GIL-GALaD: Gender Inclusive Language - German Auto-Assembled Large Database

Anna-Katharina Dick, Matthias Drews, Valentin Pickard, Victoria Pierz
Seminar für Sprachwissenschaft, Eberhard Karls Universität Tübingen
Keplerstraße 2, 72074 Tübingen, Germany
{anna-katharina.dick, matthias.drews, valentin.pickard, victoria.pierz}@student.uni-tuebingen.de

Abstract

As the need for gender-inclusive language has become a highly debated topic over the years, gendered biases in speech are unfortunately often picked up and propagated by modern language models trained on large amounts of text. While remedial efforts are underway, grammatically gendered languages such as German pose some unique challenges in generating gender-inclusive language for corrective model training or fine-tuning. We assembled GIL-GALaD, a corpus of German gender-inclusive language from different sources such as social media, news articles, public speeches and academic publications. Our corpus includes the most common types of modifications of generic masculine forms of nouns and spans 30 years (1993-2023), containing over 800,000 instances of gender-inclusive language. Tools for corpus usage and extension are to be included in the release. During corpus assembly, we were also able to gain some insights into which types of gender-inclusive language were used in practice throughout the years and across different domains.

Keywords: Corpus, Inclusive language, German, Gender bias

1. Introduction

In recent years, awareness of the need for inclusive and gender-neutral communication has increased. There are many options being explored for gender-inclusive language, but not all options are equally adaptable to all language systems. In "natural gender" languages like English, semantic gender is only visible through pronouns or gendered nouns like brother or sister. In grammatically gendered languages like German, semantic gender is also reflected in other parts of speech, such as articles, adjectives, and adverbs (Hord, 2016). German nouns referring to people are also traditionally gendered in a generically masculine way, for example in job titles. As such, creating gender-neutral expressions in German poses more challenges than in English. The use of generically masculine forms to refer to persons of unspecified or unknown gender has been raising concerns about female visibility, sexism and discrimination since the late 1970s (Trömel-Plötz, 1978) and prompted German feminists to push for a move away from generically masculine forms of nouns, especially in professional fields (Hord, 2016). We now briefly describe the most common strategies for gender-inclusive language in German, using the plural noun *'Lehrer'* (teachers m./f.) as an example. A first approach is to use generically feminine forms (1) instead of masculine forms, typically this is achieved by appending the suffixes *-in* for singular or *-innen* for plural nouns. More commonly both masculine and derived feminine forms are given, either joined by an appropriate coordinating conjunction (2), or in a single word, by capitalizing the suffix boundary (so-called 'Binnen-I', 2a) or by interfixing special symbols (2b). The latter is usually intended to emphasize the existence and inclusion of non-binary genders as well. Alternatively, when possible, explicitly gender-neutral rephrasings are used, by using adjectival constructions (3), nominalized participles or adjectives (4) or abstract genderless nouns (5). The use of such expressions is not without controversy however, and there are different levels of acceptance and use of different types of gender-inclusive forms across different demographics and publications.

(1) *Lehrerinnen*
'teachers (f.)'

(2) *Lehrerinnen und/oder Lehrer*
'teachers (f.) and/or teachers (m.)'

(2a) *LehrerInnen*

(2b) *Lehrer*innen*, *Lehrer_innen*, *Lehrer/-innen*

(3) *lehrende Personen*, *lehrende Menschen*
'teaching persons', 'teaching people'

(4) *Lehrende*
'teaching ones'

(5) *Lehrkräfte*, *Lehrerschaft*
'teaching forces', 'teaching body'

To the best of our knowledge, there is no publicly available corpus of German gender-inclusive language so far. We assembled such a corpus of

