Scientists & Women Scientists:

Exploring Biases in the Use of Gendered Categories

Prototype Effects in Categories

Not all members of a category are equally representative (Rosch, 1973)

Birds

Less Typical — **More Typical**

6i, -7+9i, …
(Complex Numbers)

π, e, √2, …
(Irrational Numbers)

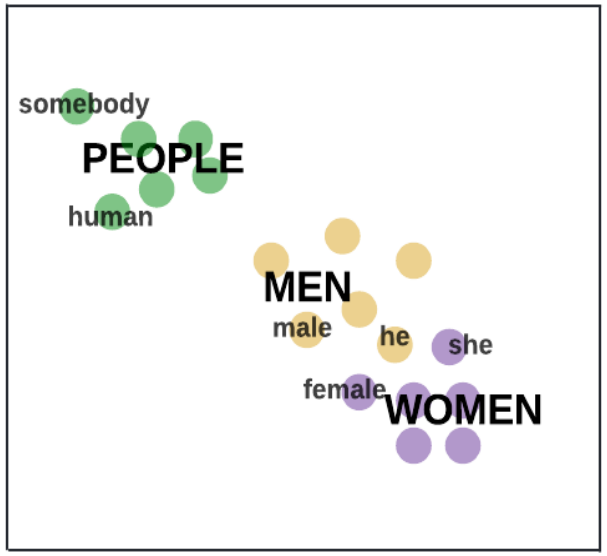
-1, -2, -3, …
(Negative Numbers)

1, 2, 3…
(Natural Numbers)

Numbers

Androcentrism

sim(PEOPLE, MEN) >
sim(PEOPLE, WOMEN)



Based on Bailey, A. H., Williams, A., & Cimpian A. (2022)

PEOPLE = MALE BIAS

(Silveira, 1980)

Table I. Percentage of Subjects Attributing Male Gender to a Gender-Unspecified Individual as a Function of Subject Gender and Script Condition

	Subject Gender (%)		
Script	Male (n = 85)	Female (n = 82)	Combined (n = 167)
Business	88	88	88
Interpersonal	67	74	70
Education	81	60	71
Mean	79	73	76

From Davis, R. M., & Kok, C. J. (1995)

(a) **woman** (16), surfer (14)

(b) **surfer** (24), man (6), person (2), boy (2)

(c) **girl** (11), skateboarder (7), skater (6), child (6)

(d) **skater** (9), skateboarder (6), boy (6), kid (3)

From Harrison, S., Gualdoni, E., & Boleda, G. (2023).

A Social Cognitive Account of Androcentrism

Bailey, A. H., LeFrance, M., & Dovidio J.F. (2019)

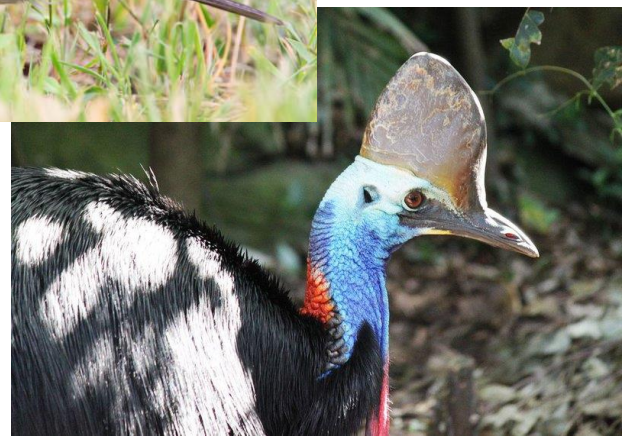
Men are more **frequently instantiated** as examples of humans

Stereotypically masculine traits & attributes form human **category ideals**

Atypicality & Language

Atypical category members are often marked by language

"Birds"



