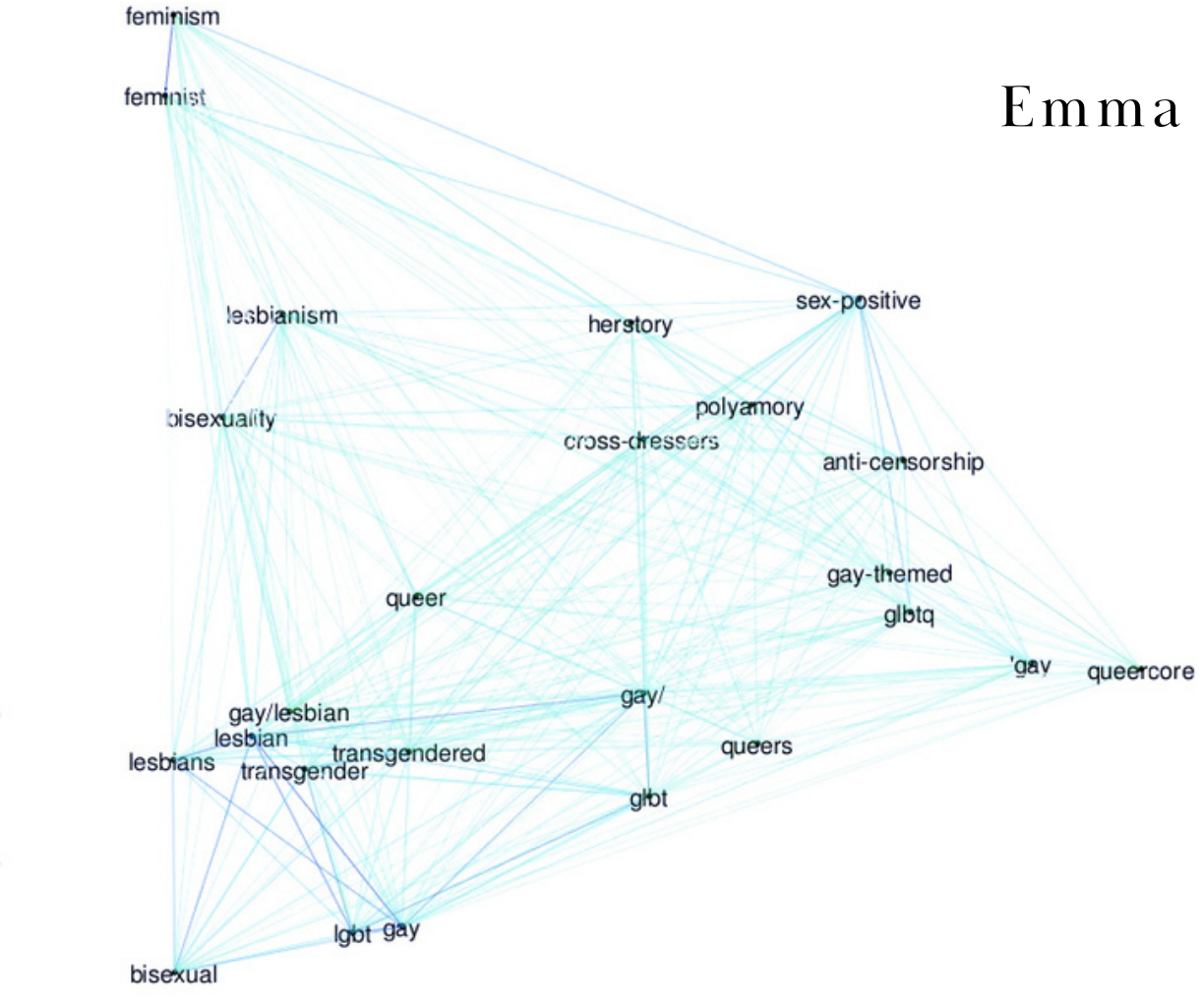
An exploration of gender biases through a distributional semantics lens



Emma Angela Montecchiari

Project Report Presentation
Human Language Technologies
Università di Trento
Cognitive Science Master's
09-02-2024

