First Women, Second Sex : Gender Bias in Wikipedia

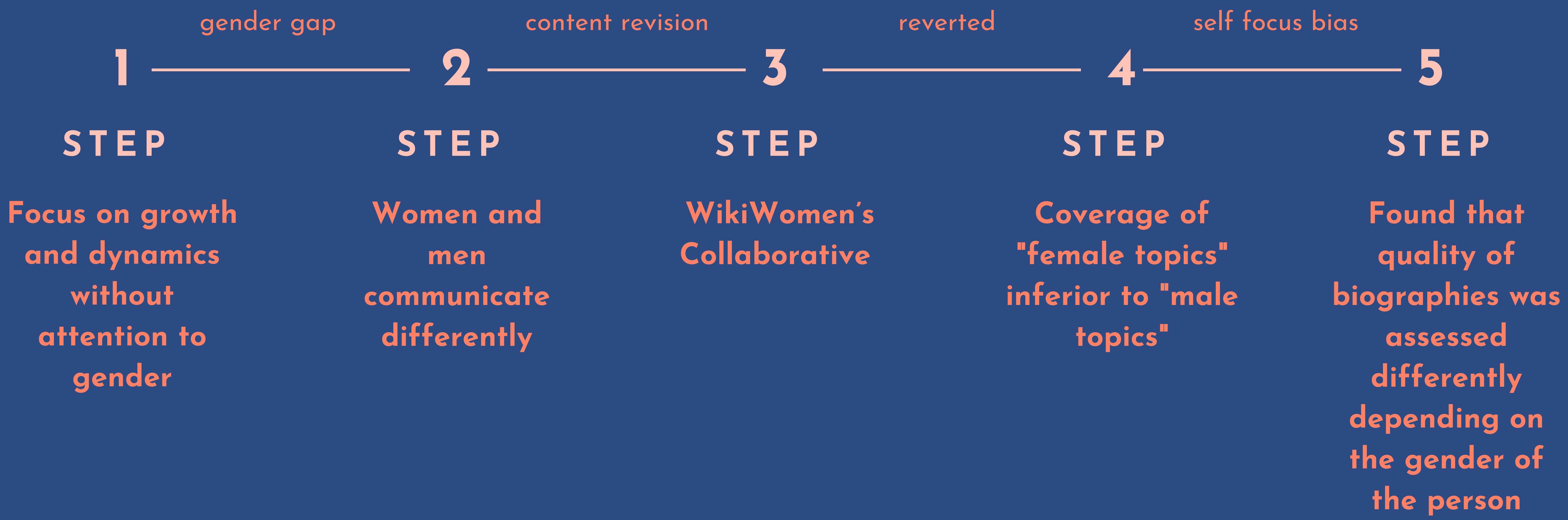
Group 10

Nandini Bajaj - 18CY20020
Pratyush Jaiswal - 18EE35014
Nuruddin Jiruwala - 18EE35022
Abhinav Japesh - 18EE35023

“it is not women’s inferiority that has determined their historical insignificance: it is their historical insignificance that has doomed them to inferiority”