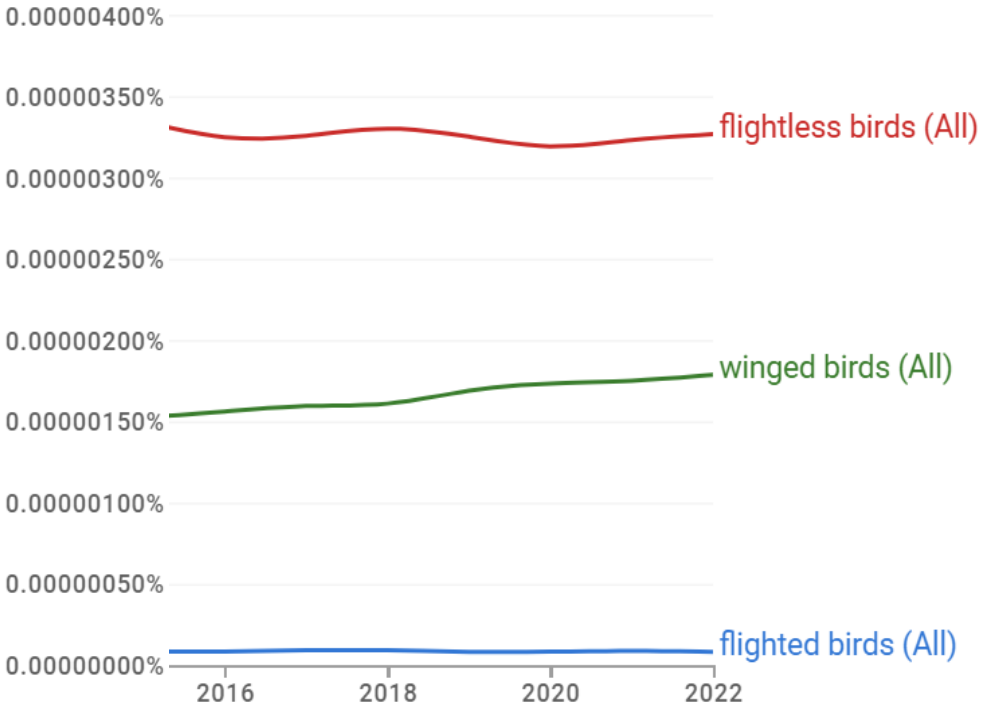
"Flightless birds"

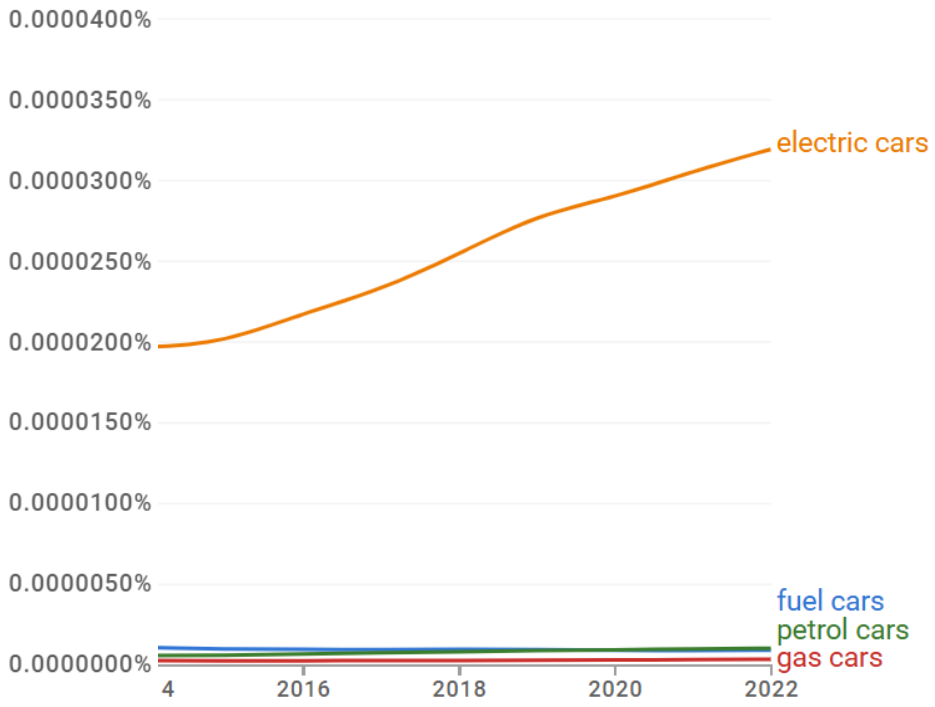
"Cars"

