How to study gender in large corpora of online data?

A sociolinguistic approach with the RedditGender corpus

Café TAL – 02/02/2022

Marie Flesch – ATILF, team “Didactique des langues et sociolinguistique”

UNIVERSITÉ DE LORRAINE

ATILF-CNRS | École doctorale SLTC (Sociétés, Langages, Temps, Connaissances)

Thèse présentée et soutenue publiquement en vue de l'obtention du titre de docteure de l'université de Lorraine, Mention « Sciences du langage », par Marie Flesch, le 16 décembre 2020

lol thats how reddit talks ;) : le site américain Reddit comme espace de variation de l'anglais.

Étude de corpus intersectionnelle et quantitative d'usages non standard, au prisme du genre, de l'âge et de l'ethnicité

Gender and lexical type frequencies in Finland Twitter English

Steven Coats

He Votes or She Votes? Female and Male Discursive Strategies in Twitter Political Hashtags

Evandro Cunha[1,2*], Gabriel Magno[1], Marcos André Gonçalves[1], César Cambraia[2], Virgilio Almeida[1]
1 Computer Science Department, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, 2 College of Letters, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

Lexi Webster*

"I am I": Self-constructed transgender identities in in communicatio

GENDERED C ITUALS ON EFFECTIV ON MORE THA WHAT ONE IS

ELSEVIER

Author gender identification from text☆

Na Cheng, R. Chandramouli, K.P. Subbalakshmi*
Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboke

'Bad language' in the Nordics: profanity and gender in a social media corpus

Steven Coats

Gender Classification using Tw

Publisher: IEEE | Cite This | PD

Journal of Sociolinguistics 18/2, 2014: 135–160

Gender identity and lexical variation in social media[1]

David Bamman,[a] Jacob Eisenstein[b] and Tyler Schnoebelen[c]

ender differences i change commu on Twitte

Kim Holmberg
Department of Organization Sciences, VU University

hy be normal?": Language and identity practices in a community of nerd girls

MARY BUCHOLTZ

RCH ARTICLE

men are Warmer but No Less n Men: Gender and Language

[Submitted on 13 Aug 2020]

Exploration of Gender Differences in COVID-19 Discourse on Reddit

Jai Aggarwal, Ella Rabinovich, Suzanne Stevenson

Gendered Tweets: Computa Gender Differenc

Lingshu Hu, Michael

DOI:10.4236/ijis.2012

Men Are from Ma Examining Gender on Soci

arXiv.org > cs > arXiv:1805.03122 Search... Help | Advanced

Computer Science > Computation and Language

[Submitted on 8 May 2018]