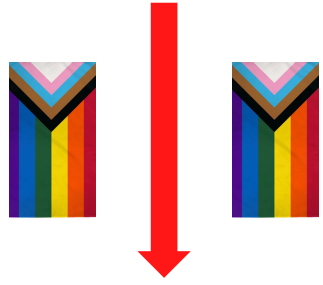
General framework

Biases against marginalised identities in language models have been well-documented, underlying implicit human biases manifesting in language regularities

Retention of voices aligned with hegemonic viewpoints (e.g. inadequately representation of social movements)

Specific framework

Majority of biases analyses focused on binary gender and associated stereotypes



To do: enhance a more accurate and inclusive comprehension of non-binary gender perspective

Specific framework

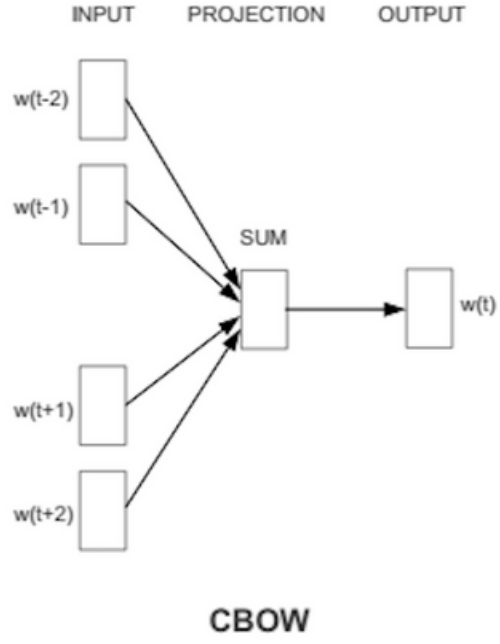
Non-binary genders encounter obstacles in media representation and access to economic and political opportunities.

- Sample size disparities (lack of understanding);
- Cyclical erasure of gender diversity (misgendering);
- Negative narratives and scarcity of positive gender non-conforming content.

Specific framwework

Proposal

Exploratory analysis on embedded contextual representations of a specific set of relevant words in gender studies



Cross-language exploration: English, German, Italian, French, Spanish and Croatian

Method

Embeddings algorithms:

- Word2Vec CBOW (target predicted based on context)
- HAL (co-occurrence patterns with a moving window)

+

Latent Semantic Analysis LSA (term-by-document matrix)

