Universität Stuttgart

Agnieszka Faleńska

Steps towards Bias-Aware NLP Systems

IRIS

NLP Group










Who can be a Doctor?

Hän on lääkäri.	→ He is a doctor.
Ona jest na sali operacyjnej. Przeprowadza właśnie operację.	→ She is in the operating room. He's in the middle of an operation.

eine Ärztin	→ lekarz

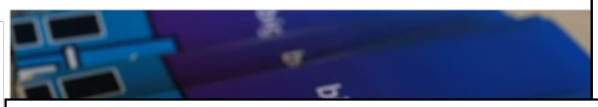
commons.wikimedia.org/

Increasing Concerns about NLP Systems' Harms

Amazon scraps secret AI recruitin
showed bias against women

By Jeffrey Dastin

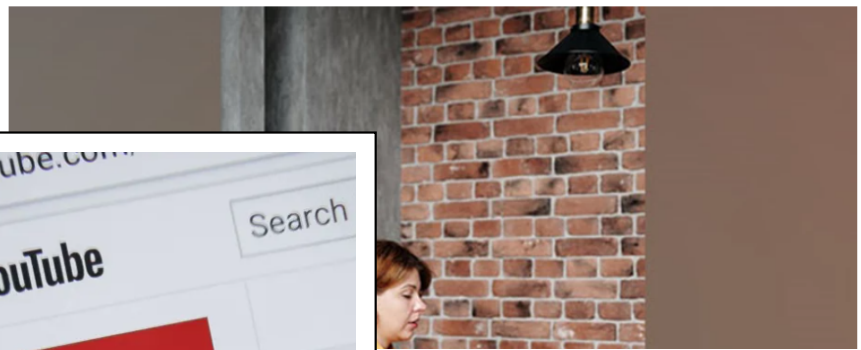
SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.
specialists uncovered a big problem: their new recruiting
women.



AI language models show bias against people with disabilities: Study

Natural language processing (NLP) is a sort of artificial intelligence that allows machines to utilise written and spoken phrases in a variety of applications, such as smart assistants or email autocorrect and spam filters, to help automate and expedite activities for individual users and companies. However, the algorithms that power this technology frequently exhibit characteristics that might be insulting or discriminatory toward people with disabilities.

ANI | Pennsylvania | Updated: 14-10-2022 08:50 IST | Created: 14-10-2022 08:50 IST



REPORT

Echo chambers, rabbit holes, and ideological bias: How YouTube recommends content to real users

Megan A. Brown, Jonathan Nagler, James Bisbee, Angela Lai, and Joshua A. Tucker · Thursday, October 13, 2022

tions

University of Washington
rctatman@uw.edu

tion in language use, for example, has been extensively studied (Trudgill, 1972; Eckert, 1989, among many others). There is also robust variation in language use by native speakers across dialect regions. For instance, English varies dramatically between the United States (Cassidy and others, 1985), New Zealand (Hay et al., 2008) and Scotland (Milroy and Milroy, 2014).

Consequences of Harmful Behaviors

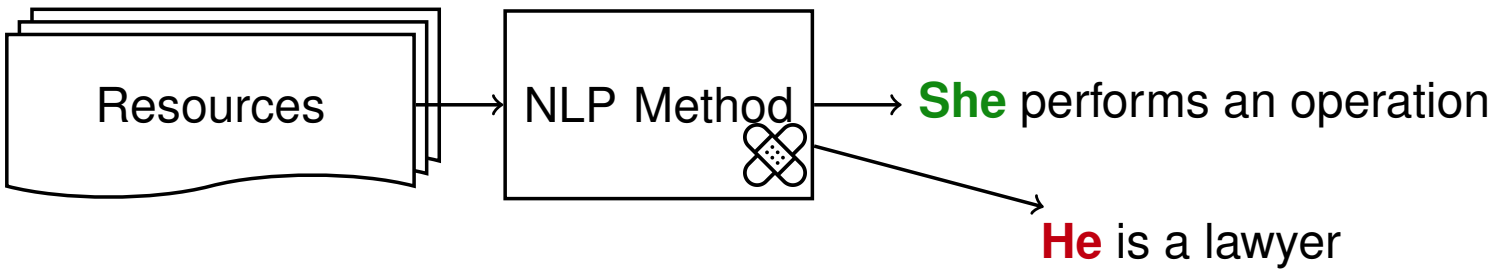
- Allocation and representational harms

  (Blodgett et al., 2020)

- Broad range of applications
  - healthcare, banking, judicial system, ...

- We use them
  - by choice: translators, voice assistants, ...
  - not by choice: CV filtering systems, political sentiment analyzers, ...

Remedies for Harmful Behaviors

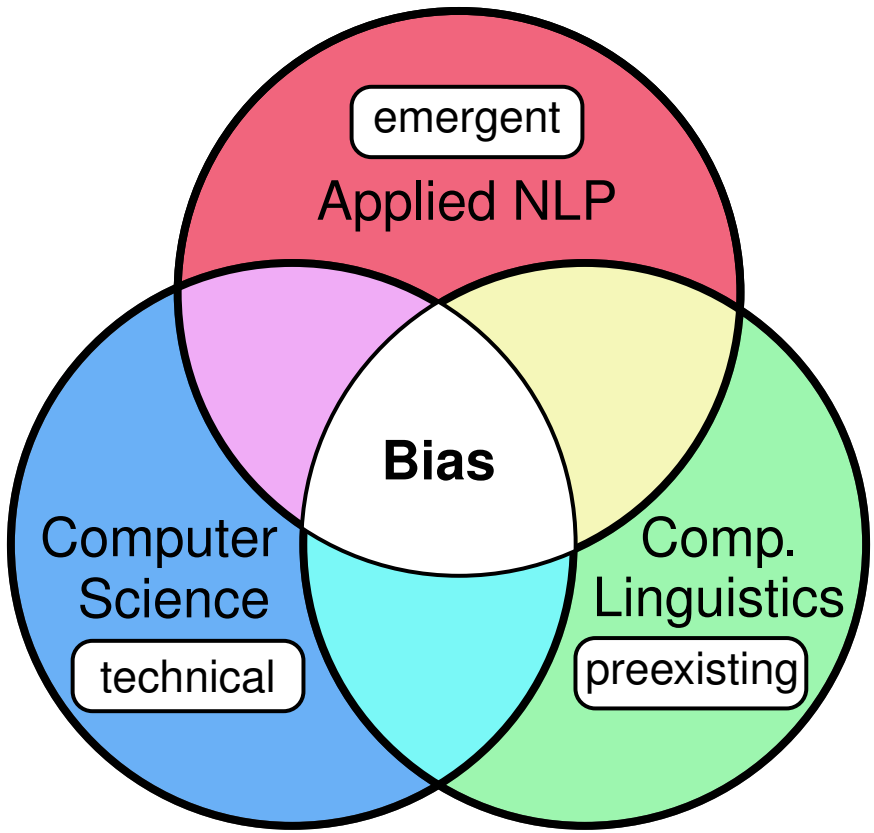


- Harmful behaviors as a technical issue
- Mitigation techniques
- Reduce observed disparities in models' outputs

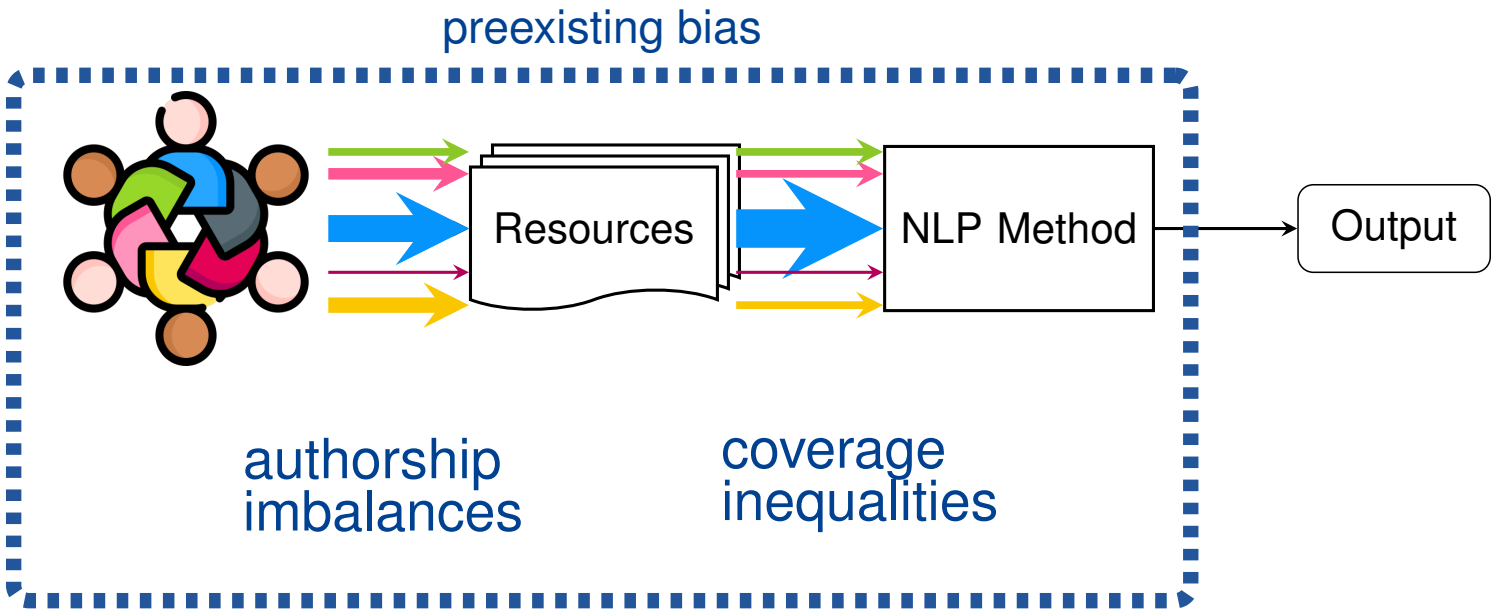
Patch icons from Freepik www.flaticon.com/free-icons/patch