"Electric cars"

Atypicality & Language

Atypical category members are often marked by language





Atypicality & Language

If women are the atypical person than there should be more categories that use gendered language to specify that these categories are for women

AND

These categories should appear more frequently in language

Project Goals

1. Quantify asymmetries in the existence of categories marked by gendered language across multiple human category systems

2. Understand how these asymmetries have changed with time

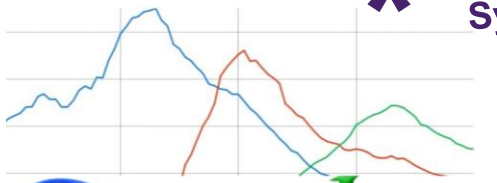
3. Understand how these asymmetries differ between more institutional and more cultural category systems

Category Systems

The LCSH

Macaws in art *(Not Subd Geog)*
Macayo Site (Nogales, Ariz.)
USE El Macayo Site (Nogales, Ariz.)
MacBean family
USE Bean family
Macbeth, Hamish (Fictitious character)
(Not Subd Geog)
UF Hamish Macbeth (Fictitious character)
MacBook (Computer) *(Not Subd Geog)*
UF Apple MacBook (Computer)

* Noun phrases from Syntactic Google Ngram (2008)



Google books
NGRAM VIEWER

Institutional ⟷ **Cultural**

WIKIPEDIA
The Free Encyclopedia

Data Collection

Based on Falenska, A., & Cetinoglu, O. (2021)

1. Collect subject headings & Wikipedia pages that contain gendered words

Men	Women
men, masculine, male, gentlemen	women, feminine, female, ladies

2. Find all parent categories of gendered categories that exist in a systems

Married women
↔
Married people

Male dentist
↔
Dentist

Data Collection

Based on Falenska, A., & Cetinoglu, O. (2021)

3. Divide into the following cases

Cases	Example (W; M; G)
WMG	women caregivers; male caregivers; caregivers
WM	wild women in art; wild men in art ; NONE
WG	women doctoral students; NONE; doctoral students
MG	NONE; male fans; fans
W	women and the sea; NONE; NONE
M	NONE; uncircumcised men; NONE

Wikipedia

From Falenska, A., & Cetinoglu, O. (2021)

	English	German	Polish	Turkish
W | M | G	129	333	128	15
W | M	18033	3480	3712	896
W | G	4438	1895	407	161
M | G	155	52	10	4
W	13641	2971	1321	1328
M	20752	2491	4144	670
TOTAL	57148	11222	9722	3074

Masculine Generic Bias

The LCSH

Cases	Counts
WMG	120
WM	78
WG	1660
MG	88
W	633
M	95
Total	2674

Gender Asymmetry:

Women	92.61%
Men	7.39%

12.5× more for women

The LCSH

Cases	Counts
WMG	120
WM	78
WG	1660
MG	88
W	633
M	95
Total	2674

Gender Markedness:

A concept is marked by gender if a gendered subcategory exists for one gender but not the other.