Simone de Beauvoir

Evolution of Wikipedia



1 STEP

Focus on growth and dynamics without attention to gender

gender gap

2 STEP

Women and men communicate differently

content revision

3 STEP

WikiWomen’s Collaborative

reverted

4 STEP

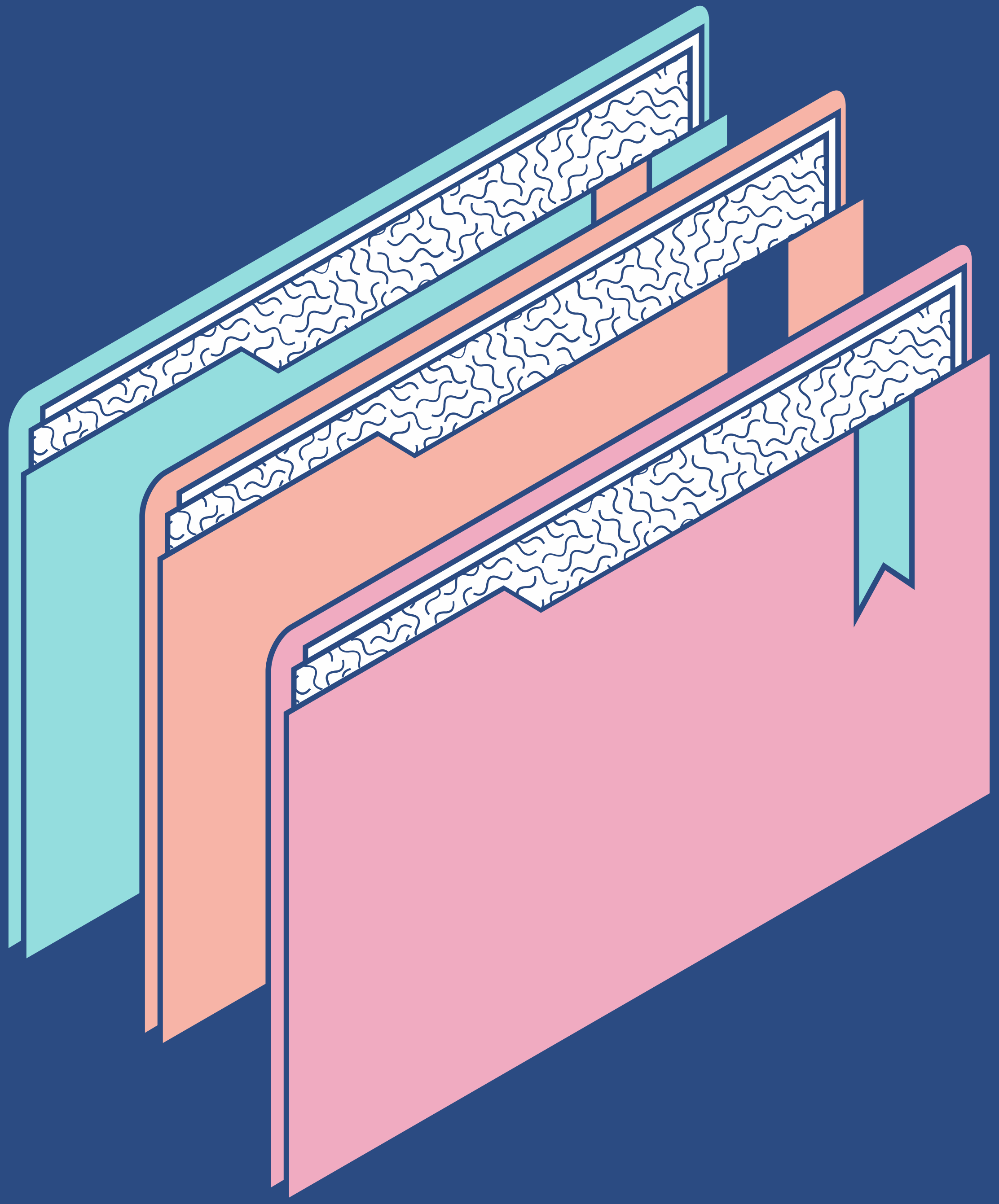
Coverage of "female topics" inferior to "male topics"

self focus bias

5 STEP

Found that quality of biographies was assessed differently depending on the gender of the person

Let's answer some why's?



WHY STUDY GENDER BIAS AND WHY BIOGRAPHIES?

