Project Report

Emma Angela Montecchiari
[emma.montecchiari@studenti.unitn.it]

An exploration of gender biases through a distributional semantics lens

University of Trento - Cognitive Science Master's degree
Course: Human Language Technologies
Winter Session Submission - February 9, 2024

Abstract

This study addresses the need to mitigate biases in language models, focusing on the often-overlooked experiences of non-binary individuals. While existing research has primarily tackled biases related to binary gender stereotypes, this study explores the broader spectrum of gender identities. Through an exploratory analysis of embedded contextual representations within distributional language models, the aim is to observe and identify biases in a multi-language set of semantic spaces.

1 Introduction

As the prevalence of language models in language processing applications increases, ensuring a higher degree of fairness in their learned representations and intervening in any biased decisions they make has become increasingly crucial. It is well-documented that large language models exhibit various biases, including stereotypical associations and intersectionality effects, disproportionately encoding biases against marginalized identities along multiple dimensions [5, 16, 30]. Understanding that implicit human biases manifest in the statistical regularities of language, studying how these biases could perpetuate harm is both feasible and important.

Despite the vastness and diversity of the internet, large datasets collected from it often fail to capture the full spectrum of viewpoints. Factors such as narrow internet participation, filtering of crawled data, and retention of voices aligned with hegemonic viewpoints skew the representation. This over-representation of specific categorical views in training data, particularly in US and UK English, not only exceeds their prevalence in the general population but also exacerbates biases in language models, potentially leading to the amplification of harms. Media coverage inadequately represents social movements, resulting in the misrepresentation of marginalized identities in language models [12, 7, 23]. Efforts have therefore been made to identify and address social biases in language processing, including quantification and mitigation strategies [27, 33, 34, 29].

While a significant portion of social bias studies on language models have focused on biases related to binary gender and associated stereotypes [33, 32], the scope of gender in these analyses and associated performance metrics predominantly revolves around binary gender. While addressing biases related to binary gender and enhancing model performance remain essential, reframing our comprehension of gender in language technologies to embrace a more accurate, inclusive, and non-binary perspective is imperative, compounding their effects through intersectionality.

Biases faced by non-binary individuals may significantly differ, with a high risk of including a cyclical erasure of non-binary gender identities[13], driven from sample size disparities, non-recognition and lack of understanding of non-binary genders [32, 30]. Non-binary individuals in fact frequently encounter obstacles in media representation and access to economic and political opportunities [26], leading to negative narratives or erasure of gender diversity within communities. Training data predominantly reflect hegemonic viewpoints and prioritize modeling systems simplicity [14], therefore it often contains negative connotations and a scarcity of positive gender non-conforming content [10, 4].

The concept of gender is multifaceted and complex, encompassing various aspects of a person's internal experience, social expression, societal expectations, and perception [10, 4]. Misgendering, whether of transgender or cisgender individuals, can have significant ramifications, perpetuating between others, psychological harm. Recent efforts aim to mitigate these harms by constructing task-specific datasets that go beyond binary gender and developing metrics to potentially measure biases against all genders [3, 27].

In this project, my focus delves in this framework within an exploratory analysis on embedded contextual representations of selected words. It is done through a distributional semantic lens, stating that there exist a

correspondence between a word's distribution over contexts and its meaning, therefore that words with similar meaning tend to occur in similar contexts. The techniques to achieve these representations are various but all based on vectorial counting of words co-occurrences in contexts, with a wide variety of what context is. I use these representation to try to analyse and understand which biases language models bound with [2, 19].

2 Methods

The investigation conducted in this work focuses on exploring distributional representations of a specific set of terms relevant to gender studies. The distributional model I chose for my analysis is Word2Vec [25]. The study would like to be a cross-language comparative analysis, so to observe how biases spread along different linguistic and cultural systems as well as in different sets of training data. Seen my cultural and linguistics mother-tongue being a Romance language, I chose to focus on a European framework. This includes Germanic, Romance and Slavic languages: English, German, Italian, French, Spanish and Croatian. The analysis has been built on an interactive code, enabling users to put their hands on data as well. The material and code has been made available on GitHub: https://github.com/memonji/gender-biases-exploration.git.

2.1 Model

The distributional model chosen for the analysis is Word2Vec, known for its classic features [25, 24] and the extensive literature on its ability to simulate human cognitive abilities [18, 1]. It is generating word embeddings within neural network architectures, such as Continuous Bag of Words (CBOW) and Skip-Gram models. In CBOW, the model predicts a target word based on its context words, while in Skip-Gram, the model predicts context words given a target word. These models are trained on large text corpora to learn distributed representations of words in a continuous vector space. Additionally, I employ the Hyperspace Analogue to Language (HAL) [21] Recurrent Neural Network model, which operates by analyzing the co-occurrence patterns of words within a moving window of text.