LREC-COLING 2024, pages 7740–7745

gender-inclusive German with two main objectives in mind. Firstly, we want to be able to observe across varied domains, which forms of and approaches to gender-inclusive language are used in practice, as well as their respective frequencies and contexts. Secondly, we intend to use our corpus as training data to create more gender-inclusive language models by enriching existing models with adaptable samples gleaned from our corpus. Problems with gendered biases in primarily English language models have been identified and discussed previously (Bolukbasi et al., 2016), and such problems become even more visible in more explicitly gendered languages such as German. Machine translation models that translate from less gender-specific languages such as English into more gender-specific languages such as German are particularly susceptible to gender bias issues (Stanovsky et al., 2019), usually preferring to resolve unspecified source gender as generic masculine in the translation. This then is not just a sociological issue but it directly affects translation accuracy, as it can break co-reference links between nouns and pronouns that would normally have to agree in gender. Here, our corpus can also provide German training and test data with aligned generically masculine, feminine or gender-inclusively written samples. The corpus and its associated tools will be provided on github. [1]

2. Data & Methods

Our goal is to draw examples of gender-inclusive language from a range of different domains and registers, to observe potential differences across those and to be able to provide varied and helpful test and training data for language models. We therefore use various sources, a large collection of German tweets to cover varied domains at a more informal register and several comparatively smaller corpora written in a more formal register. The Wortschatz Leipzig corpus comprises sentences from German newspaper articles and we also include articles from magazines by the Bundeszentrale für politische Bildung (BPB) that mostly cover social and political topics. These, together with German transcriptions from the proceedings of the European Parliament, were used to obtain longer, coherent texts embedding the desired gender-inclusive forms, as opposed to single sentences or short tweets. Lastly, we also included a small collection of publications and dissertations in humanities from the University of Tübingen. A brief description of the data sources follows.

Europarl The parallel corpus from the proceedings of the European Parliament (Koehn, 2005) contains among others, about 2 million German sentences from the years 1996 to 2003. We split those into continuous single-speaker turns, discarding procedural annotations such as session agenda and similar meta-information so that we are left with about 90,000 continuous, single-speaker contributions. Extracting gender-inclusive mentions we find about 12,000 hits across about 9,000 speaker turns.

Wortschatz Leipzig The Wortschatz corpus (Goldhahn et al., 2012) contains sentences from German newspaper articles. It covers online articles from 1995 up to 2022. We extracted the one million sentence versions of the corpus from each year, totaling 27 million sentences. To find examples of type (5) that are not as straightforward to retrieve using regular expressions, a lexicon of such forms was used.[2] Many of the gender-neutral forms contained therein are not commonly used and do not appear in the corpus at all, they often can only be used in specific contexts or have different connotations to the gendered version.

Twitter We analysed approximately 1.14 billion German tweets from June 2019 to February 2023 that were collected from the former Twitter streaming API by using a list of common German words. Due to the vast amount of data and a higher prevalence of false positives of gender-inclusive hits for types (3), (4) and (5), we only collected samples of type (1) and (2) including subtypes. This yielded about 650,000 hits across around the same number of tweets.

Bundeszentrale für politische Bildung The Bundeszentrale für politische Bildung (BPB) is the German Federal Agency for Civic Education, and publishes several free magazines either online or in print. One of these magazines is *Aus Politik und Zeitgeschichte* (APuZ), which covers both contemporary as well as specialized topics. This publication sometimes includes gender-inclusive language, although there are no official guidelines or requirements. We included the yearly anthology from 2014 until 2022 in our research (APuZ, 2014-2022).

Academic texts We chose to also include publications and dissertations from the University of Tübingen that had to be preprocessed manually, resulting in fewer samples. These are scientific texts by students in academia that may offer an interesting contrast to the other sources. We randomly chose about 30 papers from the philosophical faculty for analysis.

3. Corpus Statistics & Results

For each source text we additionally store a unique ID, a source descriptor string, the year of publication and all gender-inclusive forms contained

[1] https://github.com/iscl-lrl/gil-galad

[2] https://geschicktgendern.de/

therein. If available, the author and publication day, month and location are also recorded. Author metadata will be pseudo-anonymized for publication. Source texts vary in length from single sentences or short tweets to full news articles or academic papers. For each form, the source text ID, the type of gender-inclusive form (such as e.g. Binnen-I), the original form itself, both start and end index within the text sample and, where available, the derived exclusively masculine and feminine forms are saved. Our corpus is thus organized in two tables, one for source texts and one for associated gender-inclusive forms. We will provide those in plain text/CSV format together with tools to transform the data to other commonly used formats such as JSON or XML. In total, more than 830,000 occurrences of gender-inclusive forms were extracted from the given sources, with about 44,000 unique examples. These came from just shy of 180,000 different authors – although it should be noted that not every source had explicitly mentioned authors. The largest source by far of our corpus is Twitter, at about 80% of all gender-inclusive forms (Table 2). In contrast, the smallest source are academic texts with just a bit over 1000 forms, due to the limited number of source texts.