On pre-built semantic spaces

trained on WaC, Wikipedia, BNC, OpenSubtitles corpora

English, Italian, German, Spanish, Croatian - CBOW

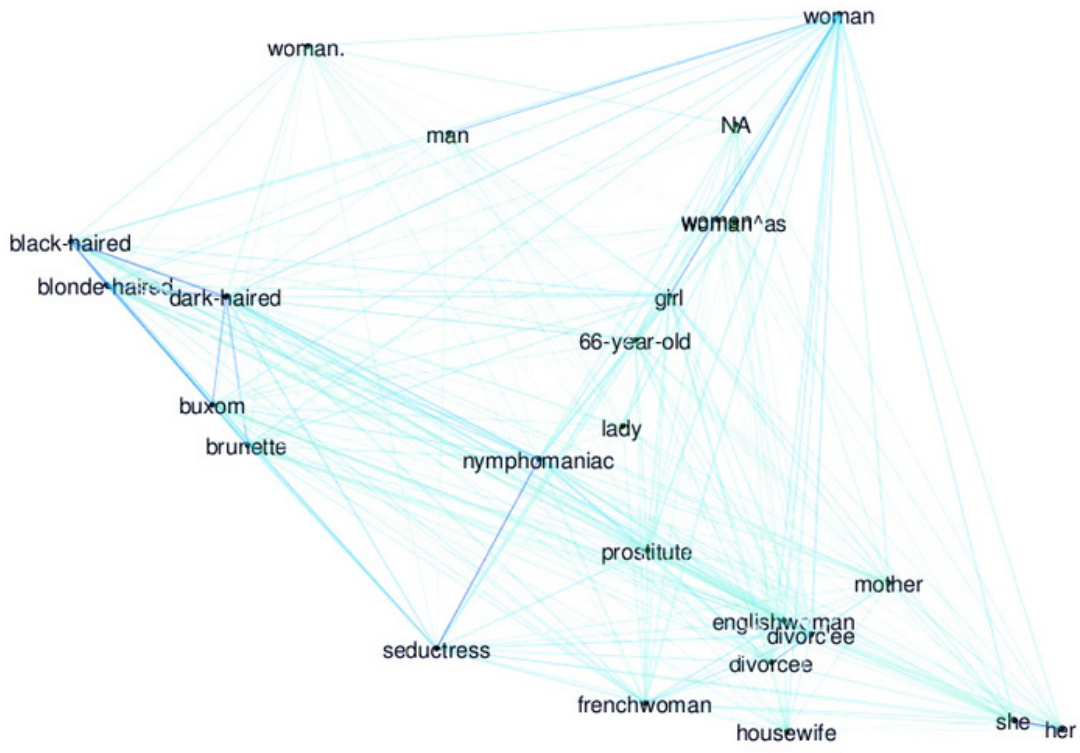
French - HAL

Procedure (A)

(A) Investigation over the entire spaces

[R implementation - LSAfunpackage]