Women	94.97%
Men	5.03%

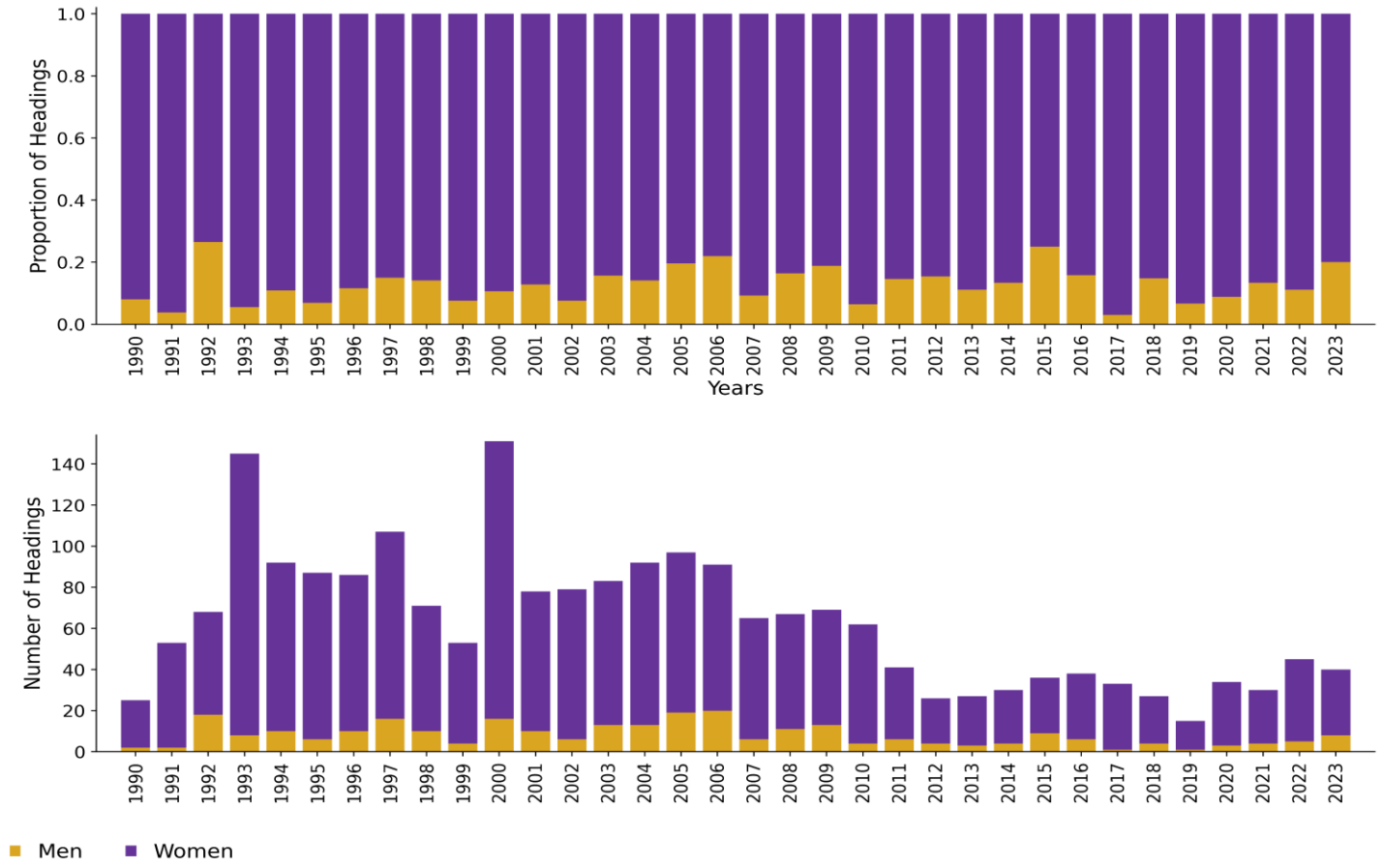
18.9× more for women

The LCSH

Headings more or less consistent with language use in Google books

	female_freq > male_freq	male_freq > female_freq
female_but_not_male	**1106**	261
male_but_not_female	13	**67**
male_and_female	101	43

Has the LCSH Changed?



Not if you look at the proportion of gendered headings added each year

Has the LCSH Changed?