Furthermore, I use a Latent Semantic Analysis (LSA) metric [8, 20], which aims to uncover the underlying semantic structure of textual data. Initially, a term-by-document frequency matrix is constructed from a text corpus. A weighting scheme, such as log-entropy weighting [22], is then applied to enhance the influence of low-frequency words, which often carry more specific meanings. Singular Value Decomposition (SVD) is then used to decompose the weighted matrix into orthogonal word and document matrices, effectively representing words and documents in a shared semantic space. The resulting vectors enable semantic similarity calculations using cosine similarity, with higher values indicating greater semantic relatedness.

2.2 Dataset

The used data have been retrieved from diverse languages pre-built semantic spaces, which I have downloaded from: Homepage of Fritz Günther. Those are:

- **baroni** English CBOW space. It is derived from a 2.8 billion word corpus, including the British National Corpus, the ukWaC corpus, and a 2009 Wikipedia dump. It employs a context window size of 11 words (5 left, 5 right) and 400-dimensional vectors [1].
- **frwak** French HAL space with 300 dimensions generated from the 1.6 billion word frWaC corpus. HAL-like moving window model with a window size of 5, incorporating the 100k most frequent words without lemmatization. A Positive Pointwise Mutual Information weighting scheme and Singular Value Decomposition were applied to reduce the space from 100k to 300 dimensions.
- **itwac** Italian CBOW space, 400 dimensions, generated from the 2 billion word itWaC corpus. Similar to the German space, it employs the CBOW algorithm with a context window size of 5 words and 400-dimensional vectors, utilizing negative sampling with $k = 10$ and subsampling with $t = 1 \times 10^{-5}$.
- **dewac** German CBOW space, 400 dimensions, derived from a lemmatized version of the 1.7 billion word deWaC corpus. This space utilizes the CBOW algorithm with a context window size of 5 words and 400-dimensional vectors, employing negative sampling with $k = 10$ and subsampling with $t = 1 \times 10^{-5}$.
- **es** Spanish CBOW space, 400 dimensions, created from a lemmatized version of the 1.5 billion word OpenSubtitles 2018 Spanish corpus. Similar to the previous spaces, it employs the CBOW algorithm with a context window size of 5 words and 400-dimensional vectors, utilizing negative sampling with $k = 10$ and subsampling with $t = 1 \times 10^{-5}$.
- **hr** Croatian CBOW space, 300 dimensions, generated from the 707 million word OpenSubtitles 2018 Croatian corpus. This space utilizes the CBOW algorithm with a context window size of 5 words, negative sampling with $k = 10$, and subsampling with $t = 1 \times 10^{-5}$. It contains vectors for 100,000 different words due to the smaller size of the source corpus.

From these spaces, a set of words was selected based on literature dealing with gender studies [31, 33, 13, 29, 26], manually ensuring parallelism across languages. Starting with Italian and English, the most relevant words were then translated into the other languages. The lists of selected words for the English language include terms such as *'abnormal', 'abuse', 'activism', 'adultery', 'aggression', 'aids', 'androgynous', 'anti-violence', 'art', 'asexual', 'assault'*, and many more.

2.3 Procedure

Within this model and dataset framework, I'm conducting my analyses using R and Python.

The initial step has been implemented on R, downloading the pre-built semantic spaces and then computing analyses using the LSAfun package [17], which is built to deal with semantic spaces. I computed neighbors analysis on specific words all over the entire semantic spaces ranges. This metric extracts the most similar words to the target one, based on similarity in vectorial representation. A 3-dimensional example of this extraction can be seen in Figure 1. The overall results are shown in Appendix A. You can find code I used for this step at: https://github.com/memonji/gender-biases-exploration/blob/main/Overall$_n eighbors_c omparison.R$.

The second step has been to extract sub-spaces focusing on language addressed in gender studies, such adjectives and nouns, associated with positive or negative connotations. Examples were chosen based on intuition, existing literature and frameworks such as Sketch Engine co-occurrences [31, 33, 13, 29, 26, 10].

In constructing the dataframes, I initially compiled lists in English and Italian languages in which I am proficient, along with French, which has abundant literature on the subject. I then aligned the lists (with translations) cross-linguistically, aiming for inclusivity while recognizing cultural and linguistic differences. Due to variations in available data and literature across languages, a limited number of terms were selected to focus on specific characteristics of interest, avoiding overly broad analyses that might obscure discriminatory language features. This is also the reason why I dive in these sub-spaces, since to be able to restrict my observation framework on specific discriminatory instances. I developed this step in: https://github.com/memonji/gender-biases-exploration/blob/main/Subspaces$_c reation.R$.