id	text	source	year	month	day	author	location
tw_0	"Werdet heute Nacht selbst zu Forschern und Forscherinnen und entdeckt was man mit Licht und einem Mikroskop so anstellen kann! [...]" *Become researchers (m.) and researchers (f.) yourselves tonight and discover what you can do with light and a microscope!*	Twitter	2019	06	14	auth_1	Dresden

id	original	start	end	masculine	feminine
tw_0	Forschern und Forscherinnen *researchers (m.) and (f.)*	29	56	Forschern *researchers (m.)*	Forscherinnen *researchers (f.)*

Table 1: Example source text and gender-inclusive form entries extracted from Twitter data

Source	Frequency
Twitter	670980
Wortschatz Leipzig	89046
Europarl	57270
APuZ magazine	17866
Academic texts	1147

Table 2: Gender-inclusive forms per source

The vast majority of all examples, around 79%, are explicit double mentions (2), as seen in Table 3. This can be explained through the comparatively long history of mentioning both male and female addressees. Several state governments, for example North Rhine-Westphalia (Justizministerium et al., 1993) require that official correspondence on all levels is written to equally address both women and men. The most frequent three examples, adding up to about 17% of all occurrences, refer to both female and male colleagues, citizens and (school) students (Table 4).

Strategy	Frequency
Explicit double mention (2)	659344
Symbols (2b)	72594
(Generically) feminine forms (1)	55150
Abstract genderless forms (5)	19548
Binnen-I (2a)	18092
Nominalized adjectives (4)	11564
Adjectival forms (3)	17

Table 3: Frequency of strategies for gender-inclusive language in our corpus

Form	Frequency
Kolleginnen und Kollegen *colleagues (f.) and (m.)*	54495
Bürgerinnen und Bürger *citizens (f.) and (m.)*	46275
Schülerinnen und Schüler *pupils/students (f.) and (m.)*	40804
Präsidentin *president (f.)*	12340
Soldatinnen und Soldaten *soldiers (f.) and (m.)*	11124
Freundinnen und Freunde *friends (f.) and (m.)*	10656
Mitarbeiterinnen und Mitarbeiter *coworkers (f.) and (m.)*	10607
Wählerinnen und Wähler *voters (f.) and (m.)*	10376
Ärztinnen und Ärzte *(medical) doctors (f.) and (m.)*	10195
Bürgerinnen und Bürgern *citizens (f.) and (m.), dative*	10140

Table 4: Ten most common forms

If we exclude those explicit double mentions, and instead look at other types of gender-inclusive forms, we once again find examples of official language like *president (f.)* and *colleagues (f.)*. We have to note that especially the former and other singular feminine forms are more often specific mentions where there referent's gender is known, than actual gender-inclusive mentions. Narrowing down results by excluding purely feminine forms the most frequent word is *Jüd*innen* - 'Jewish people' (m./f./other). The other examples include nominalized participles (*Auszubildende* – 'apprentices'), as well as abstract genderless forms (*Rettungskräfte* – 'emergency service workers'). Various gender-inclusive translations for Jewish people are present in the Top-10 list of examples using specialised

orthography (Table 6). Additional tables containing the ten most frequent forms for each source can be found in the appendix.