	Asymmetry		Markedness	
	% W	Total	% W	Total
1994	**92.14**	967	**95.77**	567
2004	**93.13**	1731	**95.53**	1185
2014	**92.87**	2201	**95.13**	1559
2024	**92.61**	2476	**94.97**	1748

Not if you look at asymmetry & markedness

Wikipedia

	English
W | M | G	129
W | M	18033
W | G	4438
M | G	155
W	13641
M	20752
TOTAL	57148

Gender Asymmetry:

Women	46.37%
Men	53.63%

1.2× more for men

Wikipedia

	English
W | M | G	129
W | M	18033
W | G	4438
M | G	155
W	13641
M	20752
TOTAL	57148

Gender Markedness:

Women	96.63%
Men	3.37%

28.6× more for men

Have yet to investigate if it changes!

Summary so Far

The LCSH has strong gender bias that has not changed in the past 30 years

Wikipedia has strong **gender markedness** but not **gender asymmetry**

What’s Going on with Wikipedia?

Method for data processing

Has individual category members

Not constrained by literary warrant

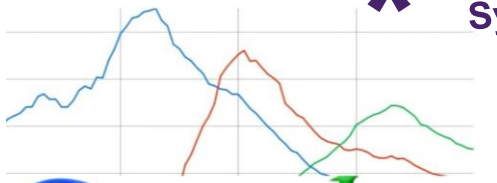
More modern, less institutional

Category Systems

The LCSH

Macaws in art *(Not Subd Geog)*
Macayo Site (Nogales, Ariz.)
USE El Macayo Site (Nogales, Ariz.)
MacBean family
USE Bean family
Macbeth, Hamish (Fictitious character)
(Not Subd Geog)
UF Hamish Macbeth (Fictitious character)
MacBook (Computer) *(Not Subd Geog)*
UF Apple MacBook (Computer)

* Noun phrases from Syntactic Google Ngram (2008)

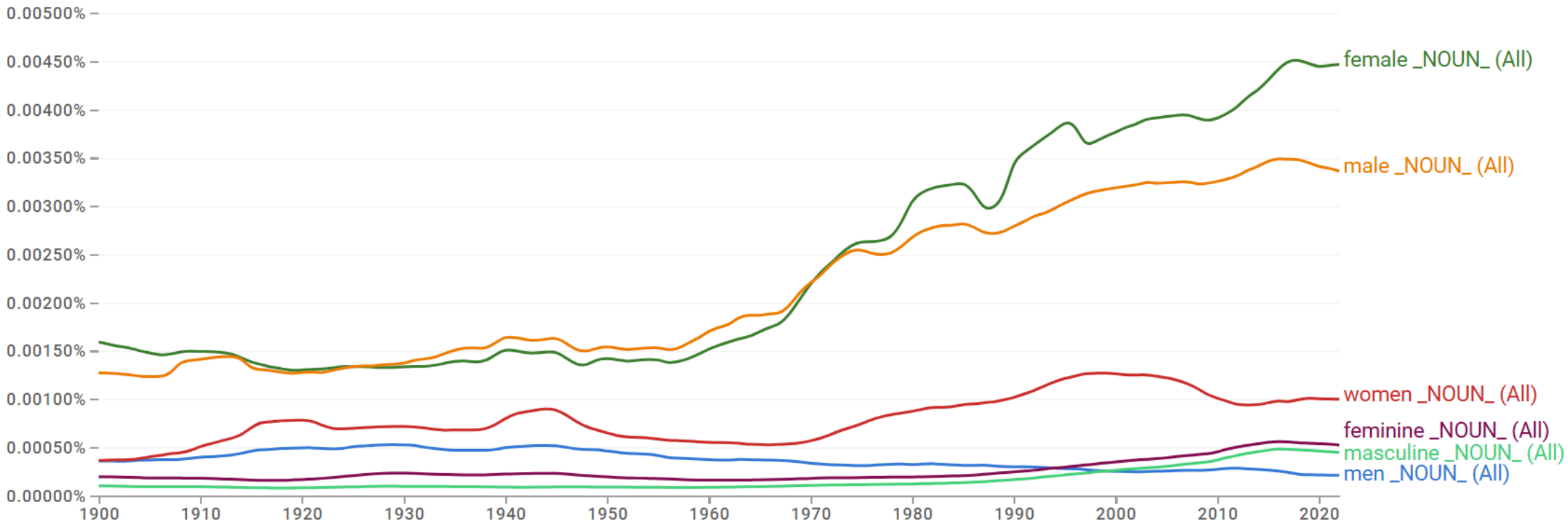


Google books
NGRAM VIEWER

Institutional ⟷ **Cultural**

WIKIPEDIA
The Free Encyclopedia

What about Language?