The process involved interactive functions in R code to align and order the terms, resulting in multi-cosine similarity matrices for each language, which could easily be imported in other programming language, i.e. Python. Matrices are available at: https://github.com/memonji/gender-biases-exploration/tree/main/matrices.

Finally, Python was used to create an interactive code allowing users to explore (a) the most similar words to a target one extracted from the matrices (sub-spaces); (b) the most similar words pairs extracted from the matrices (sub-spaces); (c) heatmaps for each language, such as the one in Figure 2. The code is available on GitHub: https://github.com/memonji/gender-biases-exploration/blob/main/main.py.

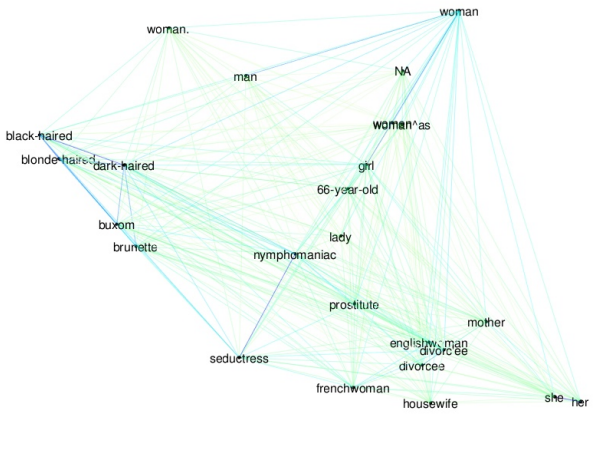

Figure 1: 3D Neighbors - Eng - 'woman'

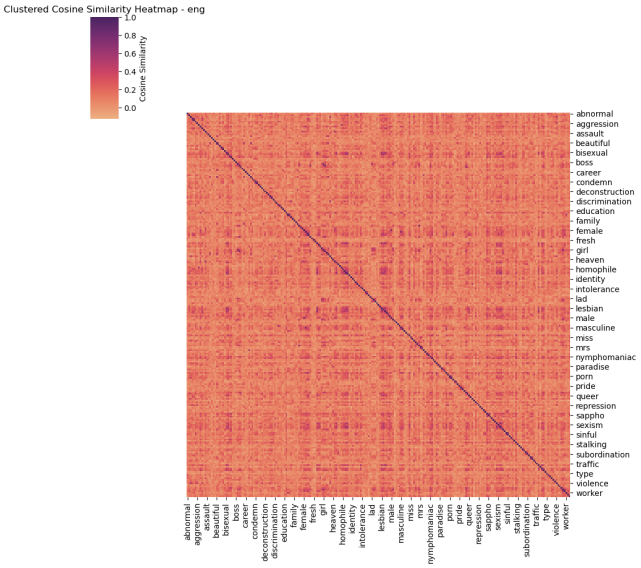

Figure 2: Eng - Multicosine - Heatmap

3 Results

3.1 Investigation over the entire space

In the first analysis, I computed over the entire semantic spaces, extracting neighbors of the target words I inputted, and extracting lists of 20 most similar terms. I selected a set of target words that seemed to me the most salient for investigating the relationship between different languages/cultures and the gender biases they might reveal. The analysis could have been broad and deep, and I chose to select the words *'man, woman, female, male, body, sex, queer, homosexual, gender, transgender'* which seemed to me canonical in the way of straightforwardly allow me to retrieve biases from most common categories. The results are reported in Appendix A, where I divide them by language and put the most similar items. I omitted the numerical differences due to space constraints.

3.2 Investigation over the sub-spaces

3.2.1 Most similar pairs

After creating these sub-spaces, I conducted two analyses. First, identifying pairs that were most similar to each other in the similarity matrix. The reported pairs are selected from the 80 most similar pairs in each language sub-space reporting some sort of biases. The cosine similarity score is plotted at the left. In German, I didn't find many biases, and those that I did find were not significant enough to report. I report the results in Appendix B.

3.2.2 Most similar words to target

For the third analysis, I focused on a similarity comparison inside the sub-spaces, so to be able to find more paradigmatic comparisons. I chose to focus on just three umbrella terms: 'transgender', 'female', and 'male'. I chose them so to observe the difference between binary vs. non-binary umbrella terms, which, as they could have been more appropriate, they can still give a good point of view. I computed the measures in the languages I'm most proficient in: Italian, English, and French.