Bleaching Text: Abstract Feature Gender Prediction

Rob van der Goot, Nikola Ljubešić, Ian Matroos, M

H. Andrew Schwartz[2], Marg
vid Stillwell[4], Lyle H. Ungar
nsylvania, Philadelphia, Penns

Gender and Cross-Cultural Social Media Disclosures of

unmun De Choudhury Sanket S. Sharma

nal of Sociolinguistics 10/4, 2006: 439–459

Gender and ge

Susan C. He
Indiana Univ

OPEN ACCESS Freely available online PLOS ONE

Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach

H. Andrew Schwartz[1,2*], Johannes C. Eichstaedt[1], Margaret L. Kern[1], Lukasz Dziurzynski[1],

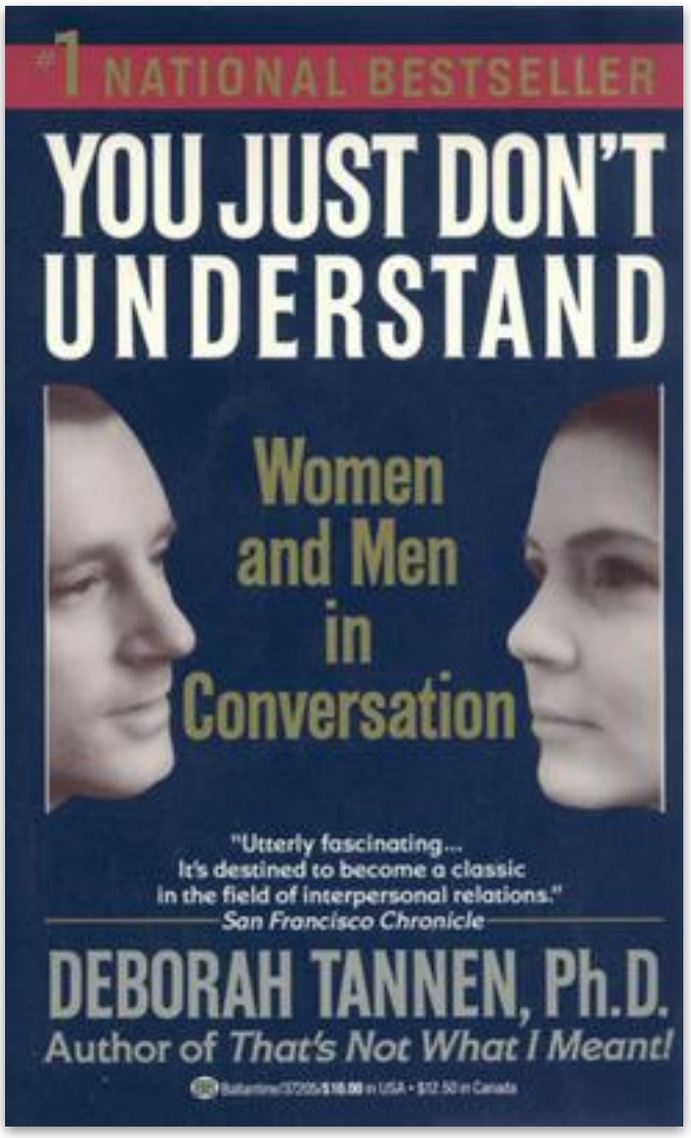
ASSESSING GENDE AUTHENTICITY I COMPUTER-MEDIATE LANGUAGE US idence From an Identity Gam

Why do we study gender?

Why do we study gender?

- NLP practicioners (people who develop commercial applications)
    - goal: get insight about consumers' behavior in order to serve the interest of a compagny/to build various types of applications
    - issue: gender used as a variable in machine learning systems can create "algorithmic discrimination" (Goodman, 2016)
- academic researchers in NLP & sociolinguistics
    - goal: contribute to science/knowledge
    - issue: research has social effects (when covered by the news, for instance); can reinforce sterotypes/have harmful effects

An example of the social effects of scientific research



1990

Sex Res Soc Policy (2010) 7:45–49
DOI 10.1007/s13178-010-0003-4

Young Heterosexual Men's Use of the Miscommunication Model in Explaining Acquaintance Rape

Susan Hansen · Rachael O'Byrne · Mark Rapley

What is gender?

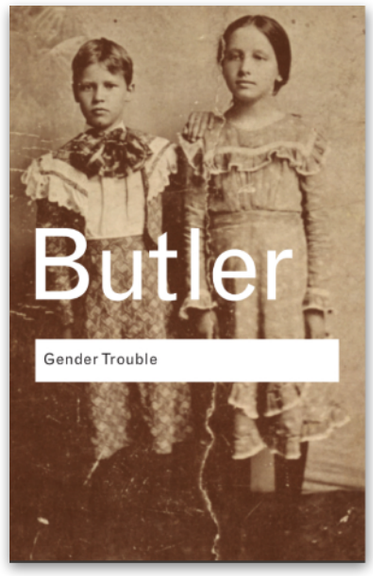
The “folk” view of gender

- “folk” view = common beliefs about gender
- conflates sex (chromosomal and biological characteristics) and gender (behaviour, appearance)
- is binary: there are only women and men
- is flawed: biological sex is not always binary
- 1.7% of babies are born with a form of intersexuality (Blackless et al. 2000)

Gender as a social construct

- Gender is not a biological concept; it is constructed.
- A process: "gender fortification" (Fausto-Sterling, 2012)

Gender as a performance



- J. Butler (1990), *Gender trouble*
- Gender is not something that we are, it is something that we do
- We create our gender identity in the way we use language (among other things)
- There is a diversity of gender identities
- Very useful when studying people who transgress the "folk" view of gender

How do we assign gender to authors?

Assigning gender: using explicit statements

- How internet users define their gender identity
- The "gold standard" for NLP research (Larson, 2017)
- Asking internet users (Finlay, 2014)
- Searching for the information in texts (manual, labor-intensive)

Assigning gender: using platforms labels

- Some social media use gender labels
- Some blogging platforms do too (Burger et al., 2011)

Assigning gender: inferring gender from user names

- A very popular approach with Twitter data (Mislove, 2011, Coats, 2017; Cunha et al. 2014)
- Automatically infers gender from first names, comparing them with a list of popular first names in a given country

Rank	Name
1	MARY
2	PATRICIA
3	LINDA
4	BARBARA
5	ELIZABETH
6	JENNIFER
7	MARIA
8	SUSAN
9	MARGARET
10	DOROTHY
11	LISA
12	NANCY
13	KAREN
14	BETTY

- But some users can misgendered, mostly women (Thelwall et al., 2018)

Is gender enough?