- community of Wikipedians is not diverse in terms of gender, as women represent only 16% of editors (gender gap)
- excluded from the category “American Novelists” included in “American Women Novelists.”
- biographies are a good source given that an article is about a specific person
- enables the identification of cultural differences in content and coverage as well as the construction of social networks of historical (and current) figures

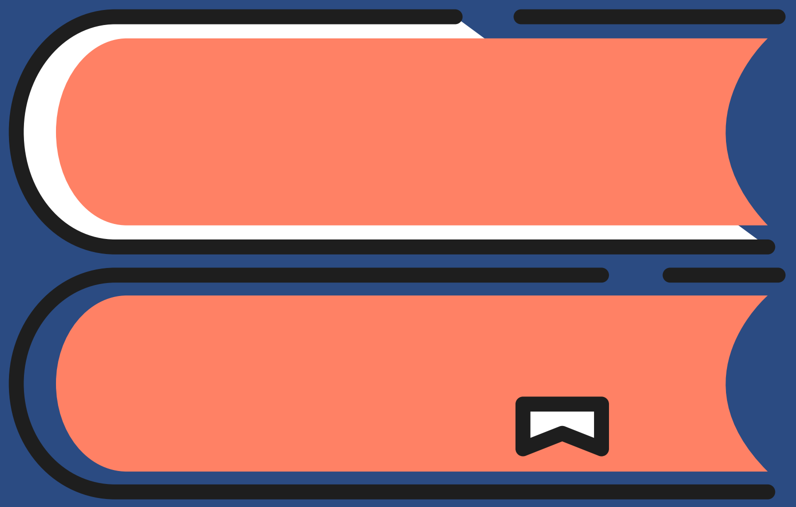
Research Questions

Is there a gender bias in biographies of men and women in Wikipedia?
If so, how to identify and quantify it?
Can it be contextualized based on social theory?

Data Sources (open source)

- provides meta-data for articles, URIs, , normalized links between articles, categorization into a shallow ontology, which includes a Person category
- processed from infoboxes
- classifies infobox matching person template to person class
- maps infobox properties to specific fields in a person's meta-data