4 Discussion

4.1 Investigation over the all space

As previously mentioned, the initial analysis delved into the overall semantic spaces, exploring the neighbors of a selected set of words: *'man', 'woman', 'female', 'male', 'body', 'sex', 'queer', 'homosexual', 'gender', 'transgender'*. Across all languages, I computed the most similar neighbors for each word in the set. The results, detailed in Appendix A, reveal nuanced associations within each language.

A comparative examination across languages reveals intriguing insights. For instance, the term *'man'* tends to carry positive connotations across languages. In Italian and French, it evokes associations with godly or universal human attributes, emphasizing its universal significance.

Conversely, *'woman'* carries more complex social connotations, often entangled with themes of oppression and control. In English, terms like *nymphomaniac* and *prostitute* reflect historical denigration and limited social recognition. Similarly, in German, 'woman' (*'frau'*) is associated with 'sexual' (*'geschlechtlich'*), highlighting societal attitudes towards gender and sexuality.

The term *'female'* elicits varied associations, from notions of social oppression to feminist empowerment. In English, it is linked to *'male-dominated'*, reflecting struggles against gender inequities. Interestingly, queer-related terms like *'cross-dresser'* and *'polygynous'* also emerge, suggesting intersections with diverse gender identities.

Conversely, *'male'* in Italian is associated with terms like 'chauvinism' *'maschilismo'* and *'virile'*, reflecting stereotypical gender norms. This contrasts with English, where *'male'* tends to evoke less loaded associations.

Exploring *'queer'* reveals a range of associations, from positive cultural movements to prejudiced attitudes. English terms like *'queercore'* and *'sex-positive'* highlight positive LGBTQ+ representation, whereas Italian and French terms like 'sanctimony' *'bigottismo'* suggest entrenched biases.

The term *'homosexual'* evokes negative connotations across languages, with associations like *'pedophilia'* and *'anti-gay'*. In French, particularly, the associations include stigmatizing terms like *'rapist'* and *'criminal'*, highlighting societal prejudices.

'Gender' is more connotatively rich in English, with associations to *'ethnicity'* and *'race'*. This suggests a broader discourse around intersectionality and identity in English-speaking contexts.

Interestingly, *'transgender'* elicits fewer negative associations in Italian and none in French, indicating varied societal attitudes towards transgender identities across languages.

4.2 Sub-spaces investigation

4.2.1 Most similar Pairs

The results are plotted in Appendix B, where pairs revealing perpetuation of social biases are reported. A noticeable polarization of gender identities emerges, particularly in pairs like *'female - male'* and *'feminine - masculine'*, which exhibit similar distributions and are perceived as opposites. This binary representation excludes non-binary identities, contributing to the marginalization of communities outside the traditional male/female spectrum.

Signs of discrimination are evident in several pairs. For instance, in English, pairs like *'nymphomaniac-tranvestite'*, *'nymphomaniac-prostitute'*, and *'lesbianism-misogyny'* suggest stigmatization of sexual expression and gender identities. Similarly, in French, pairs like *'machisme-misogynie'* and *'homophobie-sexisme'* reflect entrenched gender biases.

In Spanish, pairs like *'mujer-zorra'* ('woman-prostitute') highlight derogatory associations with femininity and sexuality. Croatian pairs reveal pervasive discrimination against women, mothers, and lesbians, exemplified by pairs like *'djevojka-drolja'* ('girl-slut'), *'kurva-lezbijka'* ('prostitute-lesbian'), and *'dama-drolja'* ('lady-slut').

4.2.2 Most Similar words to target

For the third analysis, the results are plotted in Appendix C. Here I focused on a similarity comparison inside the sub-spaces, so to be able to find more paradigmatic comparisons. I chose to focus just on the three most typical binary vs. non-binary umbrella terms, such as 'transgender', 'female', and 'male', for English, French and Italian. I had to choose 'queer' instead of 'transgender' for French since the word was not represented in the downloaded pre-built French semantic space.

In the **Italian** language, under the 'transgender' umbrella term, I find terms related to self-consciousness, feminism, and pride, associated with the LGBTQ+ communities. These include terms like *queer, pride, feminism, drag, sub-culture,* and *deconstruction*. However, discriminatory terms such as *maschilism, slut, stereotype, invisible, mysogyny,* and *bitch* are also present. Similarly, under the 'feminine' umbrella term, Italian lists contain terms related to the feminist movement, alongside discrimination-related terms like *maschilism, mysogyny, stereotype, segregation, anti-violence,* and *discrimination*. For the term 'male', both feminism and mysogyny are prominent, along with terms like *sexism, segregation, incest, contraceptive,* and *pronoun*. Queer terms such as *gay, asexual,* and *transexual* are also present but appear towards the bottom of the similarity items.