Harmful Behaviors are Symptoms of Bias

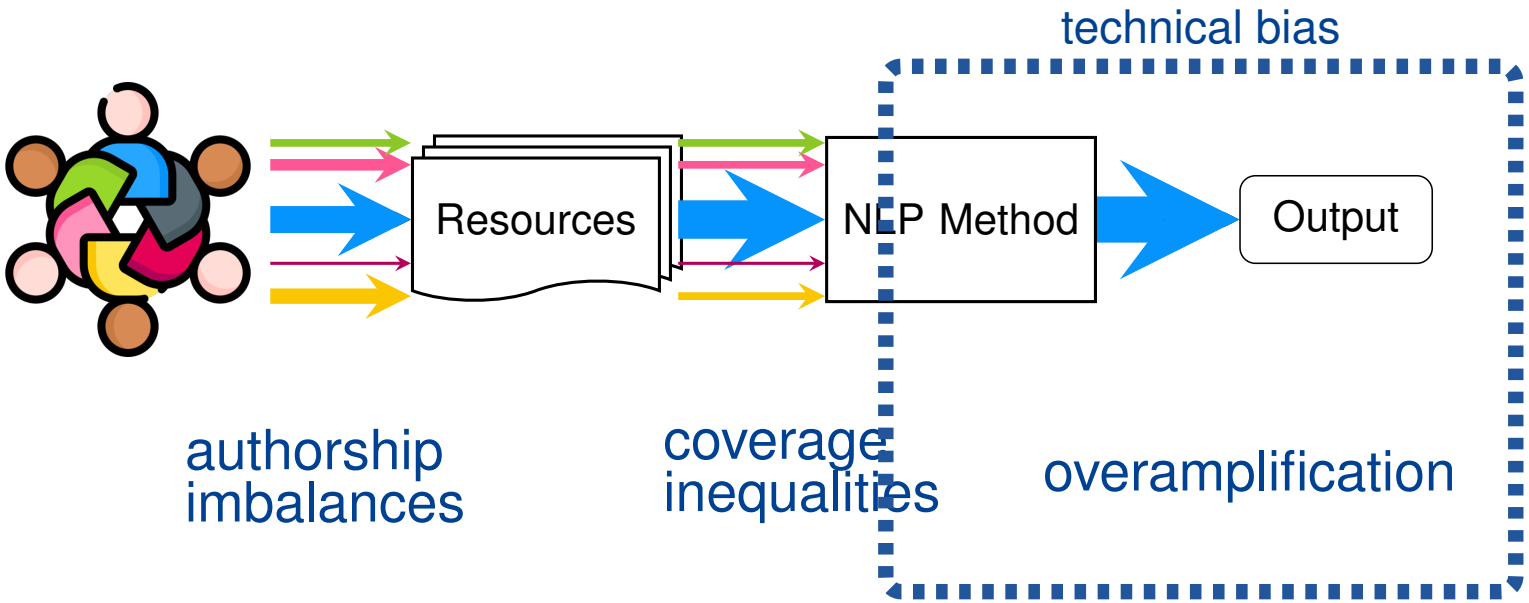
Bias – systematic preference or discrimination against certain groups of users (Friedman and Nissenbaum, 1996)



NLP Systems vs. Friedman and Nissenbaum (1996)

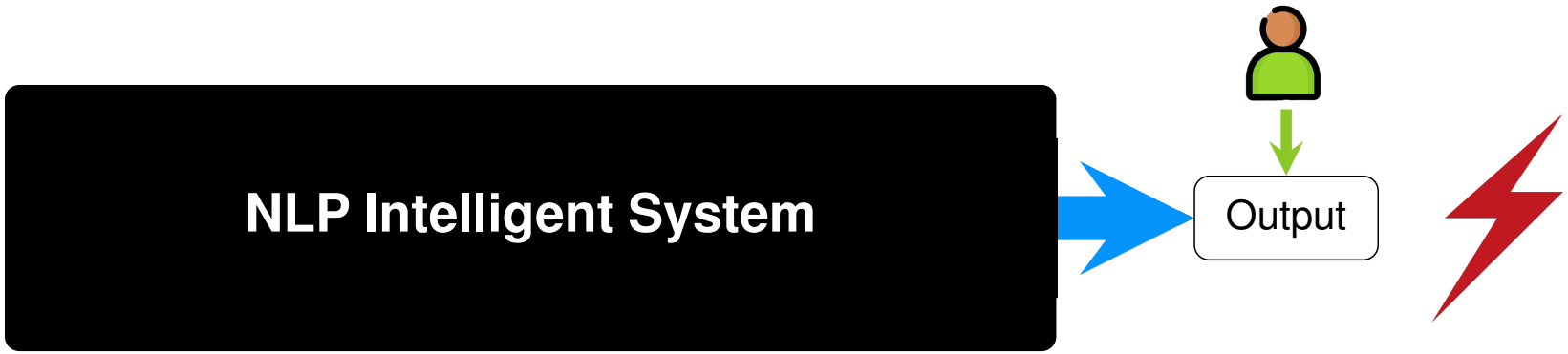


NLP Systems vs. Friedman and Nissenbaum (1996)



NLP Systems vs. Friedman and Nissenbaum (1996)

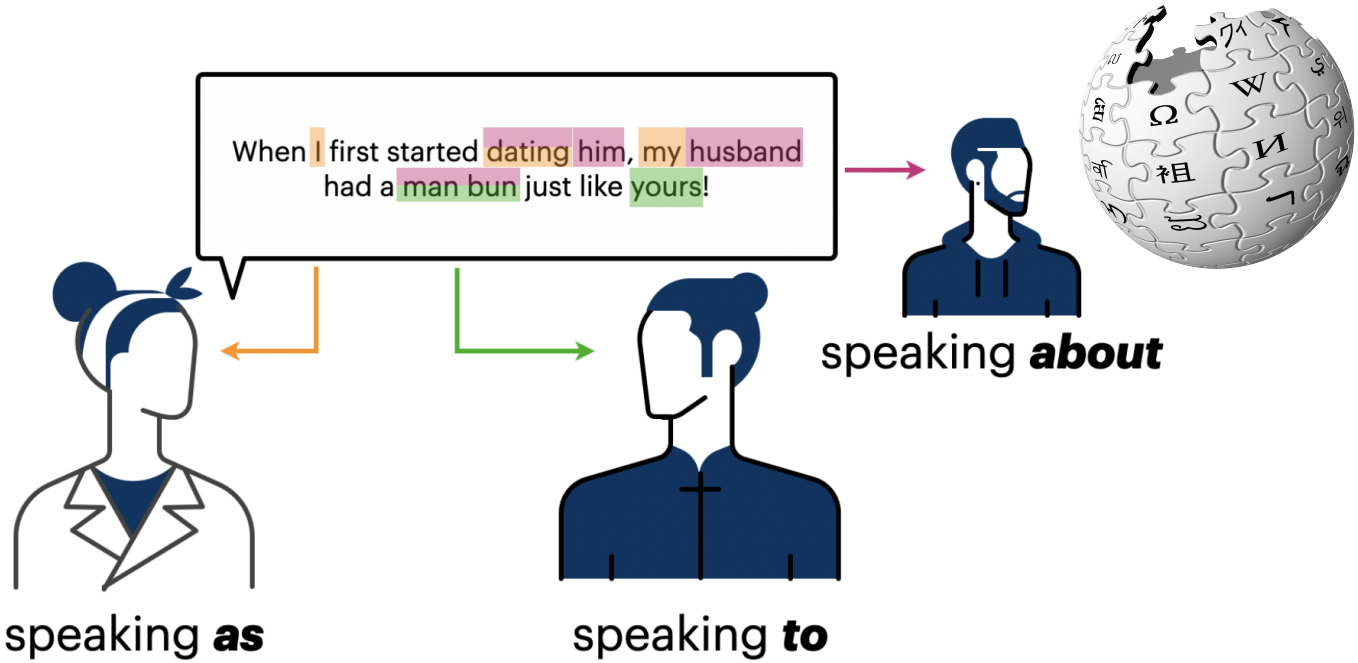
NLP Systems vs. Friedman and Nissenbaum (1996)



- No impact from underrepresented populations (Bender et al., 2021)
- Models reflect demographic imbalances (Hovy and Yang, 2021)
- Speech recognizers with lower accuracy for female voices (Tatman, 2017)

Preexisting Bias

Gender-related Inequalities in Primary Data



Framework for Gender Bias by Dinan et al. (2020)

Assessing Gender Bias in Wikipedia: Inequalities in Article Titles

Agnieszka Faleńska

Özlem Çetinoğlu

Wikipedia

- Commonly used source of training data for NLP models
  (Devlin et al., 2019; Webster et al., 2019; Zeldes, 2017)
- Inequalities in representations of genders
  (Callahan and Herring, 2011; Wagner et al., 2015; Schmahl et al., 2020)
- Focus on biographies

Are Inequalities Present only in Biographies?

England national football team

From Wikipedia, the free encyclopedia

This article is about the men's team. For the women's team, see England women's national football team.

The **England men's national football team**, and informally known as the **Three Lions**, has

England

England women's national football team

From Wikipedia, the free encyclopedia

"Lionesses" redirects here. For other uses, see Lioness (disambiguation).

The **England women's national football team** has been governed by the Football Association (FA)

England

RQ1 How frequently Wikipedia titles describe concepts in an asymmetrical way?

RQ2 Which domains?

Example from Perez (2019)

Data

- Wikipedia in: Turkish, English, German, Polish
- Filters and heuristics

  Step 1 Filtering Gender-Related Titles
  Step 2 Assigning Meta-Categories
  Step 3 Grouping Concept-related Titles

Step 1: Filtering Gender-Related Titles

Women articles:

- Human female sexuality
- Women in Islam
- ...

Men articles:

- Argentina men's national softball team
- List of male jazz singers
- ...

Step 1: Result's Size

	English	German	Polish	Turkish
TOTAL	75310	15035	13562	3985
%WIKIPEDIA	1.23%	0.67%	0.98%	1.04%

Step 2: Assigning Meta-Categories

Three main meta-categories:

- SPORTS – sports teams or events
- LISTS – listings of people or organizations
- SOCIAL – history, awards, gender issues, etc.