DBpedia 2014

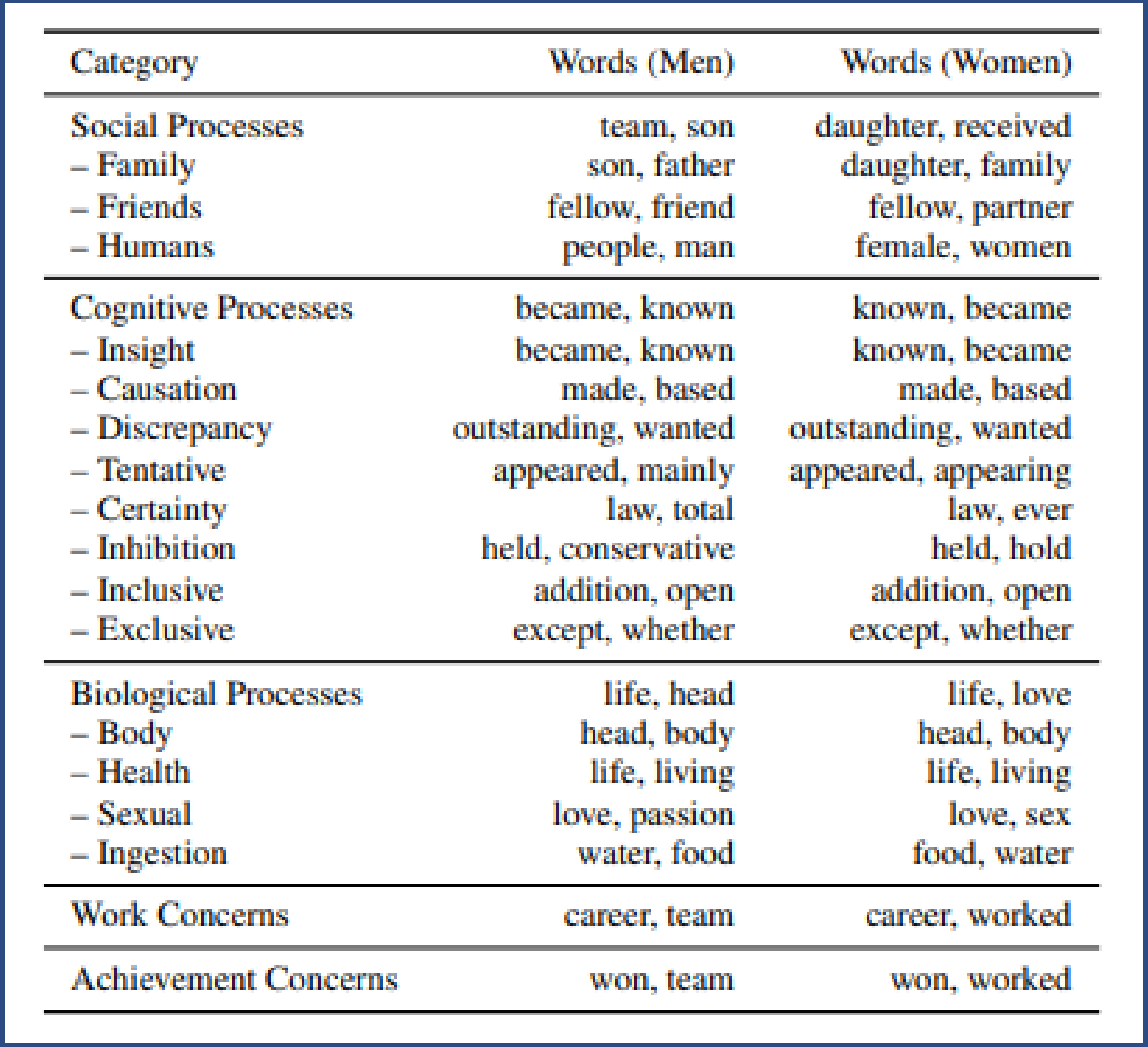


Wikimedia Datadump

- overview- analyze full vocabulary
- wikipedia has a manually built overview which might include bias
- full text- analyze words pertaining to LIWC dictionaries
- LIWC- to find if different genders have different characterizations according to those semantic categories

Linguistic Inquiry and Word Count (dictionaries)

- contains a list of words and prefixes for every category
- Social Processes, Cognitive Processes, Biological Processes, Work Concerns, and Achievement Concerns
- Other categories not considered are Positivity, Negativity, Relativity, Religion, and Death as they can be used in different context
- Data cleaning like: Virginia matches virgin* from the Sexual category, Victoria matches victor* from the Achievement Concerns category
- manually cleaned dictionary contains 2877/4500 words



Category	Words (Men)	Words (Women)
Social Processes	team, son	daughter, received
– Family	son, father	daughter, family
– Friends	fellow, friend	fellow, partner
– Humans	people, man	female, women
Cognitive Processes	became, known	known, became
– Insight	became, known	known, became
– Causation	made, based	made, based
– Discrepancy	outstanding, wanted	outstanding, wanted
– Tentative	appeared, mainly	appeared, appearing
– Certainty	law, total	law, ever
– Inhibition	held, conservative	held, hold
– Inclusive	addition, open	addition, open
– Exclusive	except, whether	except, whether
Biological Processes	life, head	life, love
– Body	head, body	head, body
– Health	life, living	life, living
– Sexual	love, passion	love, sex
– Ingestion	water, food	food, water
Work Concerns	career, team	career, worked
Achievement Concerns	won, team	won, worked

Inferred gender for Wikipedia biographies by Bamman and Smith

- contains inferred gender for biographies based on count of pronouns used
- tested on 500 random bigraphies with 100% precision and 97.6% recall
- considers only male and female
- used to match article URIs with this dataset to get gender meta-data