Is gender enough?

- Gender is only one component of your identity
- Only taking gender into account is reductive: women and men are not homogeneous groups
- Gender is the easiest variable to collect
- Age: much more challenging
- Race: Mislove et al. 2011 (database with ethnic makeup of US last names); Eisenstein, 2015 (GPS data in tweets + geographical statistics of the US census)
- Very few multifactorial studies of language and gender (Nguyen et al., 2016)

Why take gender into account at all?

Why study gender at all?

- NLP: not including gender when not a relevant variable (to avoid algorithmic discrimation) (Larson, 2017)

- Sociolinguistics: studies about gender that don't use gender as a starting point/ as a independent variable

  Bamman, Eisenstein & Schnoebelen (2014)

  Clusterered Twitter users based on their use of lexical features

  Were able to show the multifaceted nature of gendered language styles

- Other categories may be more relevant: for instance, on a web forum, moderator/ordinary user (Androutsopoulos, 2014)

My intersectional study of a corpus of Reddit comments

The PhD

- My goal: to analyze the correlations between nonstandard and innovative linguistic pratices in English & age, gender, and race on Reddit

Reddit:
a community website (not social media)
forums (=subreddits)
pseudonymous

The linguistic variables:
emoticons: :-) **women**
emoji 🤣, ❤️
letter and punctuation lengthenings: *nooooo*, *!!!!!!!!!!*) **women**
abreviations: *lol*, *idk*, *omg* **men** **women**
phonetic spellings: *gonna*, *kinda*, *ya*
apostrophe omission: *im*, *cant* **men**
lowercase instead of uppercase: *i don't know*
interjections: *duh*, *ugh*, *wow* **women**
all caps: *what REALLY helped me*
g-dropping: *doin'*, *goin* **men**

My approach: the intersectional approach

- Intersectionality: origins in Black Feminism (Crenshaw, 1989)

- When applied to linguistic research:

> "The belief that no one category (e.g. 'woman' or 'lesbian') is sufficient to account for individual experience or behavior." (Levon, 2015)

The sample

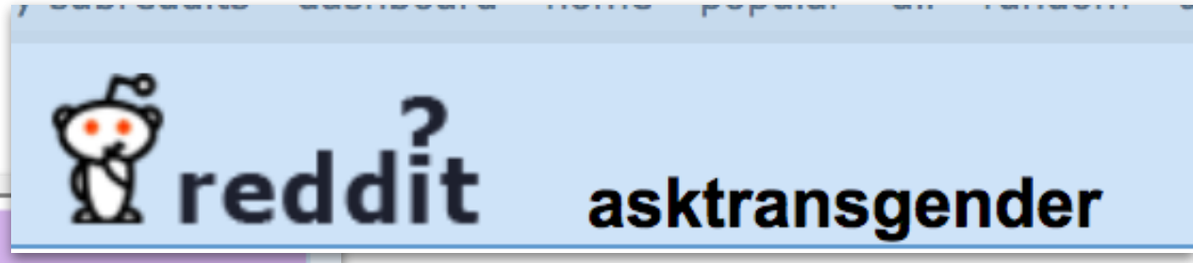
- Reddit users: mostly white, male, and young (Barthel et al. 2016)

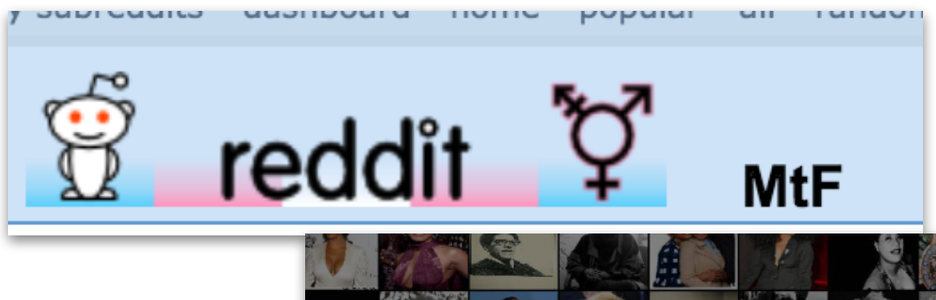
- Decision made to over-represent some categories of users (women, trans individuals, Blacks, Asians, Hispanics, etc.)