Step 2: Assigning Meta-Categories

Women articles:

- Human female sexuality – SOCIAL
- Women in Islam – SOCIAL
- ...

Men articles:

- Argentina men's national softball team – SPORTS
- List of male jazz singers – LISTS
- ...

Step 3: Grouping Concept-related Titles

SOCIAL

WOMEN: Human female sexuality
MEN: Human male sexuality
GENERIC: Human sexuality

SOCIAL

WOMEN: Women in Islam
MEN: –
GENERIC: Islam

NAMES

WOMEN: –
MEN: List of male jazz singers
GENERIC: –

Which Groups Describe Concepts Asymmetrically?

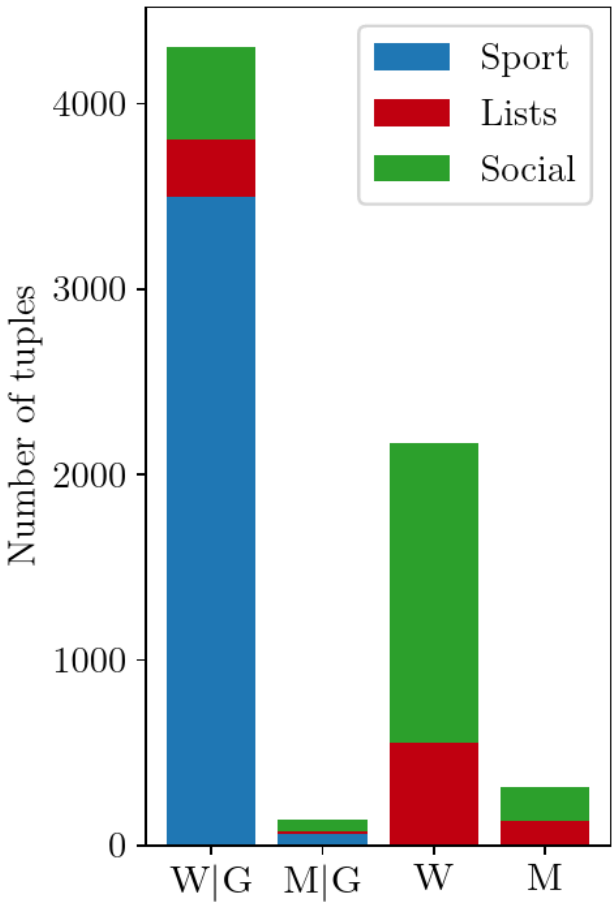
	LISTS
W|G	WOMEN: List of Albanian women writers MEN: – GENERIC: List of Albanian writers

	SOCIAL
M|G	WOMEN: – MEN: Men's health in Australia GENERIC: Health in Australia

	SOCIAL
W	WOMEN: Violence against women in Guatemala MEN: – GENERIC: –

	LISTS
M	WOMEN: – MEN: List of male jazz singers GENERIC: –

RQs: How Frequently Wikipedia Titles Describe Concepts Asymmetrically? Which Domains?

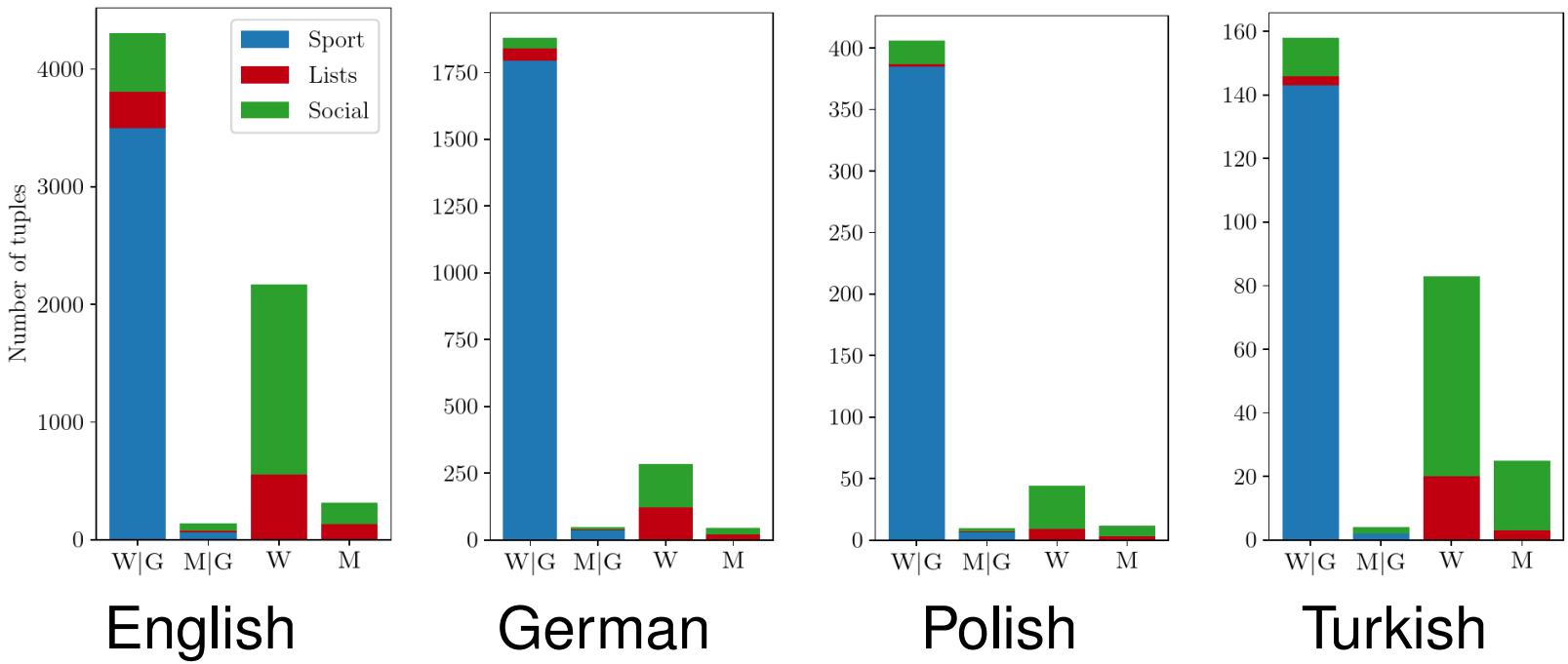


- majority include WOMEN articles
- W|G and SPORTS most frequent

→ England women's national football team

- then SOCIAL, especially in W

Frequency of Asymmetrical Groups – across Languages

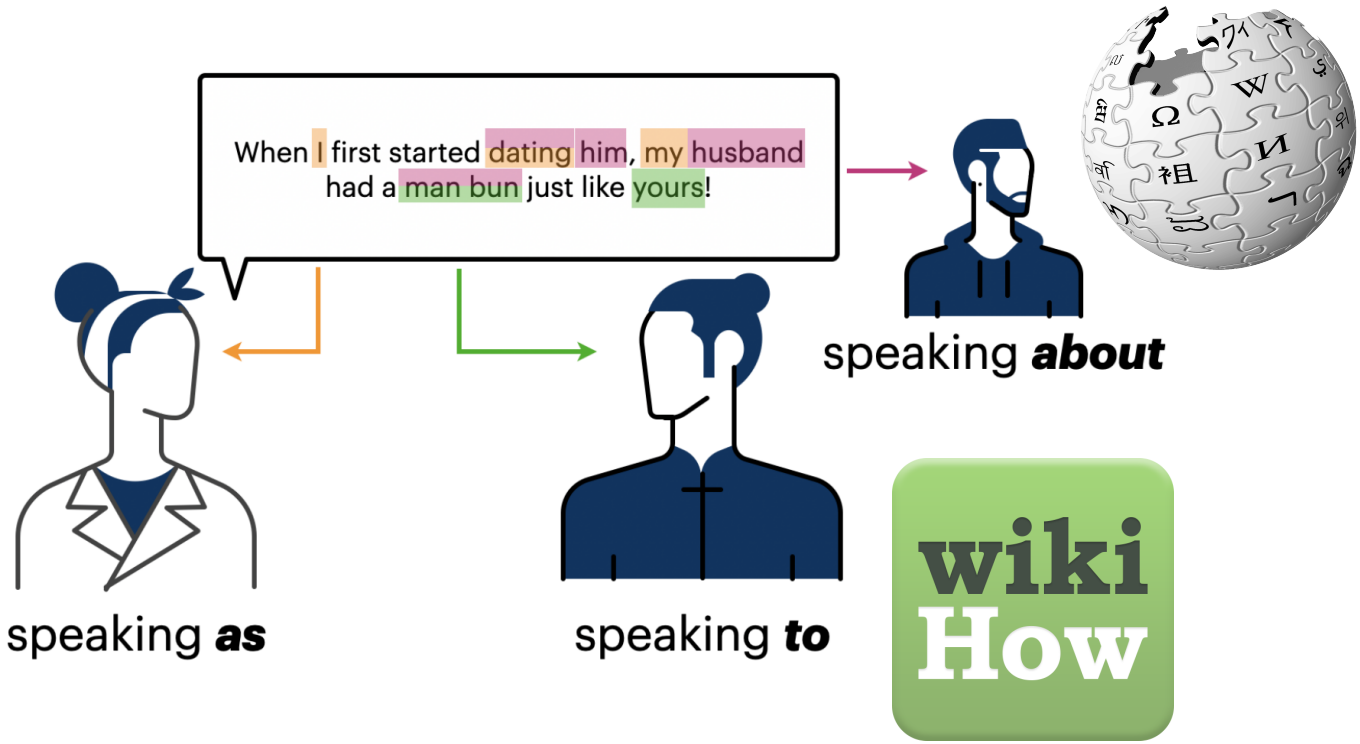


- The same pattern across languages

Takeaways: Inequalities in Wikipedia

- Gender-related inequalities in Wikipedia extend beyond biographies
- Systematic asymmetries in article titles
- Present especially in sports and social issues

Gender-related Inequalities in Primary Data



Framework for Gender Bias by Dinan et al. (2020)

How-to Guides for Specific Audiences: A Corpus and Initial Findings

Nicola Fanton

Agnieszka Faleńska

Michael Roth

How to Repair a Flat Tire

Inflate the tire. In order to find a leak the tire must be properly pressurized. You should inflate your tire with air until it reaches the appropriate pressure (measured in psi) specified in your vehicle's service manual.