In the **English** language, under the term 'transgender', definitions of non-binary sexualities such as *transexual, lesbian, bisexual, intersex, gay, homosexuality,* and *lgbtq* are prominent. Discriminatory-related terms like *hiv, discrimination, objectification,* and *anti-violence* are also present. Similarly, the term 'female' is directly related to non-binary characteristics such as *hermaphroditic, androgynous, effeminate,* and *intersex*. Discriminatory terms like *nymphomaniac, objectification,* and *mutilation* are mostly present for the female term but are also found in the male frame.

In the **French** language, the term 'queer' is mostly associated with negative terms such as *bitch, porno, slut, deviance,* and *madness*. Terms used in postcolonial gender studies such as *feminism* and *postcolonial* are also present. For female attributes, there is a strong association with sexual imagery, fantasy, and pornography. Similar associations are found for male terms, which are also related to gay and androgynous identities.

5 Conclusion and future directions

This project delved into an exploratory observation of denigratory stereotypical biases in language models. While results are present, it is crucial to continue the analysis to mitigate discriminatory harms. Future work could focus on refining the evaluation metrics used to assess biases, expanding the dataset to include a more diverse range of languages and cultural contexts. More precisely:

Navigating complexity in gender representation: Gender representation in language models mirrors the complexity of gender itself. Attempts to categorize gender into fixed, discrete categories risk marginalizing segments of the population. Given the documented harms of misgendering and erasure, future research must carefully consider the conceptualization and modeling of gender in language representations and tasks.

Linking with diverse data sources: Augmenting traditional datasets with insights from surveys of LGBTQ+ social media discourse could enhance the representation of diverse gender identities. However, it is essential to recognize the limitations of online data, which may not fully capture the nuances of non-binary experiences. Moreover, critical attention should be paid to how institutional and cultural channels perpetuate misgendering practices, particularly within systems such as job recruitment and healthcare [28, 11].

Exploration of pronouns diversity: Mitigating harms work on pronouns has been done. Yet, there exists a rich diversity of non-cis identities globally [9, 4]. In languages devoid of referential gender or where pronouns are sparingly used (e.g., Estonian), pronouns may hold less centrality to an individual's gender identity [6]. Non-binary individuals often navigate multiple pronoun sets, adapting their usage based on context and personal preference [15]. Future research could delve into the distribution and usage patterns of different pronoun sets across contexts such as creative language use and occupational environments. This investigation could shed light on the intersectionality of gender identity and professional roles.

A Appendix A

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**
	'male'	masculin féminin masculine féminine sexe singulier qualificatif significant pluriel
	'body'	corps visage mains âme bras yeux sorte peau sang intérieur pieds propre force humain
	'sex'	sexe sexualle sexualité sexuel sexes masculin féminin couple homosexualité âge
	'queer'	queer revival melting mainstream **freak lovers teen glam hype mood** fever gypsy gothic
	'homosexual_a'	homosexuel hétérosexuel **pédophile** homosexuelle **immigré violeur** obsédé notoire
	'homosexual_b'	**délinquant** travesti **drogué** célibataire repenti quinquagénaire hétéro polygame
	'gender'	genre bref vrai surtout aussi plutôt voilà autant sûr oublier mais vraiment juste sait
GER	'man'	man aber einfach es so wenn da etwas was vielleicht eben dann nicht gar also doch
	'woman'	weiblich geschlecht weibliche weiblichkeit frau lady maskulin **geschlechtlich**
	'female'	weiblich geschlecht weibliche frau maskulin **geschlechtlich** fraunrolle feminin **androgyn**
	'male'	male lof bouledogue labrador males **dogue** femmelle chiot yorkshire retriever teckel
	'body'	korper **irrationale fetischs kunsttheoretisch** abstraktion materialismus
	'sex'	sex erotik erotisch **porno** seitensprung one-night-stand selbstbefriedigung
	'queer'	queer gender **feministisch** queeren **technoscience** geschlechterkonstruktion lesbisch
	'homosexual'	homosexuell gleichgeschlechtlich homosexuelle heterosexuell schwul lesbisch sexuell
	'gender'	genre film stilistisch literarisch filmisch comic stil klassiker story gattung thriller
	'transgender'	transgender transsexuellengesetzes **lesbisch antidiskriminierungsgesetzgebung**
SP	'man'	hombre él quien tipo como verdadero un y mujer tonto alguien un y mujer tonto
	'woman'	mujer ella marido chica amante enamorado novia **madre enamorada joven** ama hija
	'female'	hembra macho crīa cabrío animal apareamiento criatura vagina mamífero placenta
	'male'	macho hembra cabrío animal mono mamífero **excita** lémur hetero celo cuidao pene
	'body'	cuerpo cadáver alma humano mente corazón pecho muerte cerebro
	'gender'	genero naturaleza fluya género humano nutren intima ficcion abarcar
CRO	'man'	covjek covek decko djecak mladic muskarac muz tip momak nacin zivot
	'woman'	zena devojka devojcica prica djevojka znas kaze cerka kcerka musterija cura mozda
	'male'	muski obican zenski jedobar mozda godisnji napustis skolski jebeni kaze zatvoris ocito
	'body'	tijelo telo truplo srce tjelo lice dijete meso tkivo mjesto gnijezdo stvorenje prestaro
	'sex'	seks sex seksa **brak** razgovor odnos orgazam spoj seksu utroje polijubac provod
	'homosexual'	homoseksualac pacifist katolik **narkoman** prijatelj gay brat glumac gej