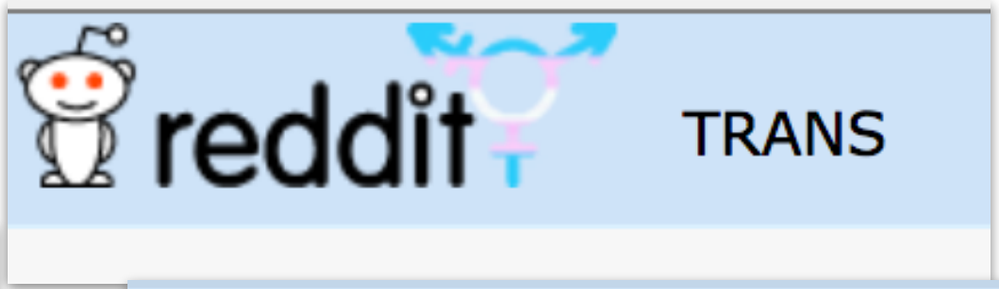
The sample

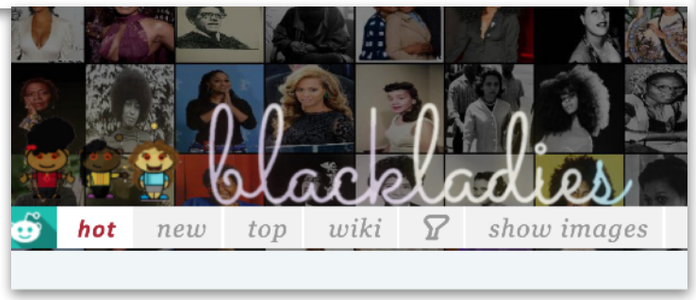

ftm

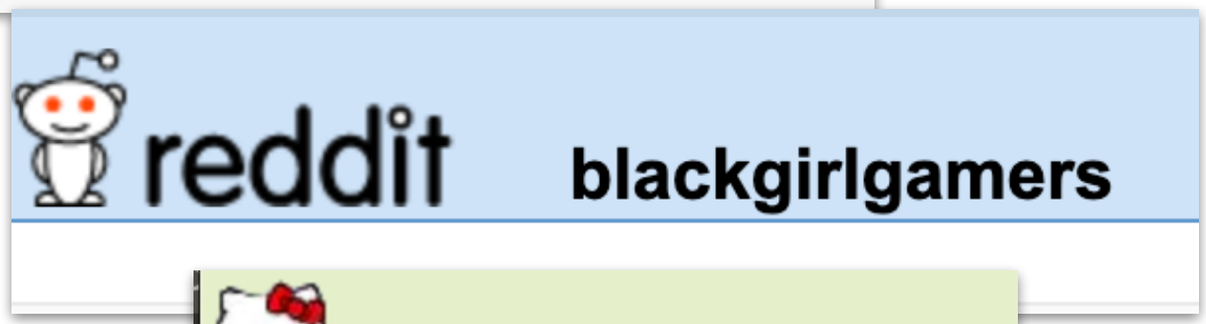

reddit
GENDERQUEER


reddit
asktransgender

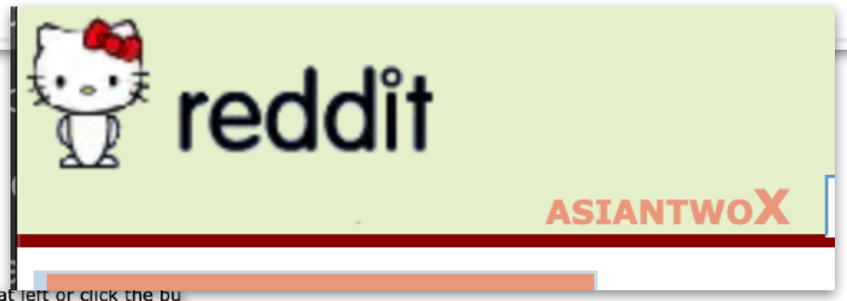

reddit
MtF


reddit
TRANS


blackladies
hot
new
top
wiki
show images


reddit
blackgirlgamers

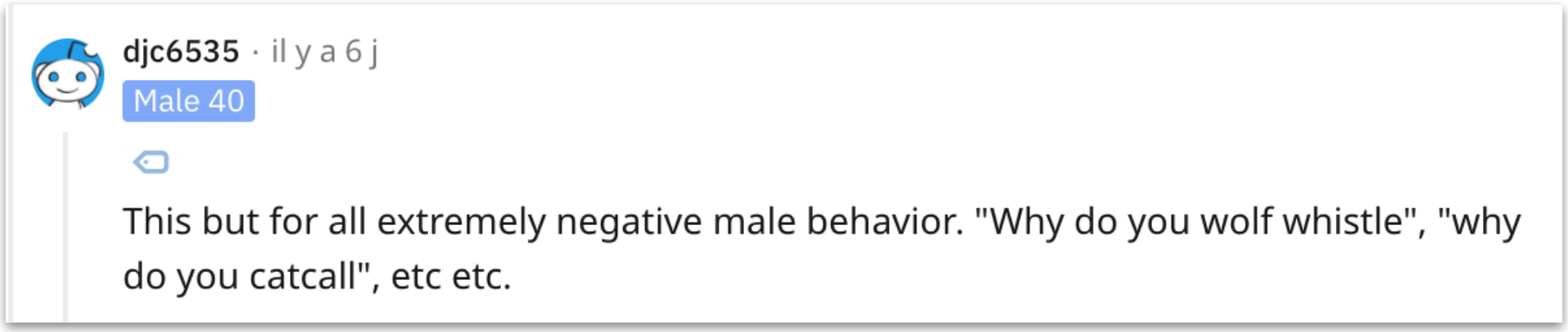

LatinoPeopleTwitter


reddit
ASIANTWOX


asianamerican
show images

Collecting sociodemographic data

- Use of “flairs”



- Are context-dependent (not used on all subreddits, vary depending on subreddits)

Collecting sociodemographic data

Collecting sociodemographic data

I'm a SAHM
I'm a black man
I'm a transwoman
I'm Hispanic
I'm a Chinese-Canadian software developer
I'm a 50s hetero male
I'm a 31 year old accountant
I'm a 23 year old trans woman
I'm a 45-yr-old female
I'm a 20s black woman in California
I'm a bisexual female
I'm Korean American
I'm a lesbian
I am a cis hetero girl
I'm a military veteran
I'm a nonbinary trans woman
I'm a straight trans guy
I'm a white girl

The RedditGender corpus

- 19.33 million tokens, 1044 internet users
- Gender: 372 cis women, 372 cis men, 100 trans women, 100 trans men, 100 nonbinary individuals
- Age: 14-20 (14.08%), 21-30 (49.52%), 30+(36.40%)
- Race: white (19.44%), Black (8.81%), Asian (6.61%), Hispanic (6.51%), unknown/ other (59.58%)