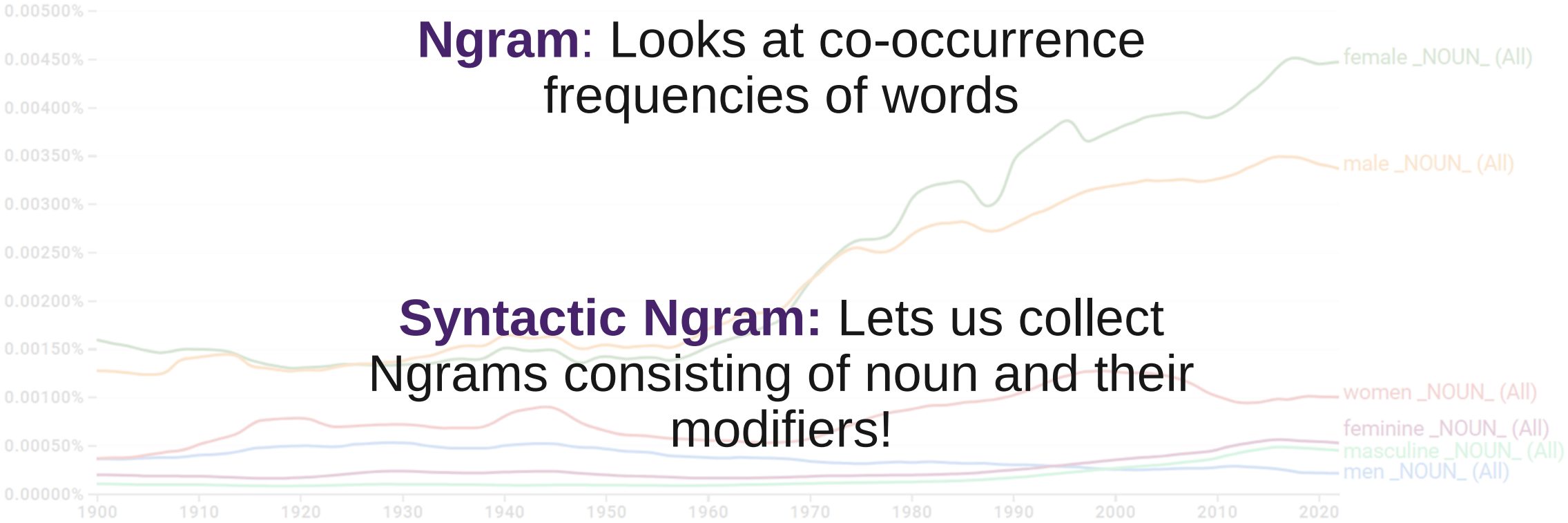


From English Google Ngram Viewer

What about Language?

Ngram: Looks at co-occurrence frequencies of words

Syntactic Ngram: Lets us collect Ngrams consisting of noun and their modifiers!



From English Google Ngram Viewer

Syntactic Ngram

1. Collect gendered noun phrases where a non-gendered noun is the head and a gendered word is one of the modified

```
scientists   women scientists      women/NNS/nn/2 scientists/NNS/pobj/0      2499
scientists   white male scientists   white/JJ/amod/3 male/JJ/amod/3 scientists/NNS/pobj/0    31
scientists   famous women scientists famous/JJ/amod/3 women/NNS/nn/3 scientists/NNS/pobj/0   13
scientists   male scientists male/JJ/amod/2 scientists/NNS/nsubj/0   459
```