Table 1: **Overall Neighbors Comparison**

B Appendix B

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
('discrimination', 'harassment')	0.5983
('nymphomaniac', 'transvestite')	0.5590
'nymphomaniac', 'prostitute'	0.5503
('lesbianism', 'misogyny')	0.5422
('misogyny', 'objectification')	0.5245
('girl', 'nymphomaniac')	0.5163

(a) Similarity Pairs in English

Pair	Similarity
'féminin' - 'masculin'	0.9298
'femme' - 'mère'	0.9081
'fille' - 'mère'	0.9078
'machisme' - 'misogynie'	0.8818
'homophobie' - 'sexisme'	0.8665
'machisme' - 'sexisme'	0.8663
'homosexualité' - 'inceste'	0.8396
'adultère' - 'inceste'	0.8366
('pédé', 'pute')	0.8235
('misogynie', 'sexisme')	0.8230
('pornographie', 'prostitution')	0.8101
('adultère', 'meurtre')	0.7920
('prostitution', 'viol')	0.7915
('homosexualité', 'prostitution')	0.7903
('inceste', 'prostitution')	0.7887
('homophobie', 'pornographie')	0.7842
('homosexualité', 'pornographie')	0.7830
('déconstruction', 'objectivation')	0.7736

(b) Similarity Pairs in French

Pair	Similarity
'femminile' - 'maschile'	0.7837
'madre' - 'padre'	0.7785
'gay' - 'omosessuale'	0.7719
'fidanzato' - 'ragazza'	0.6696
'matrimonio' - 'nozze'	0.6134

(c) Similarity Pairs in Italian

Pair	Similarity
Pair: ('madre', 'padre')	0.7174
Pair: ('atractiva', 'chica')	0.4701
Pair: ('encantadora', 'joven')	0.4685
Pair: ('mujer', 'zorra')	0.3850
Pair: ('hombre', 'viejo')	0.3753

(d) Similarity Pairs in Spanish

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

Figure 3: **Comparison of Most Similar Pairs Items**

C Appendix C

Item	Similarity
lgbt	0.5999
transsexual	0.5616
lesbian	0.5479
bisexual	0.5221
intersex	0.5114
gay	0.4791
sexuality	0.4423
queer	0.4274
homophobia	0.4204
homosexuality	0.4130
bisexuality	0.4090
lgbtq	0.4034
homophile	0.3873
gender	0.3869
anti-violence	0.3767
feminism	0.3716
feminist	0.3605
homosexual	0.3561
lesbianism	0.3548
discrimination	0.3492
transvestite	0.3492
objectification	0.3481
heterosexual	0.3475
genital	0.3453
equality	0.3415
hiv	0.3393
sexual	0.3311
patriarchy	0.3297
identity	0.3219
sexism	0.3101
activism	0.3050
subculture	0.3037
masculinity	0.3018
sex	0.3007
harassment	0.2997
activist	0.2868
male	0.2866
vagina	0.2749
normative	0.2692
effeminate	0.2688

Most Similar Items to 'transgender'

Item	Similarity
male	0.7899
women	0.4615
hermaphroditic	0.4112
woman	0.4110
androgynous	0.4039
effeminate	0.3880
sex	0.3752
feminine	0.3603
gender	0.3558
heterosexual	0.3551
feminist	0.3494
masculine	0.3461
asexual	0.3434
intersex	0.3433
transvestite	0.3413
sexual	0.3351
girl	0.3238
lesbianism	0.3223
genital	0.3210
sapphic	0.3182
prostitute	0.3171
harem	0.3080
vagina	0.3044
sexuality	0.3036
lady	0.2972
nymphomaniac	0.2964
lesbian	0.2962
transsexual	0.2936
patriarchy	0.2903
homosexual	0.2843
objectification	0.2781
young	0.2730
sexy	0.2710
bitch	0.2659
mutilation	0.2627
bitchy	0.2597
homophile	0.2583
chastity	0.2560
masculinity	0.2511
transgender	0.2461

