klenner (2023).

13. Conclusions

In this paper, we carried out the first empirical investigation on gender-specific (polar) gender-reference in texts. We have verified that the resources we used are sound and that the procedures we applied have a good (improved) performance and we have based our claims on an established statistical test to ensure reliability. We also have made an error analysis and specified the limitations of our work. Our work shows that gender-reference in German-language newspapers is gender-specific. This means not only that each gender significantly more often focuses on their own gender than the other one, but it also has a qualitative dimension. For instance, positive actorship supremacy is claimed gender-wise while at the same time, more negative effects are seen intra-gender than for the other gender. Female authors see more negative relations originating from males targeting females, while male authors see it the other way round. Polar actorship, polar effects, and polar relations all are in a particular way perceived and focused on differently depending on the gender of the author. We have found out that this is not newspaper-specific but is a global trend independent of the political orientation. However, since these trends are often mutually inverse (e.g. females see female supremacy, males see male supremacy), we cannot claim that there is a feministic-oriented (re)framing happening. For such a claim, a much lower male-male orientation would be needed, but parity is given.

14. Ethical Statement

We have followed the guidelines of Larson (2017) for using gender as a variable in NLP: We pointed out in section 2 that our notion of gender reference here is binary and that this is caused by the restriction posed by the grammatical gender of human-denoting nouns in German, which is binary: female or male. We stress again the point that we do not claim gender to be a binary class: there are more than two gender identities.

15. Bibliographical References

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for english words (ANEW): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Marc Brysbaert, Amy Warriner, and Victor Kuperman. 2013. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46.

Simon Clematide and Manfred Klenner. 2010. Evaluation and extension of a polarity lexicon for German. In *Proceedings of the First Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*.

J. Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

Jakob Fehle, Thomas Schmidt, and Christian Wolff. 2021. Lexicon-based sentiment analysis in German: Systematic evaluation of resources and preprocessing techniques. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 86–103, Düsseldorf, Germany. KONVENS 2021 Organizers.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, Spain.

Verena Henrich and Erhard Hinrichs. 2010. Gernedit - the germanet editing tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, pages 2228–2235, Valletta, Malta.

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Manfred Klenner. 2023. Sentiment inference for gender profiling. In *Language, Data and Knowledge 2023 (LDK 2023): Proceedings of the 4th Conference on Language, Data and Knowledge*, Vienna, Austria. NOVA FCSH - CLUNL.

Manfred Klenner, Don Tuggener, and Simon Clematide. 2017. Stance detection in Facebook posts of a German right-wing party. In *LSDSem 2017/LSD-Sem Linking Models of Lexical, Sentential and Discourse-level Semantics*, Valencia, Spain. Association for Computational Linguistics.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 German lemmas. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2595–2598, Portorož, Slovenia. European Language Resources Association (ELRA).

Samuel Kotz, editor. 1982. *Encyclopedia of statistical sciences*. A Wiley-Interscience publication. Wiley, New York, NY [u.a.].

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

J. Lüdtke and K.G. Hugentobler. 2022. Using emotional word ratings to extrapolated norms for valence, arousal, imagebility and concreteness: The German list of extrapolated affective norms (GLEAN). In *Proceedings of KogWis2022, the 5th Biannual Conference of the German Society for Cognitive Science*. Albert-Ludwigs-Universität Freiburg, Germany.

J. R. Martin and P. R. R. White. 2005. *Appraisal in English*. Palgrave, London.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

C.E. Osgood, G.J. Suci, and P.H. Tenenbaum. 1957. *The Measurement of meaning*. University of Illinois Press, Urbana.

Amanda Ross and Victor L. Willson. 2017. *Basic and Advanced Statistical Tests: Writing Results Sections and Creating Tables and Figures*, chapter Independent Samples T-Test. pages 13–16. SensePublishers.

David Schmidtke, Tobias Schröder, Arthur Jacobs, and Markus Conrad. 2014. Angst: Affective norms for German sentiment terms derived from the affective norms for english words. *Behavior research methods*, 46.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. 2009. A new hybrid dependency parser for German. In *Proceedings of the German Society for Computational Linguistics and Language Technology*, pages 115–124.

Peter Turney, Peter D., Littman, and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21.

Melissa Võ, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus Hofmann, and Arthur Jacobs. 2009. The Berlin affective word list reloaded (BAWL-R). *Behavior research methods*, 41.

Amy Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45.