2. Combine phrases that only differ in the form of the gendered word

women scientists + female scientists → <W> scientists

Syntactic Ngram

4. Line up in order of decreasing frequency

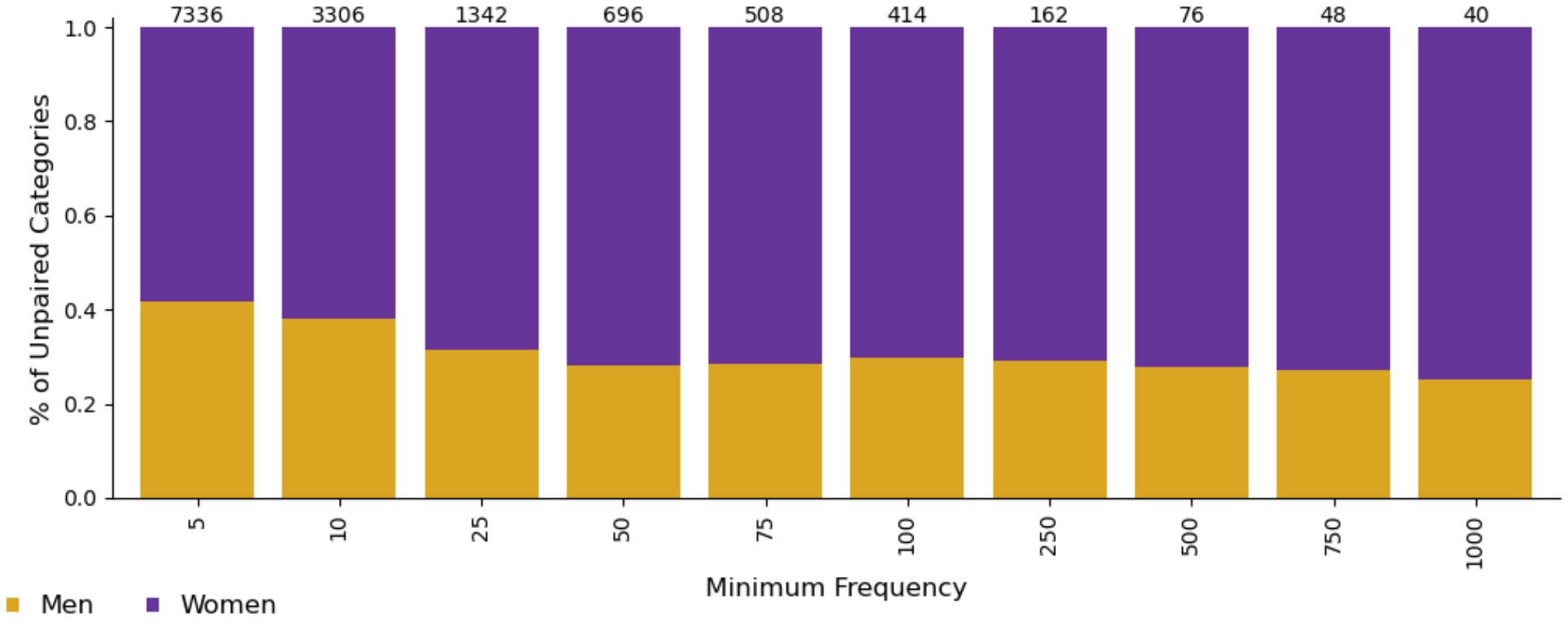
5. Count and compare unpaired NPs for women and unpaired NPs for men above various frequency thresholds