Form	Frequency
Kolleginnen *colleagues (f.)*	9676
Kommissarin *commissioner/inspector (f.)*	5819
Jüd*innen *Jewish people (any)*	5623
Berichterstatterin *reporter (f.)*	5040
Rettungskräfte *rescue workers (any)*	4314
Pflegekräfte *nursing staff (any)*	4290
Auszubildende *trainees (any)*	3958
Fachkräfte *professionals/specialists (any)*	3753
Arbeitskraft *employee (any)*	3752
Studierende *students (any)*	3643
Jüd:innen *Jewish people (any)*	3529
Lehrkräfte *teachers (any)*	2018
Asylsuchende *asylum seekers (any)*	1008

Table 5: Most common forms without double mentions, excluding *Präsidentin* from previous table

Form	Frequency
Schüler*innen *pupils/students (any)*	660
SchülerInnen *pupils/students (m. or f.)*	603
JüdInnen *Jewish people (m. or f.)*	564
Bürger*innen *citizens (any)*	550
MitarbeiterInnen *coworkers (m. or f.)*	529
Jüd_innen	521
jüd*innen *Jewish people (any)*	498
Pol*innen *Polish people (any)*	491

Table 6: Most common forms using special symbols or Binnen-I, excluding forms listed in previous table (*Jüd*innen, Jüd:innen*)

4. Discussion

We intend to explore some of the possible use cases discussed in the introduction by training or rather fine-tuning language models using our corpus. One possible application would be the creation of a German-to-Gendered-German Machine Translation model. Using our corpus as a glossary could be a quick solution, but might be too absolute in practice. Neural Machine Translation (NMT) frameworks such as MarianNMT (Junczys-Dowmunt et al., 2018) or OpenNMT (Klein et al., 2017) allow comparatively easy training of MT models, but would require more training data with both neutral and gendered segments. Such training data would not only benefit German-to-German translations, but also for example English-to-German translations. As mentioned in the Introduction section, this language pair may cause difficulties when gender ambiguities have to be resolved, having to rely on potentially biased training data. Our corpus may also help in choosing more inclusive translations instead of hard-resolving ambiguities into strictly male or female translations. A limitation of our corpus is that syntactic and semantic information and context are not yet taken into account and not all possible forms of gender-inclusive language are found. For instance, nominalized adjectival forms cannot be distinguished from gerund forms without accounting for context: *Studierende* is a quite common gender-neutral way of referring to university students, but *Studierende Männer* ('studying men') would not be gender-inclusive. Context is also crucial for distinguishing certain gender-neutral abstract forms such as *Arbeitskraft*: It is often used as a gender-neutral word for worker and is exclusively used with this meaning in the plural, but it could also mean 'capacity for work' in the singular. Similarly, for feminine nouns ending in *-in* or *-innen*, that are not part of double mentions we cannot automatically infer whether they are examples of generically feminine forms used for people of unknown or unspecified gender, or rather specifically feminine forms used for referents of known gender. As such, there are likely false positives for those types of gender-inclusive forms in the corpus, pending manual or improved automated review. Conversely, we deliberately omitted non-standard or misspelled forms of gender-inclusive forms, such as forms containing emojis or parentheses. To remedy such issues we will update and improve our extraction pipeline and publish code to easily extend our corpus with additional data, allowing users to run our extraction pipeline on new data with minimal effort. We also plan to use the same code to iteratively extend our corpus by incorporating further resources.

5. Acknowledgements

We would like to thank Dr. Çagri Çöltekin for his help and guidance in creating this corpus and paper, as well as for providing the Twitter data. We also want to thank the Bundeszentrale für politische Bildung and the editorial office of "Aus Politik und Zeitgeschichte" for allowing us to use their publications. Lastly we thank the anonymous reviewers for their constructive criticism and helpful feedback, we sincerely hope to follow up on some suggestions that we did not investigate yet in future works.

6. Ethical considerations

Biases, whether related to gender or otherwise, in natural language processing and language resources are a controversial topic, and while our corpus is intended as a resource to help mitigate those issues for gendered biases in German, we feel the need to point out, that our data may still contain examples of undesirable biases that our methods were unable to detect. Furthermore, to comply with data protection requirements we anonymize or omit authorship information on the source texts used in our corpus. We exclusively included publicly available source text data and only provide tools to build our corpus from sources that require specific consent for usage and redistribution.

7. Bibliographical References

APuZ. 2014-2022. *Aus Politik und Zeitgeschichte - Jahresausgaben 2014-22*. Bundeszentrale für politische Bildung, Bonn.