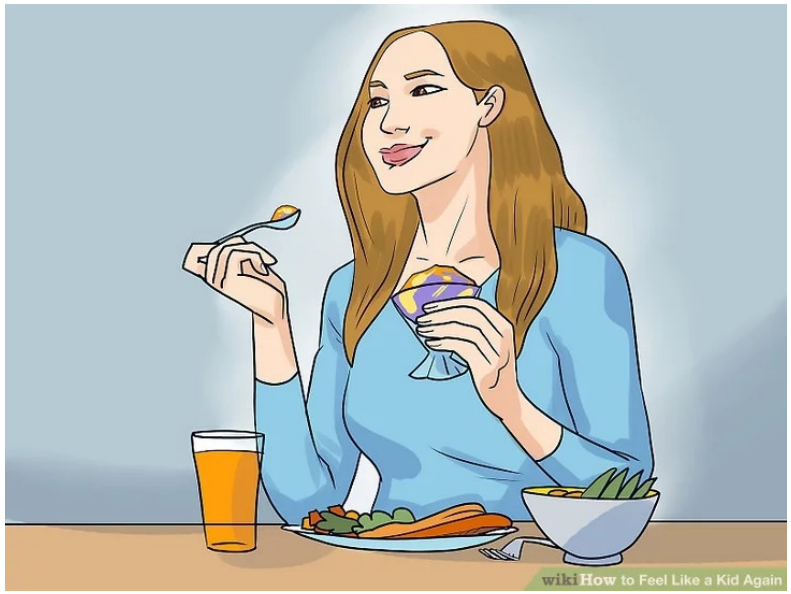
Visually inspect the tire. Before moving on to more time consuming techniques, you should take a moment to look at your tire. If you notice any holes, cuts, or objects protruding from tire then you have found your leak.

WikiHow

- Collaboratively edited online platform for instructional texts
- Niche topics and articles for minority groups
- Prominent data source for a variety of NLP tasks
  (Koupaee and Wang, 2018; Zhang et al., 2020; Cai et al., 2022)
- No previous work on biases

Guides on How to Live and Behave

How to Feel Like a Kid Again



Break a few rules, within reason. As adults we often feel like we have to follow rules all the time, but children are often more adventurous.

Climb a tree. The pride of accomplishment that comes from climbing a tree and the sense of exhilaration that you feel when sitting up high will take you back to a simpler time.

Audience-related Differences

Act Like a Kid Again (Boys)

Eat your childhood favorite food. Recollect every snack, chocolates, ice cream, candy bars, cotton candy and everything that you loved as a kid or would make you feel pampered.

Act Like a Kid Again (Girls)

Eat well and exercise, but don't obsess about your body. Be healthy without stressing too much about it. (...) go for lots of fruits and veggies. And even though kids love sugar, don't eat too much of it!

RQ How do wikiHow guides differ when they aim different audiences?

Data: Find Audience-related Guides

- Start from WIKIHOWTOIMPROVE (∼250k guides)

(Anthonio et al., 2020)

- Search for audience indicators:
  Act Like a Kid Again (Boys)

Girls	370	Guys	35
for Girls	284	for Women	35
for Kids	182	Women	34
Kids	114	UK	34
Teens	110	for Men	31
Teen Girls	100	Christianity	31
for Teens	73	Men	29
USA	49	for Beginners	29
for Guys	42	Boys	25
Windows	38	Teenage Girls	25

Data Size

- Groups: Women, Men, Teens, Kids
  → although more Girls and Boys than Women and Men

- Total 2k articles

	Kids	Teens	Women	Men
Articles	499	411	993	209
Sentences per article	29	43	40	50
Words per article	352	544	509	682

Case Study: Guides on How to Live and Behave

How to Be …

How to Be Smart in School (Girls)

How to Be Photogenic (Men)

Step 2: Results

How to Be ...

Kids vs. Teens

- good, comfortable, safe

How to Be Good With Money (for Kids)
How to Be a Good Friend (Teens)

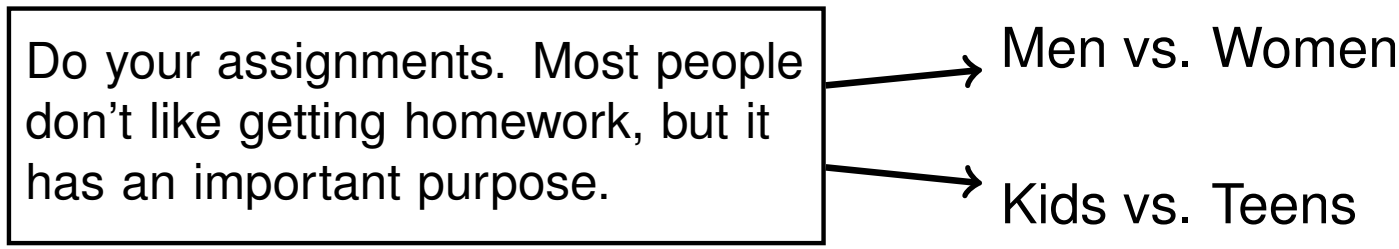
Women vs. Men

- cute, popular
- cool, more

Women:	Be Popular and Athletic Be Cute at School
Men:	Be Cool in High School Be More Physically Attractive

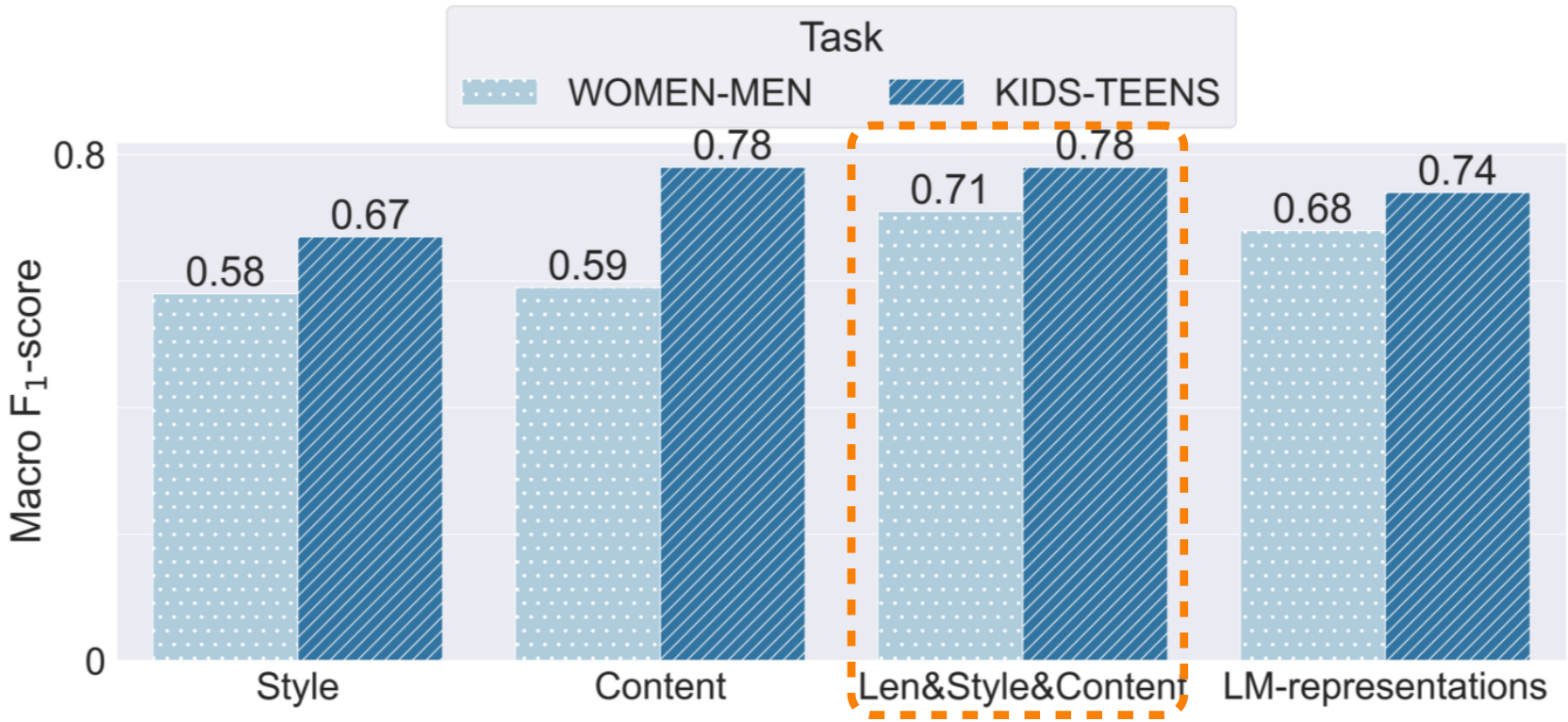
Generalization to All Guides

- Binary classification task



- Features
  - length
  - style (Sari et al., 2018)
  - content (uni-grams and bi-grams)
  - RoBERTa (Liu et al., 2019)

Can we Automatically Predict Audiences?



- Kids vs. Teens is an easier task
- All features are helpful

RQ: How do the Guides Differ?

- 10 most predictive features (highest weights)
- Kids vs. Teens
  - → audience indicators: kid, teen

Make Money (Kids)

> . . . even if you're a kid, there are ways to bank a few extra bucks.

Stay Active After School (Teens)

> When you're a teen with a busy schedule, it can be difficult to find time to be active.

RQ: How do the Guides Differ?