> 5000

> 2500

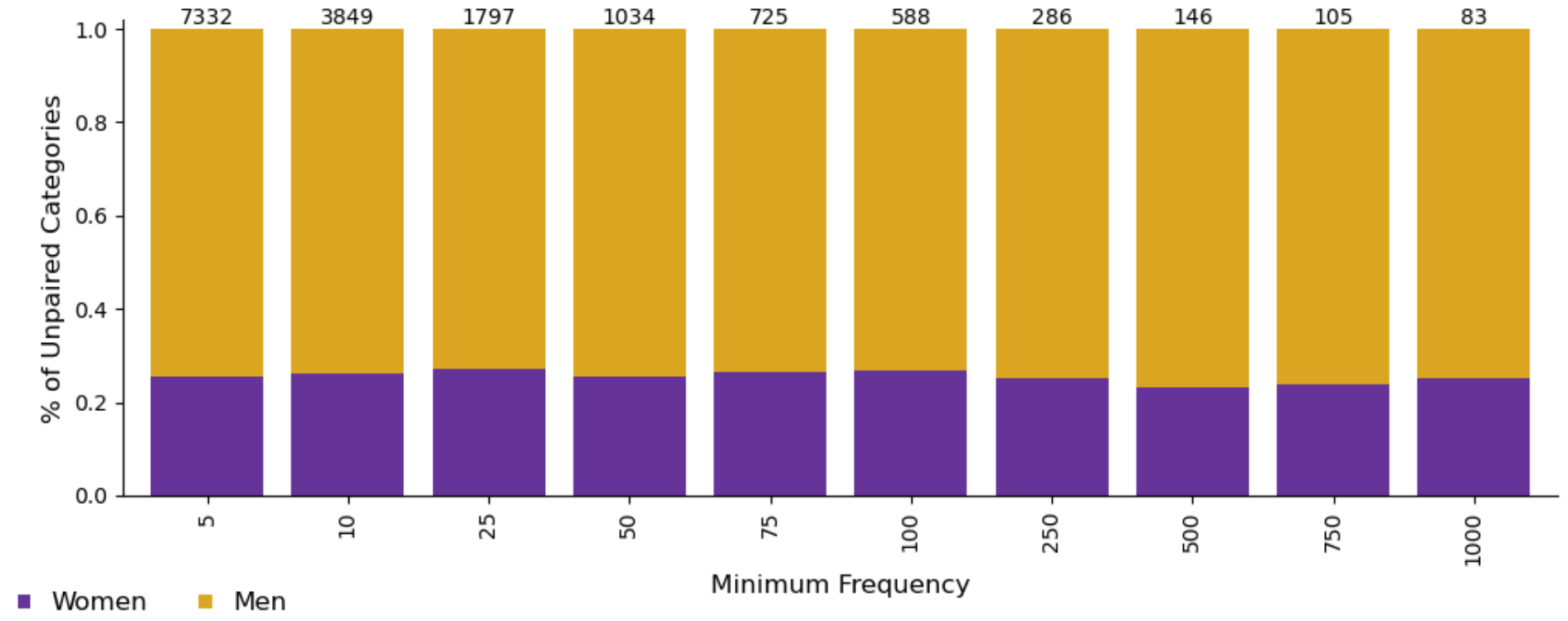
	HEAD	PHRASE	FREQ
2	minds	<M> 's minds	6536
3	rights	<W> 's rights	6512
4	sex	<W> sex	3520
5	movement	<W> 's movement	3393
6	workers	<W> workers	3318
7	hearts	<M> 's hearts	3060
8	lives	<W> 's lives	2846
9	students	<W> students	2661
10	voice	<W> voice	2657
11	sex	<M> sex	2432
12	body	<W> body	2425
13	work	<W> 's work	2346
14	souls	<M> 's souls	2211
15	education	<W> education	2193
16	issues	<W> 's issues	2151

Syntactic Ngram



An Adjective Modifying Men/Women

i.e. Canadian women, young men



Atypicality & Language

If women are the atypical person than there should be more categories that use gendered language to specify that these categories are for women

AND

Yes, especially in the LCSH

These categories should appear more frequently in language

I think so

What's Next

Run my own data collection on Wikipedia

Look at both Wikipedia and Syntactic Ngram across time

Take a closer look at the kinds of headings/pages/noun phrases that are marked

Concerns I have

1. I’m just counting things
2. It’s not distinct enough from existing work in this space
3. I’m picking and choosing to much when it comes to what Ngrams I consider relevant
4. Wikipedia, LCSH, Syntactic Ngram are all over different time spans

Thoughts?