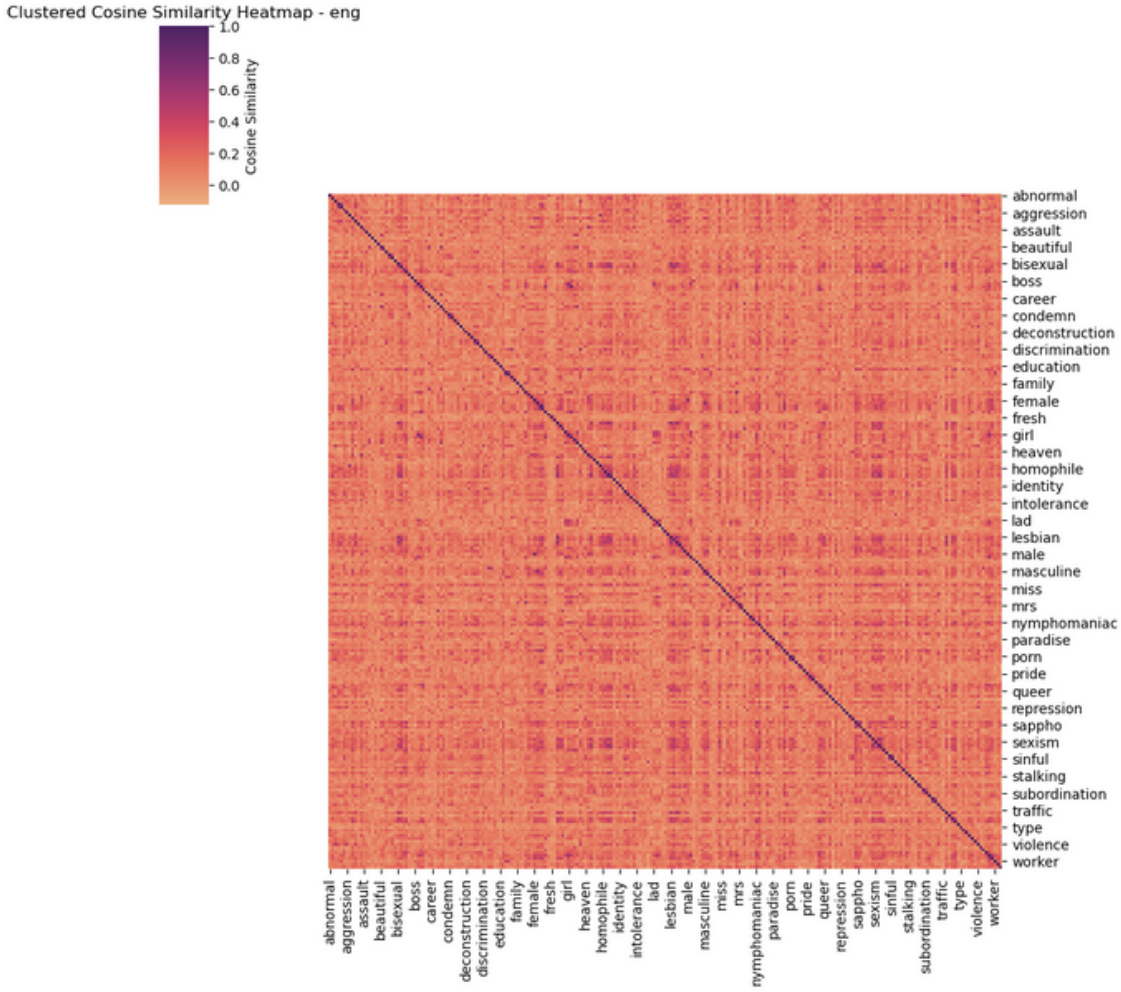
neighbours of specific words



Procedure (B)

(B) Extraction of cross-lingual sub-spaces of specific words: multicosine similarity matrices

[to restrict the focus on discriminatory or lgbtq+ language]:

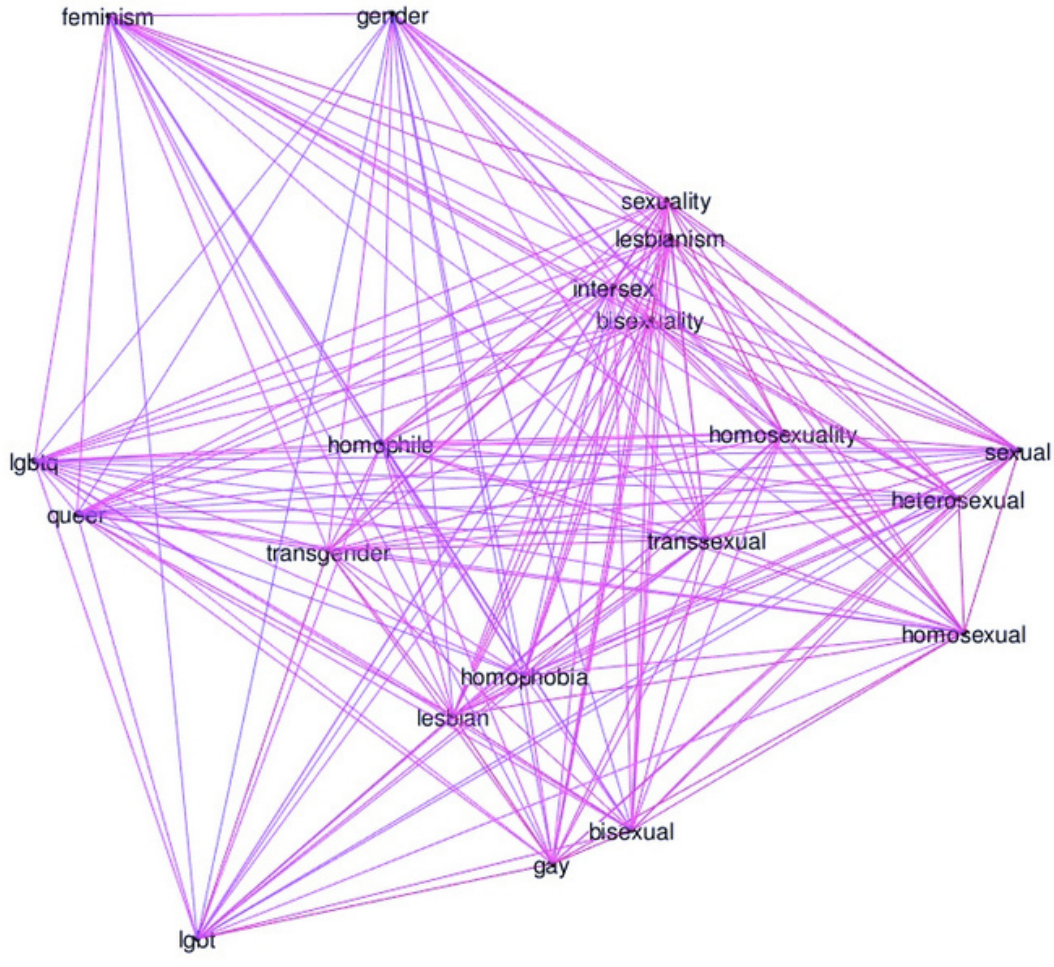


Procedure (C)

(C) **Sub-spaces** - (i) Neighbours to target words

(ii) Most similar matrices pairs

[Python implementation - interactive code]



Results - (A) Entire spaces

Entire semantic spaces:

Extraction of 20 most similar items to target words for each language.