Women vs. Men

- Women: stereotypes
  - → *cute, makeup, skirt, outfit*
  - → 'cute' can be used pejoratively as a form of social control

    (Talbot, 2019)

- Women: negations
  - → *hadn't, wasn't*
  - → negations serve a stereotype-maintaining function

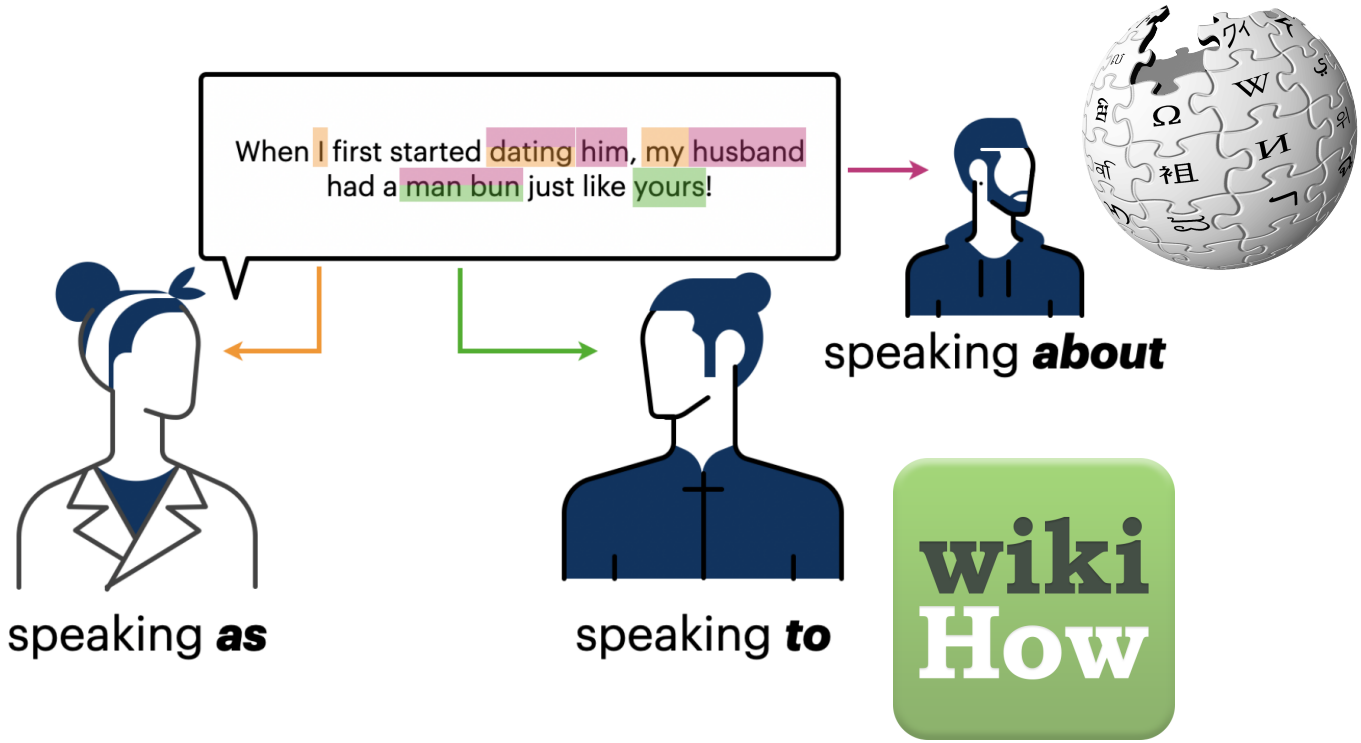
    (Beukeboom et al., 2010, 2020)

- Men: pronouns
  - → heteronormative assumption: *hers*
  - → characteristics of gender-inclusive language: *theirs*

Takeaways: Inequalities in WikiHow

- Disparate standards and treatment
- Inequalities in guides for different audiences
  - who is being more instructed
  - the number of instructions
- Subtle stereotypes
  - ‘cute’, negations

Gender-related Inequalities in Primary Data



Framework for Gender Bias by Dinan et al. (2020)

Self-reported Demographics and Discourse Dynamics in a Persuasive Online Forum

Agnieszka Faleńska

Eva Maria Vecci

Gabriella Lapesa

/r/ChangeMyView subreddit (CMV)

CaptainMalForever • 2 days ago

First, outdoor cats have a large territory not because they need the exercise or the space to be happy, but they need the space in order to get enough food to eat. I provide my cat all the food he needs and thus, he needs less territory.

Delta(s) from OP - Fresh Topic Friday

squidkyd • 2 days ago

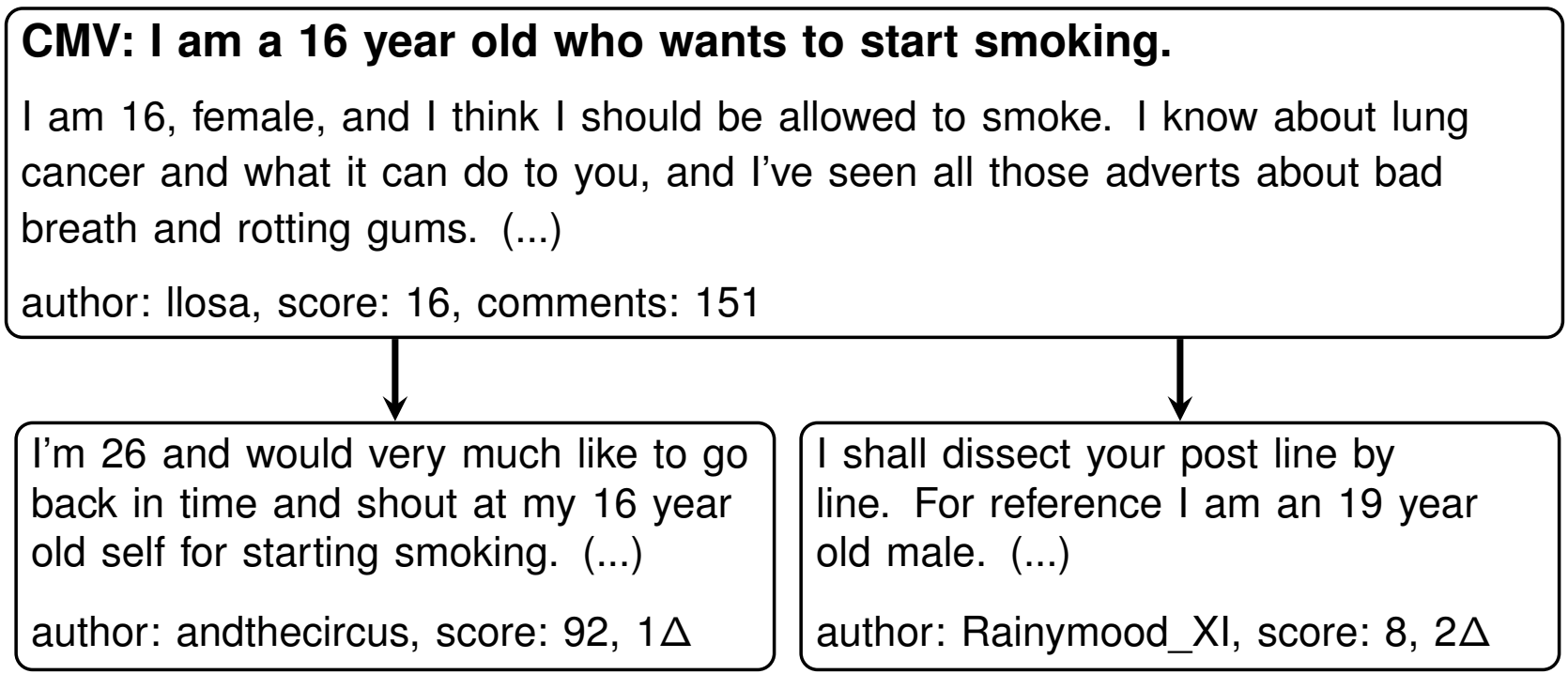
My cat escaped my house in July of 2020

A repairman left a door open and she wandered outside. You could say she was “advocating for herself” and following her instincts.

/r/ChangeMyView subreddit (CMV)

- Online forum targeted at persuasion
- Crucial for research on argument mining
  (Morio et al., 2019; Egawa et al., 2020; Dayter and Messerli, 2022)
- Bias in argument mining
  (Spliethöver and Wachsmuth, 2020; Manzoor et al., 2022)
- No focus on the influence of speakers demographics on CMV discourse

Speakers in ChangeMyView



Explicit Gender Disclosures in ChangeMyView

RQ1 When and why do people disclose their gender?

RQ2 How does the forum community react to gender disclosures?

RQ3 Are there stylistic differences related to authors gender?

Data: Find CMV Posts with Gender Disclosures

- Start from CMV (∼20k posts and 1M replies)

(Tan et al., 2016)