Most Similar Items to 'female'

Item	Similarity
female	0.7899
hermaphroditic	0.4628
heterosexual	0.4507
effeminate	0.4283
gender	0.3907
masculine	0.3808
intersex	0.3691
homosexual	0.3649
sex	0.3636
sexuality	0.3610
asexual	0.3601
androgynous	0.3582
harem	0.3445
sexual	0.3419
women	0.3374
feminine	0.3286
transsexual	0.3247
patriarchy	0.3212
transvestite	0.3212
woman	0.3165
genital	0.3120
sapphic	0.3074
masculinity	0.3018
young	0.3004
lesbianism	0.2923
gay	0.2892
transgender	0.2866
mutilation	0.2863
objectification	0.2840
bisexual	0.2811
castration	0.2774
homosexuality	0.2773
nymphomaniac	0.2721
prostitute	0.2679
lesbian	0.2673
stereotype	0.2629
fecundity	0.2600
sexism	0.2533
subordination	0.2524
feminist	0.2470

Most Similar Items to 'male'

Figure 4: **English Sub-space Most Similar Items**

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
gender	0.1967
decostruzione	0.1910
affascinante	0.1840
misoginia	0.1840
mascolino	0.1803
prostituzione	0.1784
eterosessuale	0.1782
puttana	0.1755
violentare	0.1752
matrimonio	0.1716
oggettivazione	0.1711
flirtare	0.1647
vogue	0.1628
uomo	0.1581

Most Similar Items to 'transgender'

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
bisessuale	0.2548
sentimentale	0.2544
uguaglianza	0.2541
ruolo	0.2504
discriminazione	0.2420
mutilazione	0.2419
ragazza	0.2406
saffico	0.2354
eterosessuale	0.2350
genere	0.2298
sport	0.2255
effeminato	0.2184
arte	0.2147
vogue	0.2141

Most Similar Items to 'femminile'

Item	Similarity
maschile	0.7837
sesso	0.5467
donna	0.4951
mascolino	0.4474
maschilismo	0.4269
sessuale	0.4051
genitale	0.3962
femminismo	0.3583
misoginia	0.3570
androgino	0.3562
femminista	0.3544
stereotipo	0.3144
eterosessuale	0.3075
sessismo	0.2932
uomo	0.2909
effeminato	0.2877
bisessuale	0.2819
ragazza	0.2776
sentimentale	0.2706
immaginario	0.2690
omosessuale	0.2673
lesbismo	0.2666
gender	0.2589
transgender	0.2581
segregazione	0.2565
genere	0.2550
giovane	0.2548
harem	0.2429
corpo	0.2374
incesto	0.2374
gay	0.2277
asessuale	0.2269
transessuale	0.2209
lesbica	0.2176
stupro	0.2082
contraccettivo	0.2058
pronome	0.2058
madre	0.2036
prostituta	0.2034
saffico	0.2034

Most Similar Items to 'maschile'

Figure 5: **Italian Sub-space Most Similar Items**

Item	Similarity
bitch	0.6714
gender	0.6649
féminisme	0.5777
bimbo	0.5747
porno	0.5704
gay	0.5679
postcolonial	0.5678
lesbienne	0.5640
féministe	0.5458
glamour	0.5355
vogue	0.5218
flirt	0.5183
travesti	0.5061
androgyne	0.5056
activiste	0.4989
sexy	0.4982
stéréotype	0.4910
lady	0.4870
pédé	0.4699
machisme	0.4698
misogynie	0.4667
sentimental	0.4598
hermaphrodite	0.4505
déviant	0.4452
orgie	0.4428
déviance	0.4363
grinçante	0.4335
hétérosexuel	0.4191
homosexualité	0.4167
délire	0.4141
activisme	0.4137
sexisme	0.4056
pute	0.4040
homosexuel	0.4033
déconstruction	0.3988
pornographie	0.3973
lgbt	0.3951
salope	0.3924
féminin	0.3919
argot	0.3919

Most Similar Items to 'queer'

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
travesti	0.5220
genre	0.5198
fantaisie	0.5151
désir	0.5135
invisible	0.5091
maîtresse	0.5090
femmes	0.5072
charmant	0.5061
amour	0.5053
hermaphrodite	0.5038
stéréotype	0.5037
homosexuel	0.4987
lady	0.4981
lesbienne	0.4977
garçon	0.4974
mariage	0.4961
délire	0.4948
paradis	0.4940
art	0.4898
honneur	0.4897
argot	0.4872
beau	0.4846
spectaculaire	0.4838
jeune	0.4834