Target words: *'man'*, *'woman'*, *'female'*, *'male'*, *'body'*, *'sex'*, *'queer'*, *'homosexual'*, *'gender'*, *'trans- gender'*.

Language	Target Term	Neighbors
ENG	'man'	man woman gentleman gray-haired boy person lad men girl stranger
	'woman'	woman girl man divorcée englishwoman lady **nymphomaniac prostitute**
	'female'	female male **cross-dressers incubates polygynous** spermatophore **male-dominated**
	'male'	male female male-to-female sub-adult **unringed polygynous** african/caribbean lekking
	'body'	body bodies torso corpse limbs incorrupt viscera musculature endoskeleton
	'sex'	sex sexual sexuality homosexual masturbation heterosexual male-male
	'queer'	queer lesbian gay feminist lgbt bisexual trasgender **queercore sex-positive herstory**
	'homosexual'	homosexual homosexuality sexual heterosexual gay sam-gender **pedophilia**
	'gender'	gender gendered **ethnicity** sexuality masculinities **race** subjectivities **racialised**
	'transgender'	transgender transgendered lgbt transsexual lesbian bisexual intersex transmen
ITA	'man'	uomo **dio vita** creatura **umano anima** donna **destino divino** amore vivente
	'woman'	donna femminile uomo bambina ragazza madre maschio marito **sesso incinto**
	'female'	femminile maschile donna giovanile **femminista** femminilizzazione **emancipazione**
	'male'	maschile femminile maschio sesso donna mascolinità **maschilismo** sessuato **virile**
	'body'	corpo membra corporeo anima eterico chakras prana uomo vivente **penetrare**
	'sex'	sesso maschio maschile sessuale donna femminile eterosessuale omosessuale erotico
	'queer'	queer lesbian gay cinematic prejudice **bigottismo** videos coming folk
	'homosexual'	omosessuale gay eterosessuale lesbica sessuale **omofobia anti-gay** transessuale
	'gender'	genere tipo simile semplicemente spesso piuttosto certo raramente ovvio banale
	'transgender'	transessuale haraway lesbica arcilesbica femminista transex trans ermafrodito cyborg
FR	'man'	homme **esprit âme** humain femme père comme mort peuple monde seul humanité
	'woman'	femme fille **mère** jeune mari doeur compagne elle garçon amie maïtresse père princesse
	'female'	feminine femininpersonnalite emitie mariee debuter reussie represente evolue **fidele**

Results - (B) Sub-spaces

Sub-spaces with selected words:

80 pairs with the most similar words in the similarity matrices .

In Appendix B (and example): only pairs reporting biases.

Pair	Similarity
Pair: ('djevojka', 'drolja')	0.4782
Pair: ('kurva', 'lezbijka')	0.4694
Pair: ('kurva', 'majka')	0.4428
Pair: ('djevojka', 'prostitutka')	0.3973
Pair: ('drolja', 'lezbijka')	0.3800
Pair: ('drolja', 'majka')	0.3774
Pair: ('dama', 'drolja')	0.3686
Pair: ('grijeh', 'ponos')	0.3518
Pair: ('lezbijka', 'majka')	0.3512

(e) Similarity Pairs in Croatian

Results - (B) Sub-spaces

Item	Similarity
masculin	0.9298
sexe	0.6835
glamour	0.6500
vogue	0.6211
sexy	0.5963
gay	0.5948
androgyne	0.5867
sexualité	0.5862
sexuel	0.5812
sport	0.5784
imaginaire	0.5693
beauté	0.5688
femme	0.5584
sentimental	0.5425
diable	0.5302
porno	0.5281

French sub-space - Most similar to 'féminine'

Sub-spaces with selected words:
Neighbours analysis which aims to be more paradigmatic on biases.

