

GIL-GALaD

**Gender Inclusive Language - German Auto-Assembled
Large Database**

Our team

We are a group of students of Computational Linguistics at the University of Tübingen whose coursework evolved into this conference submission.

Valentin

Anna-Katharina

Matthias

Victoria

Pickard

Dick

Drews

Pierz

Email: *firstname.lastname@student.uni-tuebingen.de*

Introduction to gender-neutral German

- German nouns referring to people are gendered in a generically masculine way: “Leser” *reader(s), male reader(s)*; “Leser**in(nen)**” *female reader(s)*
- Exceptions only for female-coded professions: “Putzfrau” *cleaning lady*
- Concerns about female visibility i.e. in the professional field, and of people of other gender identities, propagation of harmful stereotypes
- Various strategies for more gender-inclusive language
- Highly debated topic, some recent backlash

Our project

- Extract sample texts containing instances of gender-inclusive mentions from various sources
- Make them available in a unified, convenient format so they can be used to develop language models that are more sensitive to gender inclusivity
- Sample different source domains and different text lengths
- Incidentally observe trends in use of such forms

Strategies for gender-inclusive language

(1) Generic feminine

- Using the feminine form as default instead of the masculine form
- “Lehrer (sg./pl.)” (*male*) *teacher/teachers* becomes “Lehrerin (sg.)”, or “Lehrerinnen (pl.)” (*female*) *teacher/teachers*
- Easy to implement, easily readable
- Can be seen as ‘reverse discrimination’ and perceived as exclusive to non-binary genders as well as males

Strategies for gender-inclusive language

(2) Double mentions

- Using both masculine and feminine forms together explicitly
- “Lehrer (pl.)” (*male*) *teachers* becomes “Lehrer und Lehrerinnen” *male teachers and female teachers*
- Can reduce readability and be awkward and tedious to implement especially for singular forms and with additional articles & adjectives that need to be inflected by gender
- Sometimes perceived as exclusive to non-binary genders

Strategies for gender-inclusive language

(2a) Double mentions with Binnen-I

- Dropping the conjunction and capitalizing the morpheme boundary
- “Lehrer (pl.)” (*male*) *teachers* becomes “LehrerInnen” *male teachers and female teachers*

(2b) Double mentions with special characters

- Dropping the conjunction and marking the morpheme boundary with special characters such as * to include non-binary genders
- “Lehrer (pl.)” (*male*) *teachers* becomes “Lehrer*innen” *teachers of any gender*

Strategies for gender-inclusive language

(3) Neutral adjectival constructions

- “Lehrer (pl.)” (*male*) *teachers* becomes “lehrende Personen” *teaching persons* or “lehrende Menschen” *teaching people*

(4) Neutral nominalized participles

- “Lehrer (pl.)” (*male*) *teachers* becomes “Lehrende” *teaching ones*

(5) Neutral abstract nouns

- “Lehrer (pl.)” (*male*) *teachers* becomes “Lehrkräfte” *teaching forces* or “Lehrkörper” *teaching body*

Varying levels of acceptance, abstract forms not always feasible/available

Methods

- Preprocess corpora, segment into appropriate samples (speaker turns, tweets, sentences)
- Scan & extract inclusive mentions using regular expressions and word lists
- Manually discard false positives, create shared exclusion list (e.g *LinkedIn*)
- Format, store and label data
- Compute some preliminary statistics

Data sources

Twitter (tweets)

- 1.14 billion German tweets from June 2019 to February 2023
- Collected from former Twitter streaming API by using a list of common German words, only scan for subtypes 1 & 2
- 650k hits across about the same number of tweets

Wortschatz Leipzig (sentences)

- Corpus of 27 million sentences from (online) newspaper articles (1995-2022)
- Extract all subtypes, almost 90k hits

Data sources

Europarl (speaker turns)

- Parallel corpus from the proceedings of the European parliament
- 2 million German sentences from 1996 to 2003
- 12k hits across 9k speaker turns

APuZ magazine (sentences)

- Free magazine published by the German Federal Agency for Civic Education
- Covers contemporary topics, yearly anthology from 2014-2022
- ~17k hits found

Additionally: Sentences from ~30 academic texts by students of Uni Tübingen

Results - overview

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

Gender-inclusive mentions found
per source

- over 800,000 inclusive mentions total
- still contains some false positives
- available in convenient table format
- extraction and some analysis scripts will also be made available for easy extension and comparison (<https://github.com/iscl-lrl/gil-galad>)

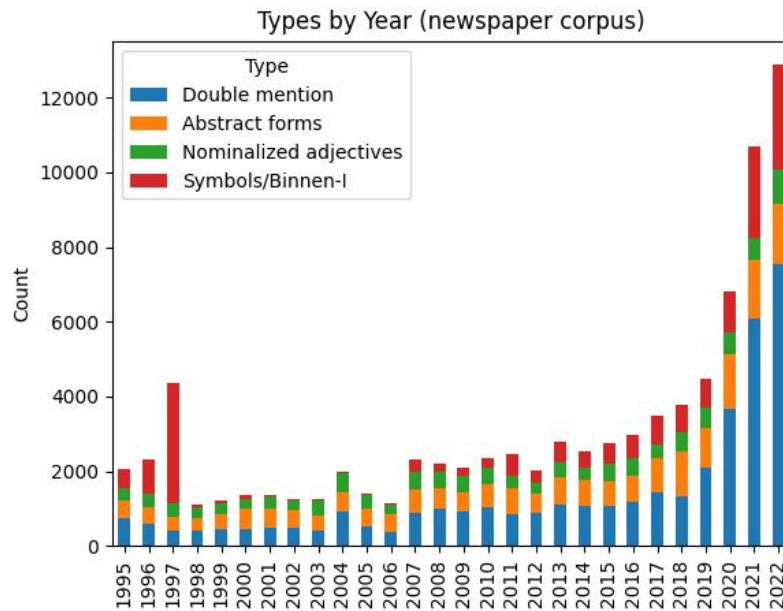
Results - data format

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! [...]" <i>Become researchers (m.) and researchers (f.) yourselves tonight and discover what you can do with light and a microscope!</i>	Twitter	2019	06	14	auth_1	Dresden

id	original	start	end	masculine	feminine
tw_0	Forschern und Forscherinnen <i>researchers (m.) and (f.)</i>	29	56	Forschern <i>researchers (m.)</i>	Forscherinnen <i>researchers (f.)</i>

Results - some observations

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



Acknowledgements

We would like to thank Dr. Çagri Çöltekin for his help and guidance in creating our 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.