Statistical methods

- "regardless of type of regression model, no main effects model represents an intersectional approach" (Bauer, 2014)
- Multiple regression with interactions
  2 or 3 interaction terms: gender * age, gender * age *race
- Skewed data/a lot of dispersion: negative binomial and zero-inflated models

A very brief overview of results

- 5 "gendered" variables (when interaction with age not significant):

  Women: emoticons, !!!!!!!!!!!!
  Men: g-droppings (*doin'*) ; lowercase *i*; apostrophe omission

- Interaction with age nuances results

Some results: comparing cis women and men

TABLEAU 12.1 – Variables les plus fréquemment utilisées par les femme et les hommes cisgenres

Groupes d'âge	Femmes	Hommes	Pas de différence significative
Tous	**Émoticônes** (1.81) **Étirements de ponctuation** (1.57)	G-droppings (1.59) *i* minuscule (2.13) Omissions d'apostrophe (1.34)	***All caps***
14-20 ans	-	Abréviations (1.38) Graphies phonétiques (1.81)	**Étirements** de **lettres** **Interjections** **Émojis** *Tout le Netspeak*
21-30 ans	**Étirements de lettres** (1.36) **Interjections** (1.26)	Graphies phonétiques (1.32)	Abréviations **Émojis** *Tout le Netspeak*
31 ans et +	**Étirements de lettres** (1.62) **Émojis** (4.41) **Interjections** (1.52) Abréviations (1.19) *Tout le Netspeak*	-	Graphies phonétiques