Chosen 3 binary vs. non-binary umbrella terms: 'transgender' 'male' 'female'
Languages: Italian, French, English

Discussion- (A) Entire spaces

'man'	uomo **dio vita** creatura **umano anima** donna **destino divino** amore vivente
'woman'	woman girl man divorcée englishwoman lady **nymphomaniac prostitute**
'male'	maschile femminile maschio sesso donna mascolinità **maschilismo** sessuato **virile**
'female'	female male **cross-dressers incubates polygynous** spermatophore **male-dominated**
'queer'	queer lesbian gay feminist lgbt bisexual trasgender **queercore sex-positive herstory**
'gender'	gender gendered **ethnicity** sexuality masculinities **race** subjectivities **racialised**
'homosexual_a' 'homosexual_b'	homosexuel hétérosexuel **pédophile** homosexuelle **immigré violeur** obsédé notoire **délinquant** travesti **drogué** célibataire repenti quinquagénaire hétéro polygame
'transgender'	transessuale haraway lesbica arcilesbica femminista transex trans ermafrodito cyborg

Discussion - (B) Sub-spaces

Pair	Similarity
('boyfriend', 'girlfriend')	0.7948
('female', 'male')	0.7899
('feminine', 'masculine')	0.7364
('boy', 'girl')	0.7291
('father', 'mother')	0.7254
('feminism', 'feminist')	0.7118
('gay', 'lesbian')	0.6894
('lesbian', 'lgbt')	0.6545
('asexual', 'hermaphroditic')	0.6081
('lgbt', 'transgender')	0.5999

'femme' - 'mère'	0.9081
'fille' - 'mère'	0.9078
'machisme' - 'misogynie'	0.8818
'homophobie' - 'sexisme'	0.8665
'machisme' - 'sexisme'	0.8663
'homosexualité' - 'inceste'	0.8396

('nymphomaniac', 'transvestite')	0.5590
'nymphomaniac', 'prostitute'	0.5503
('lesbianism', 'misogyny')	0.5422
('misogyny', 'objectification')	0.5245
('girl', 'nymphomaniac')	0.5163

('dama', 'drolja')	0.3686
('drolja', 'lezbijka')	0.3800
('kurva', 'lezbijka')	0.4694
('mujer', 'zorra')	0.3850

Discussion - (C) Sub-spaces

Item	Similarity
transessuale	0.3762
lesbica	0.3520
femminista	0.3092
omosessuale	0.2891
gay	0.2886
femminile	0.2850
travestito	0.2703
immaginario	0.2655
maschile	0.2581
autocoscienza	0.2494
androgino	0.2483
maschilismo	0.2437
pride	0.2436
queer	0.2415
lesbismo	0.2349
femminismo	0.2342
prostituta	0.2243
sessuale	0.2236
donna	0.2217
stereotipo	0.2217
drag	0.2185
sottocultura	0.2183
saffico	0.2150
invisibile	0.2117
bisessuale	0.2092
sesso	0.2002

Item	Similarity
femminile	0.7837
maschile	0.5467
donna	0.4951
femminista	0.4723
femminismo	0.4441
sesso	0.4398
maschilismo	0.3930
mascolino	0.3904
sessuale	0.3806
gender	0.3483
androgino	0.3472
genitale	0.3421
giovane	0.3408
stereotipo	0.3308
misoginia	0.3243
immaginario	0.2983
lesbismo	0.2979
transessuale	0.2956
lesbica	0.2918
transgender	0.2850
segregazione	0.2751
antiviolenza	0.2746
sessismo	0.2640
omosessuale	0.2592
harem	0.2582
prostituzione	0.2573

Future directions:

- Navigating complexity in gender diversity;
- Linking with diverse data sources;
- Exploration of pronouns diversity.

Thank you for your attention!

The code is available at:

https://github.com/memonji/gender-biases-exploration.git