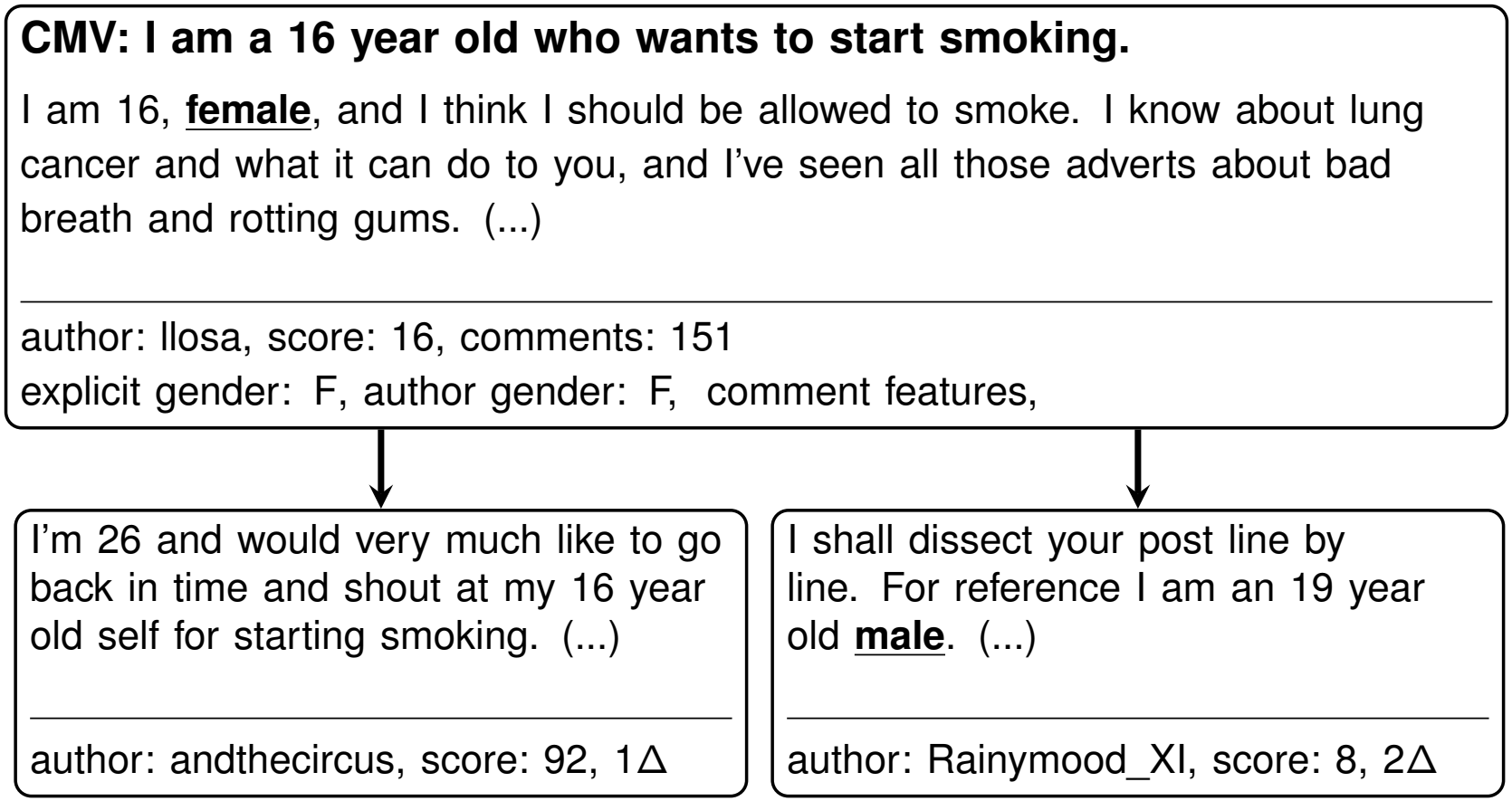
- Filter all posts:

- Two human annotators
  - quoted speech: *They don't want to hear 'I'm nonbinary*
  - hypothetical situations: *If I am female*

Data: Initial Size

	Posts	Replies	Discussions	Authors
total	396	3,235	1,812	2,456
male	299	1,953	1,357	1,640
female	89	961	693	674
other	8	321	175	158

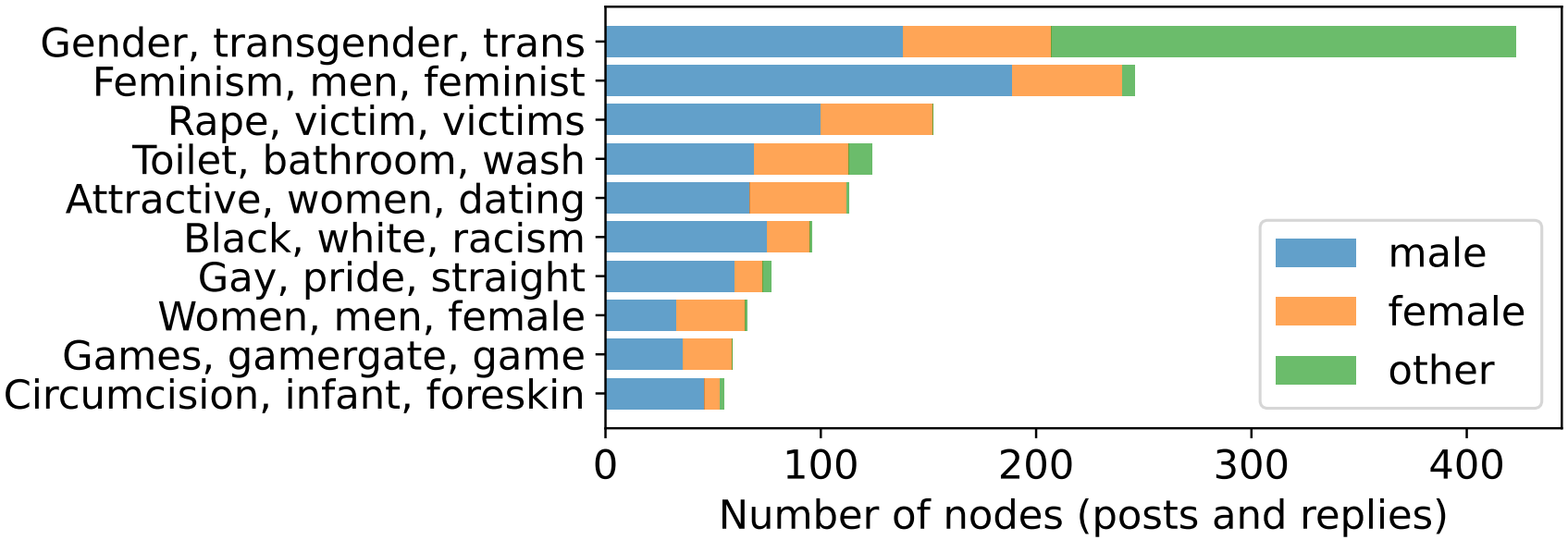
Data: Annotate Additional Features



- Extended author gender information

	Posts	Comments
male	2,253	227,261
female	396	53,042

RQ1: When Do People Explicitly Mention their Gender?



- Topics relate to gender or situations in which the set of experiencers is unbalanced
- All topics addressed by males and females

RQ1: Why Do People Explicitly Mention their Gender?

- Establishing the speaker's credibility (Falk and Lapesa, 2022)

> **CMV: If I was raped or sexually assaulted I probably wouldn't report it.**
>
> Preface: I'm a 23 year old **woman**. I believe that my life would be far worse off (...)

- An implicit rebuttal (Habernal and Gurevych, 2017)

> **I think that feminism currently uses hate speech as a way to advance its goals. In fact, this attitude hurts the advancement of women. CMV**
>
> I'll start by saying I'm 26 **male**. (...)

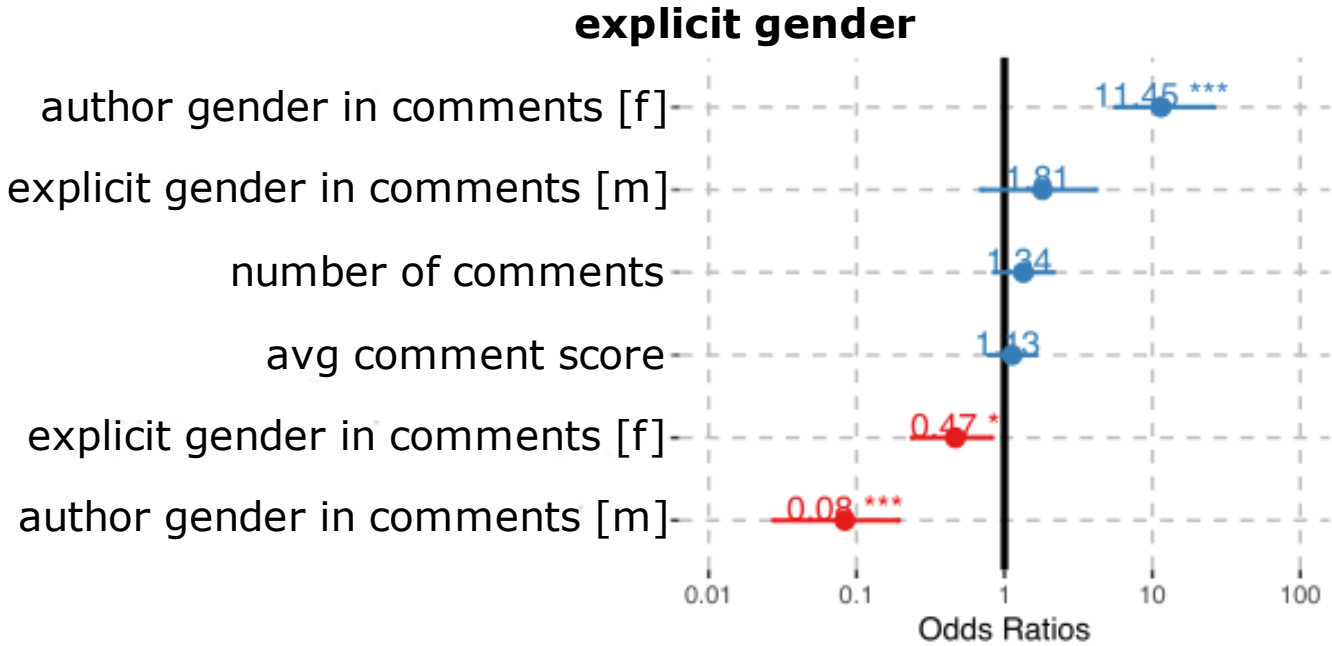
- A combination of a credibility with a concession (Musi, 2018)

> **CMV: I think the feminist movement was detrimental to society.**
>
> Firstly I'd just like to point out that I am **female**. Secondly I'd like to clarify that I'm all for equality between all people. However, (...)

RQ2: Who Reacts to Explicit Gender Mentions?

- Logistic regression model:
  - explicit gender as dependent variable
  - all features as independent variables (no style features)
- Standardized beta values for significant ($Pr(|z|) < 0.05$) terms
- Pseudo $R^2 = 52\%$

RQ2: Who Reacts to Explicit Gender Mentions?