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. *Advances in neural information processing systems*, 29.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Levi CR Hord. 2016. Bucking the linguistic binary: Gender neutral language in English, Swedish, French, and German. *Western Papers in Linguistics*, 3(1).

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia.

NRW Justizministerium et al. 1993. Gleichstellung von Frau und Mann in der Rechts-und Amtssprache. Runderlass des Justizministeriums... des Ministerpräsidenten und aller Landesministerien vom 24.03. 1993. *Ministerialblatt Nordrhein-Westfalen*, (31).

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. *arXiv preprint arXiv:1906.00591*.

S. Trömel-Plötz. 1978. Linguistik und Frauensprache. *Linguistische Berichte, 57*.

7.1. Appendices

Twitter	No.	Europarl	No.
Kolleginnen und Kollegen *colleagues (f.) and (m.)*	44040	Präsidentin *president (f.)*	12305
Bürgerinnen und Bürger *citizens (f.) and (m.)*	42000	Kolleginnen *colleagues (f.)*	9641
Schülerinnen und Schüler *students (f.) and (m.)*	37496	Kolleginnen und Kollegen *colleagues (f.) and (m.)*	9375
Soldatinnen und Soldaten *soldiers (f.) and (m.)*	10790	Kommissarin *commissioner/inspector (f.)*	5814
Freundinnen und Freunde *friends (f.) and (m.)*	10535	Berichterstatterin *reporter (f.)*	5040
Ärztinnen und Ärzte *doctors (f.) and (m.)*	9773	Kollegin *colleague (f.)*	2034
Wählerinnen und Wähler *voters (f.) and (m.)*	9710	Bürgerinnen *citizens (f.)*	1598
Mitarbeiterinnen und Mitarbeiter *colleagues (f.) and (m.)*	9251	Bürgerinnen und Bürger *citizens (f.) and (m.)*	1179
Bürgerinnen und Bürgern *citizens (f.) and (m.)*	9050	Ratspräsidentin *president of the council (f.)*	725
Schülerinnen und Schülern *students (f.) and (m.)*	8978	Haushaltsdisziplin *budgetary discipline*	352

Table 7: 10 most common occurrences for Twitter and Europarl, note the included false positive *Haushaltsdisziplin*

Wortschatz	No.	APuZ	No.	Dissertations	No.
Rettungskräfte *rescue workers (any)*	4314	Bürgerinnen *citizens (f.)*	457	Freundin *friend (f.)*	96
Pflegekräfte *nursing staff (any)*	4290	Bürgerinnen und Bürger *citizens (f.) and (m.)*	340	Autorin *author (f.)*	41
Auszubildende *trainees (any)*	3958	Bürger*innen *citizens (any)*	318	Kassiererinnen *cashiers (f.)*	33
Fachkräfte *professionals/specialists (any)*	3753	Schülerinnen *students (f.)*	292	Proband*innen *test subject (any)*	32
Arbeitskraft *employee (any)*	3752	Schülerinnen und Schüler *students (f.) and (m.)*	222	Kaiserin *empress (f.)*	28
Studierende *students (any)*	3643	Professorin *professor (f.)*	179	Erzählerin *narrator (f.)*	27
Schülerinnen und Schüler *students (f.) and (m.)*	3078	Jüdinnen *Jewish people (f.)*	168	Ich-Erzählerin *first-person narrator (f.)*	19
Bürgerinnen und Bürger *citizens (f.) and (m.)*	2755	Bundeskanzlerin *federal chancellor (f.)*	157	Übersetzerin *translator (f.)*	19
Lehrkräfte *teachers (any)*	2018	Mitarbeiterin *coworker (f.)*	142	Lebensgefährtin *significant other (f.)*	18
Mitarbeiterinnen und Mitarbeiter *coworker (f.) and (m.)*	1292	Schüler*innen *students (any)*	137	Ensslin *Gudrun Ensslin, German terrorist (RAF)*	18

Table 8: 10 most common occurrences for Wortschatz, APuZ and Dissertations, note the included false positive *Ensslin*