A very brief overview of results

- Interaction with race also nuances results

TABLEAU 12.5 – Différences significatives entre femmes et hommes, par groupe ethnique

Variables	Blancs	Afr.Am.	Asiatiques	Hispaniques
		AVEC INTERACTION		
Émoticônes	-	**F** (2.94)	-	**F** (2.69)
Émojis	-	**F** (3.76)	**F** (10.37)	-
Étirements de ponctuation	**F** (1.81)	**F** (2.31)	**F** (2.44)	-
G-droppings	-	-	-	**H** (4.92)
Interjections	**F** (1.52)	-	**F** (1.40)	-
Graphies phonétiques	-	**H** (2.06)	**H** (1.62)	**H** (1.55)
Omissions d'apostrophe	-	**H** (2.06)	**H** (6.14)	-
Mots en majuscules	-	-	-	-
		SANS INTERACTION		
i	**H** (2.25)			
Étirements de lettres	**F** 21-30 ans et 31 ans et + (1.41 ; 1.71)			
Abréviations	-			

A very brief overview of results

- Do trans people align their linguistic pratices with that of cis people?

"Gendered variables"	Trans men and nonbinary individuals	Trans women
Emoticons	No alignment	Align with cis women
Punctuation lengthenings		Align with cis men
i		
Letter lengthenings		Align with cis men (21+)
Apostrophe omission		No alignment
Interjections		

New corpus, new challenges

A new Reddit corpus in French

- Weekly scraping of r/france with RedditExtractoR
- Corpus processed with R and the quanteda package
- 1st exploration with automatic method

je suis un homme / je suis une femme
je suis papa / je suis maman
je suis content / je suis contente
je suis fatigué / je suis fatiguée

- Issue: back to the binary…
- Solution: more data; or asking Redditors (sociolinguistic questionnaire)

Conclusion

NLP x sociolinguistics

- "Computational sociolinguistics" (Nguyen et al., 2016)
- Cross-fertilization between NLP and sociolinguistics
- For NLP: need to be more careful in the way they define gender and assign gender identity, and to state limitations of their research

But do we really need NLP research about trans & nonbinary people?

- For sociolinguists: need to work with the NLP community/or learn NLP tools to scrape data, process corpora, perform statistical analyses
- Esp. in France (very few computational socioliguistic studies)

Thx! 👍

References

Androutsopoulos, J. (2014). Computer-mediated communication and linguistic landscapes. In J. Holmes & K. Hazen (Éds.), *Research methods in sociolinguistics: A practical guide* (p. 74–90). Wiley-Blackwell Oxford.

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, *18*(2), 135–160.

Barthel, M., Stocking, G., Holcomb, J., & Mitchell, A. (2016, février 25). Reddit news users more likely to be male, young and digital in their news preferences. *Pew Research Center's Journalism Project*.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on Twitter. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1301–1309.

Butler, J. (2006). *Gender trouble* (3rd ed.). Routledge.

Coats, S. (2017). Gender and lexical type frequencies in Finland Twitter English. *Studies in Variation, Contacts and Change in English*, *19*.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University Of Chicago Legal Forum*, *1*, 140–167.

Cunha, E., Magno, G., Gonçalves, M. A., Cambraia, C., & Almeida, V. (2014). He votes or she votes? Female and male discursive strategies in Twitter political hashtags. *PLoS ONE*, *9*(1), e87041.

Eisenstein, J. (2015). Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, *19*(2), 161–188.

References

Fausto-Sterling, A. (2012). Sex/gender: Biology in a social world. Routledge.

Finlay, S. C. (2014). Age and gender in Reddit commenting and success. *Journal of Information Science Theory and Practice*, *2*(3), 18–28.

Goodman, B. W. (2016). A step towards accountable algorithms?: Algorithmic discrimination and the European Union general data protection. *29th conference on neural information processing systems (NIPS 2016), barcelona. NIPS foundation*.

Larson, B. (2017). Gender as a variable in natural-language processing: Ethical considerations. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 1–11.

Levon, E. (2015). Integrating intersectionality in language, gender, and sexuality research. *Language and Linguistics Compass*, *9*(7), 295–308.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2011). Understanding the demographics of Twitter users. *ICWSM*.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics*, *42*(3), 537–593.

Thelwall, M., & Stuart, E. (2018). She's Reddit: A source of statistically significant gendered interest information? *arXiv:1810.08091 [cs]*.