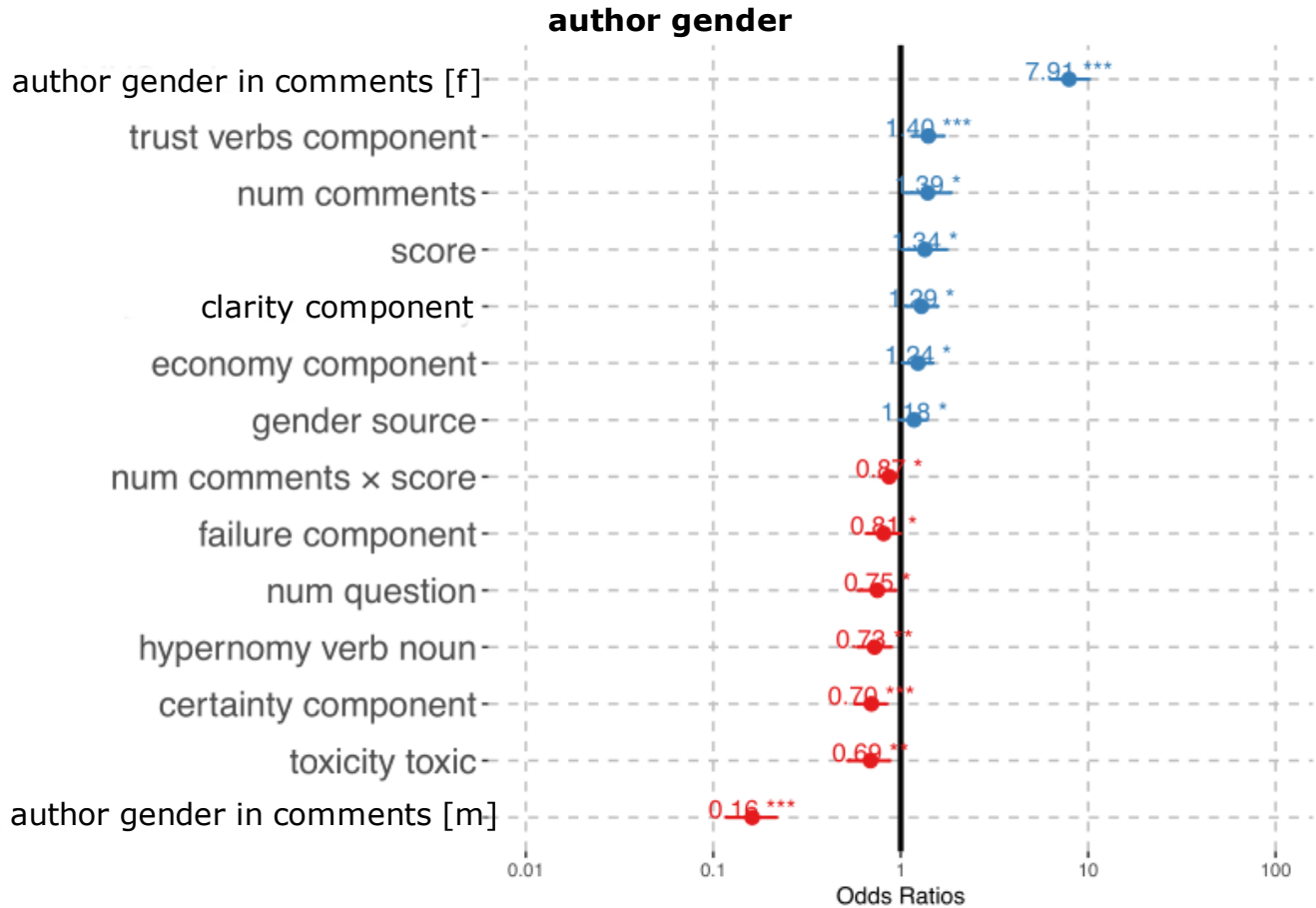


- Explicit mentions strongly define the commenting population
- Argument-counterargument discourse
- Quantity and quality of interaction higher for female posts

RQ3: Are there Style Differences?

- Logistic regression model
  - author gender as dependent variable
  - include style features in independent variables
- Standardized beta values for significant ($Pr(|z|) < 0.05$) terms
- Pseudo $R^2 = 61\%$

RQ3: Are there Style Differences?

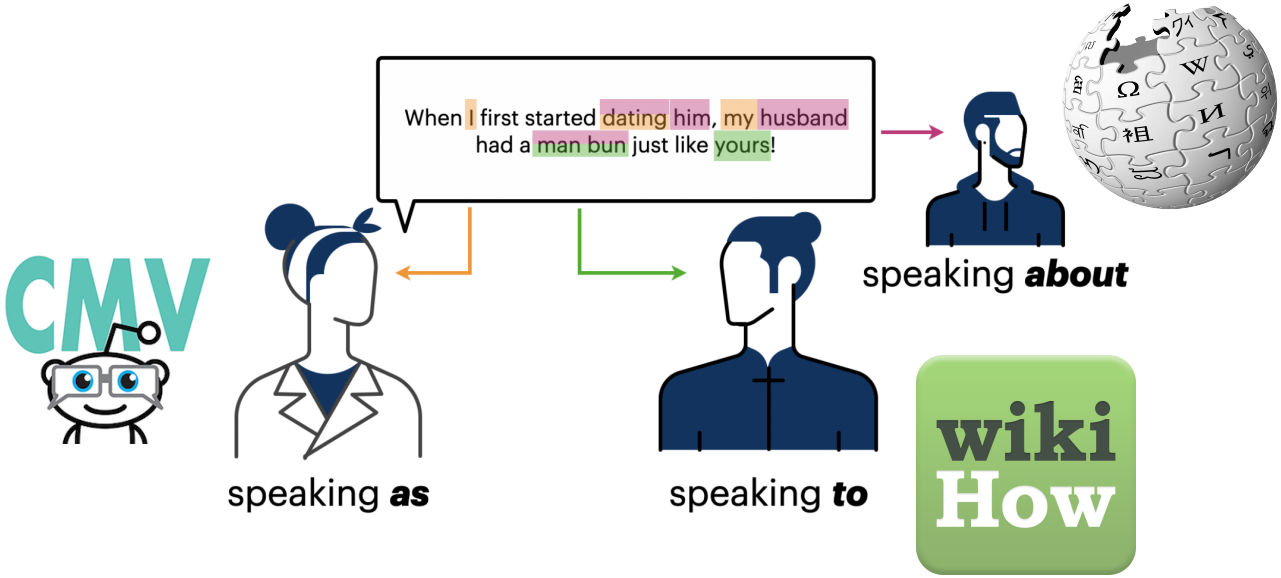


- gender-dependent engagement stays, but not in explicit mentions
- style-related differences (Morales Sánchez et al., 2022)

Takeaways: Inequalities in CMV

- Explicitly mentioning gender has a persuasive function
  - Reddit is male-skewed
- Reaction imbalances
  - comments from users of the same gender
  - explicit gender mentions from users of the opposite gender
- Gender-related style differences

Gender-related Inequalities in Data



- Gender-related inequalities and subtle biases
- Need for a discussion within the communities
  - → England men's national football team
  - → List of female jazz singers
  - → Women's suffrage → Men's suffrage?

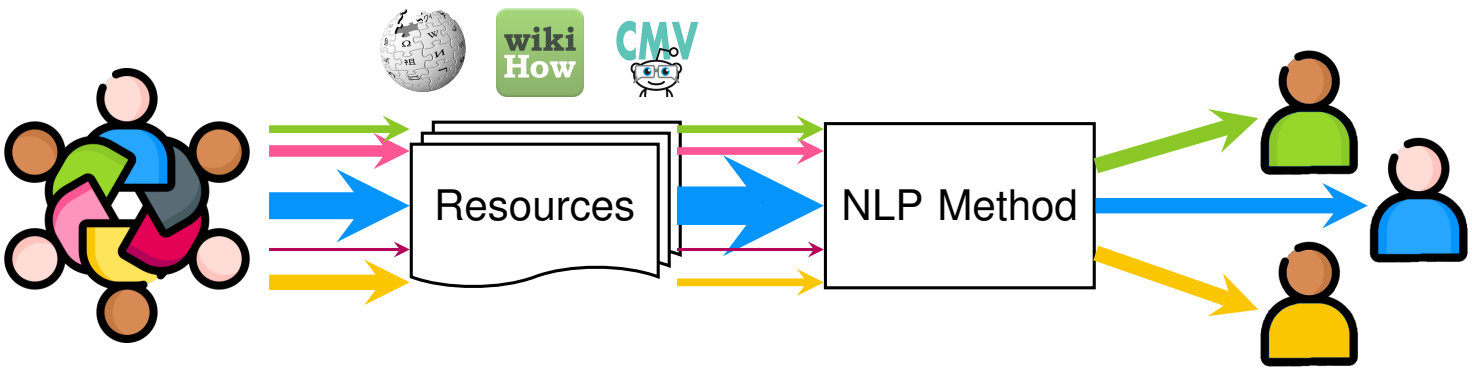
Framework for Gender Bias by Dinan et al. (2020)

How to Remove Biases from Data?

- Language is inherently biased
- *"(...) records human interpretations that are situated in a specific time, place, and worldview"*

(Haraway, 1988; Havens et al., 2020)

Steps towards Bias-aware NLP Systems



1. Diagnose biases in primary data
2. **Understand their impact on NLP methods**
3. Make biases transparent in NLP architectures
4. Assist end users in developing NLP solutions

Future Directions

Understanding Impact of Subtle Biases

Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST

IRIS

DANIS: Diversity-Aware NLP Intelligent Systems

Quy Nguyen
Pema Gurung Hongyu Chen

DFG (under review)