Most Similar Items to 'féminin'

Item	Similarity
masculin	0.9298
sexe	0.7025
sexuel	0.6193
androgyne	0.6055
glamour	0.6035
sexualité	0.5810
stéréotype	0.5706
vogue	0.5653
gay	0.5614
sexy	0.5560
travesti	0.5518
homosexuel	0.5468
sentimental	0.5401
argot	0.5397
hétérosexuel	0.5346
castration	0.5191
pronom	0.5189
hermaphrodite	0.5124
employé	0.5112
imaginaire	0.4991
lesbienne	0.4982
beauté	0.4972
garçon	0.4971
diable	0.4966
porno	0.4966
homosexualité	0.4952
sport	0.4938
parent	0.4918
femme	0.4900
paternité	0.4884
désir	0.4794
esclave	0.4792
invisible	0.4786
homme	0.4782
délire	0.4777
charmant	0.4761
radical	0.4754
déviant	0.4725
fécondité	0.4721
attirant	0.4715

Most Similar Items to 'masculin'

Figure 6: **French Sub-space Most Similar Items**

References

[1] M. Baroni, G. Dinu, and G. Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, 2014.

[2] Christine Basta, Marta R Costa-jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, 2019.

[3] Yang Trista Cao and Hal Daumé III. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, 2020.

[4] Jessica Clarke. They, them, and theirs. *Harvard Law Review*, 132:894, 2019.

[5] Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum*, page 139, 1989.

[6] E. Crouch. Being non-binary in a language without gendered pronouns – estonian. *Deep Baltic*, 2018.

[7] Christian Davenport. *Media bias, perspective, and state repression: The Black Panther Party*. Cambridge University Press, 2009.

[8] Scott C. Deerwester, Susan T. Dumais, George W. Furnas, T. K. Landauer, and Richard A. Harshman. Indexing by latent semantic indexing analysis. *Journal of the Association for Information Science and Technology*, 1990.

[9] N. Desai. The making of a feminist. *Indian Journal of Gender Studies*, 25(2):307–318, 2018.

[10] S Dev, M Monajatipoor, A Ovalle, A Subramonian, et al. Harms of gender exclusivity and challenges in non-binary representation in language technologies. *arXiv preprint arXiv:2108.12084*, 2021.

[11] Catherine D'Ignazio. *Data Feminism*. The MIT Press, Cambridge, Massachusetts, 2020.

[12] Jennifer Earl, Andrew Martin, John D McCarthy, and Sarah A Soule. The use of newspaper data in the study of collective action. *Annual Review of Sociology*, 30:65–80, 2004.

[13] Ethan Fast, Tina Vachovsky, and Michael Bernstein. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proceedings of the International AAAI Conference on Web and Social Media, volume 10*, 2016.

[14] Christine Feraday. *For lack of a better word: Neo-identities in non-cisgender, non-straight communities on tumblr*. PhD thesis, Ryerson University, 2016.

[15] V. Gautam. Guest lecture in pronouns: Vasundhara. In K. Conrod, editor, *Pronoun Studies*, 2021.

[16] Wei Guo and Aylin Caliskan. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. 2020.

[17] F. Günther, C. Dudschig, and B. Kaup. Lsafun-an r package for computations based on latent semantic analysis. *Behavior Research Methods*, 47(4):930–944, 2015.

[18] F. Günther, L. Rinaldi, and M. Marelli. Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, 14(1):101–115, 2019.

[19] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, 2019.

[20] Thomas K. Landauer and Susan T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

[21] Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28:203–208, 1996.

[22] Dian I. Martin and Michael W. Berry. Mathematical foundations behind latent semantic analysis. 2007.

[23] Douglas M McLeod. News coverage and social protest: How the media's protect paradigm exacerbates social conflict. *Journal of Dispute Resolution*, page 185, 2007.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[25] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, page 3111–3119, 2013.

[26] Micah Rajunov and Scott Duane. *Nonbinary: Memoirs of Gender and Identity*. Columbia University Press, 2019.

[27] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[28] Ricardo Simmonds. Virginia eubanks (2018) automating inequality: How high-tech tools profile, police and punish the poor. new york: St. martin's press. 260 pages. eisbn: 9781466885967. *Science & Technology Studies*, 2019.

[29] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[30] Yi Chern Tan and L Elisa Celis. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*, pages 13230–13241, 2019.

[31] Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. Rtgender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018. European Language Resources Association (ELRA).

[32] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5309–5318. IEEE, 2019.

[33] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617, 2018.

[34] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. 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)*, pages 629–634, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.