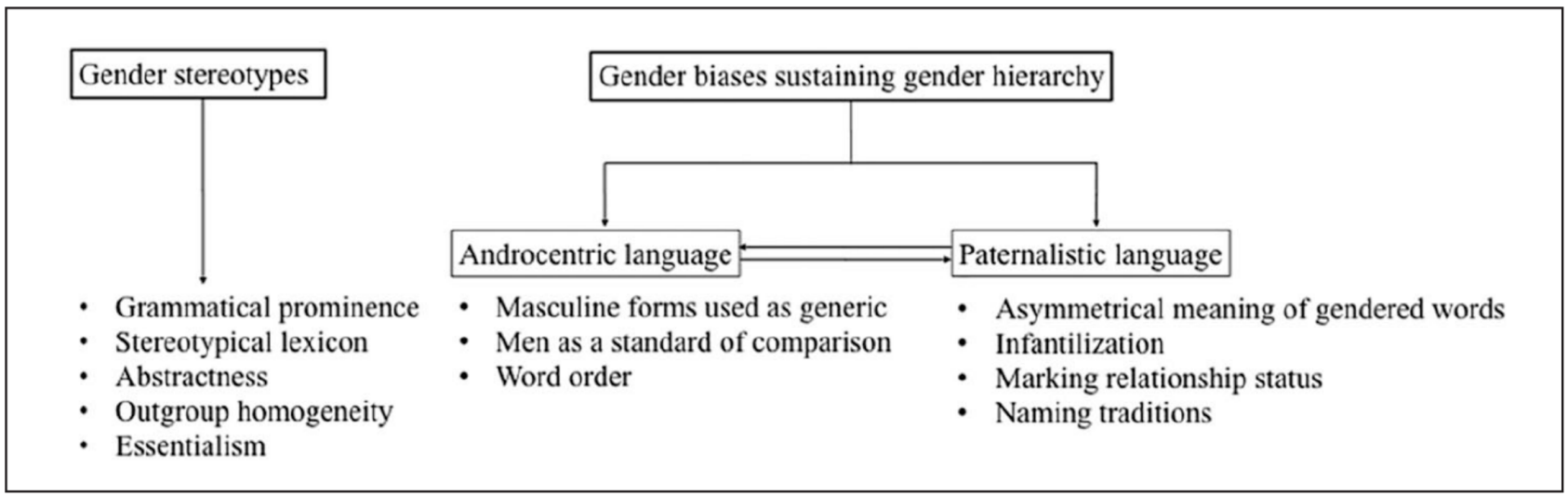
SEAL: Use and Effects of Androcentric Language

Raphael Heiberger

Androcentric Language (AL)

Linguistic constructions reflecting gender-related power asymmetries

(Friedman and Nissenbaum, 1996; Formanowicz and Hansen, 2021)



Categories of Androcentric Language

1. Male generics
   - *beim Arzt*
   - *byłem u lekarza*

   (Moulton et al., 1978)

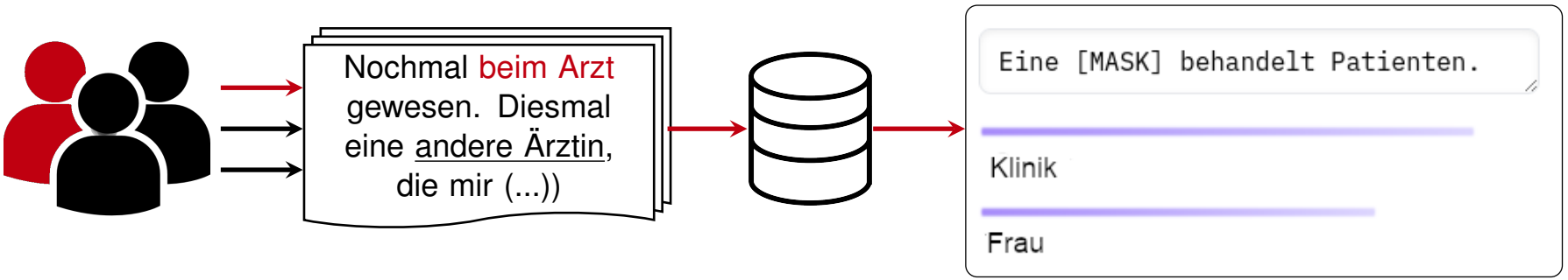
2. Mentioning men first
   - *Ärzte und Ärztinnen*
   - *Panowie i Panie*

   (Benor and Levy, 2006)

3. Comparing women to men
   - *women earn less than men*
   - *10 zaskakujących rzeczy, które kobiety robią znaczniej lepiej od mężczyzn*

   (Bruckmüller et al., 2012)

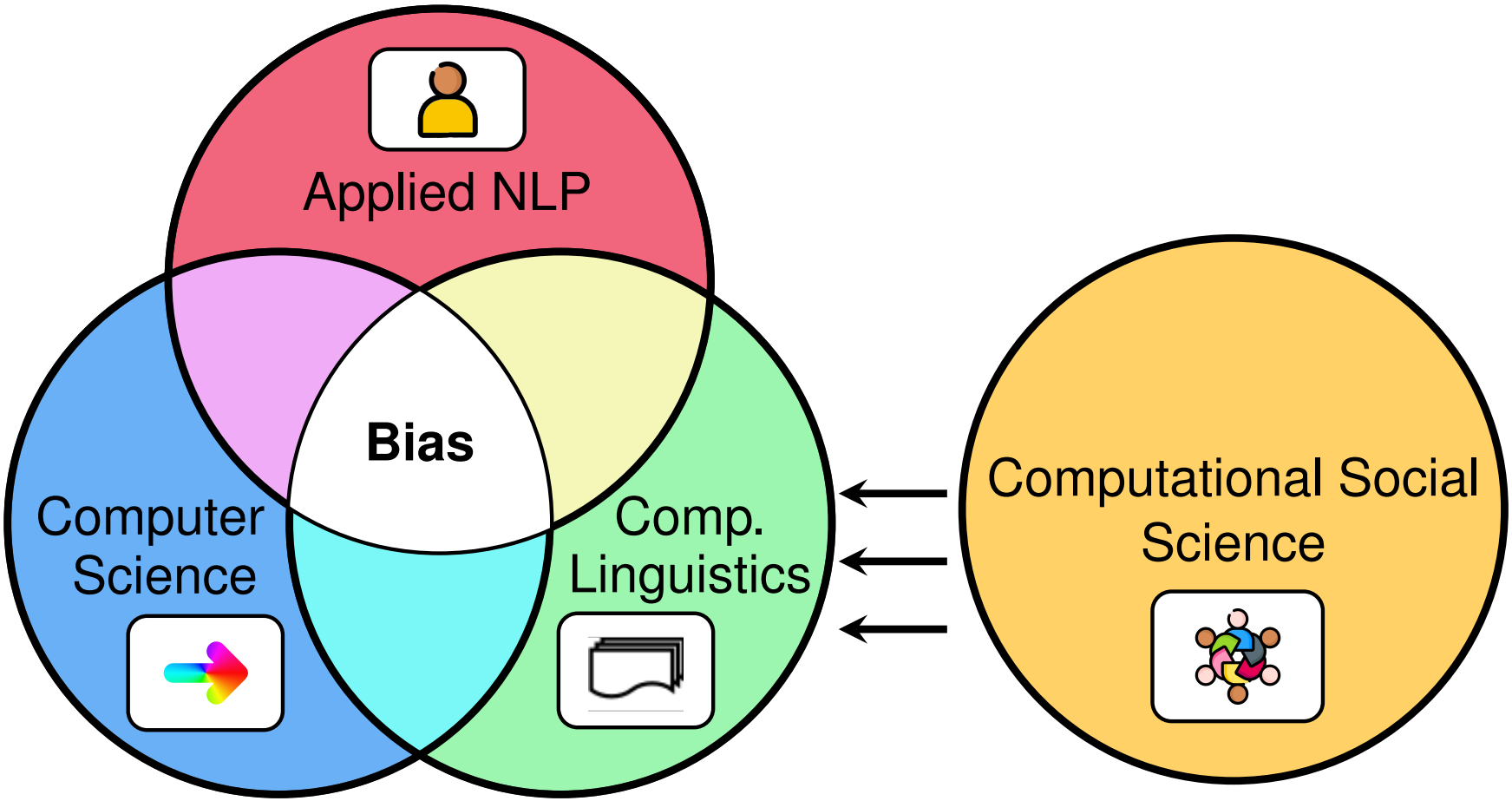
ASEAL: Use and Effects of Androcentric Language



- Which linguistic features characterize AL?
  - → English, German
- Who uses androcentric language? Which domains?
  - → public figures
- What is the impact of AL on NLP models?
  - → language modeling, question answering
- Facilitate AL-aware design
  - → tools for modeling AL, data and model statements

Take-home Message

Steps Towards Bias-Aware NLP Systems




Thank you

Universität Stuttgart

Dr. Agnieszka Faleńska
Interchange Forum for Reflecting on Intelligent Systems

e-mail agnieszka.falenska@iris.uni-stuttgart.de
telefon +49-711-685 813 58

Talita Anthonio, Irshad Bhat, and Michael Roth. 2020. wikiHowToImprove: A resource and analyses on edits in instructional texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pages 5721–5729. https://aclanthology.org/2020.lrec-1.702.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. pages 610–623.

Sarah Bunin Benor and Roger Levy. 2006. The chicken or the egg? A probabilistic analysis of English binomials. *Language* pages 233–278.

Camiel J. Beukeboom, Christian Burgers, Zsolt P. Szabó, Slavica Cvejic, Jan-Erik M. Lönnqvist, and Kasper Welbers. 2020. The negation bias in stereotype maintenance: A replication in five languages. *Journal of Language and Social Psychology* 39(2):219–236. https://doi.org/10.1177/0261927X19869759.

Camiel J. Beukeboom, Catrin Finkenauer, and Daniël H. J. Wigboldus. 2010. The negation bias: When negations signal stereotypic expectancies. *Journal of Personality and Social Psychology* 99(6):978–992. https://doi.org/10.1037/a0020861.

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*. Association for Computational Linguistics, Online, pages 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485.

Susanne Bruckmüller, Peter Hegarty, and Andrea E Abele. 2012. Framing gender differences: Linguistic normativity affects perceptions of power and gender stereotypes. *European Journal of Social Psychology* 42(2):210–218.

Pengshan Cai, Mo Yu, Fei Liu, and Hong Yu. 2022. Generating coherent narratives with subtopic planning to answer how-to questions. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), pages 26–42. https://aclanthology.org/2022.gem-1.3.

Ewa S. Callahan and Susan C. Herring. 2011. Cultural Bias in Wikipedia Content on Famous Persons. *Journal of the American society for information science and technology* 62(10):1899–1915. https://doi.org/https://doi.org/10.1002/asi.21577.

Daria Dayter and Thomas C Messerli. 2022. Persuasive language and features of formality on the r/changemyview subreddit. *Internet Pragmatics* 5(1):165–195.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pages 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, pages 314–331. https://doi.org/10.18653/v1/2020.emnlp-main.23.