META-DATA PROPERTIES

A SHORT EXPLANATION OR SUMMARY OF WHAT THE DATA IS

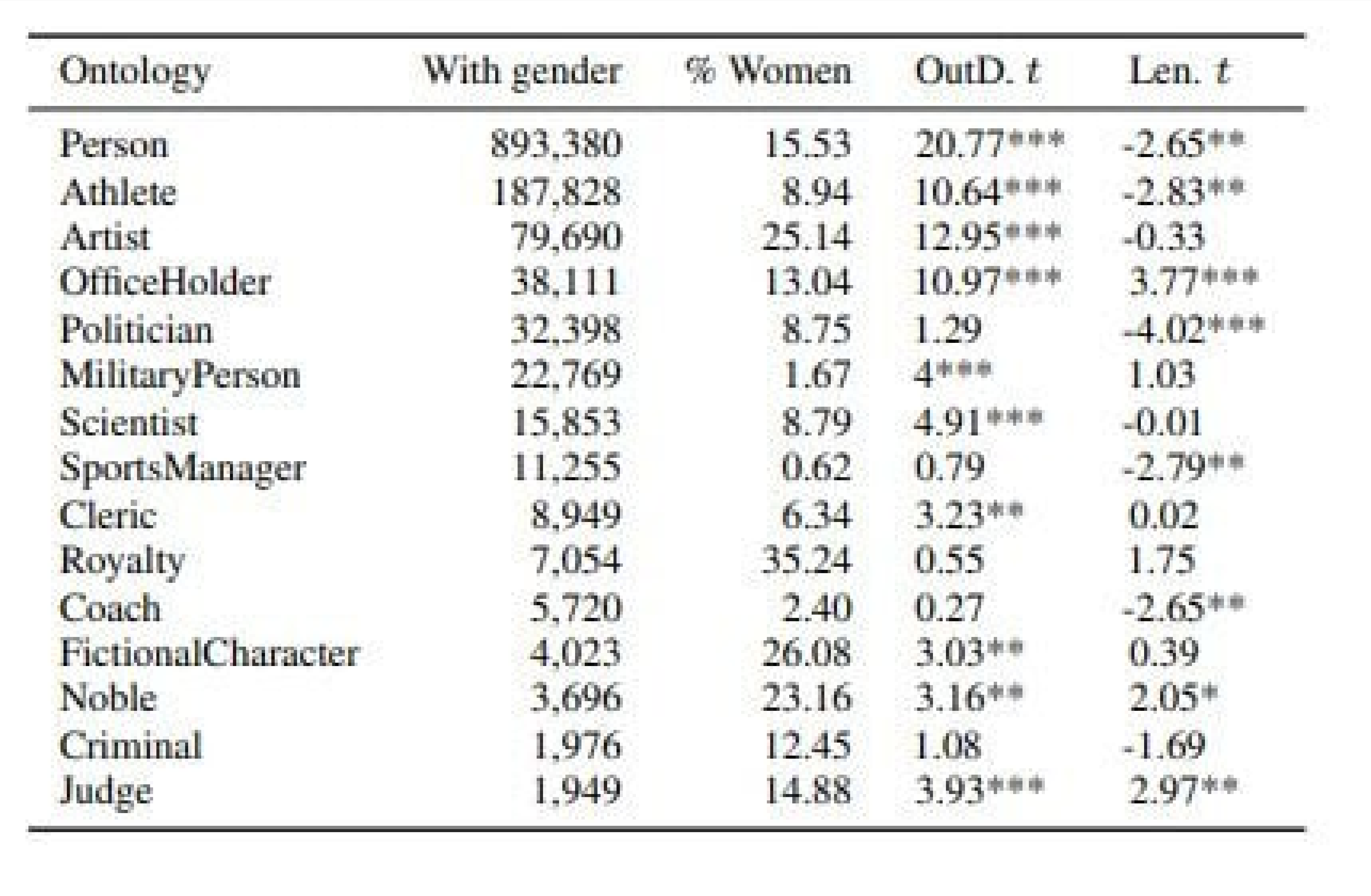
Metadata is used by browsers (how to display content or reload the page), search engines (keywords), and other web services.

Presence and Proportion According to Class

DBPedia estimates the length (in characters) of each article and provides the network of links between articles.

- Of the set of 1,445,021 biographies (articles in the DBpedia Person class), 893,380 (61.82%) have gender metadata. Of those, only 15.5% are women
- Mean article length is 5,955 characters for men and 6,013 characters for women.
- Mean out-degrees (number of links) of 42.1 for men and 39.4 for women also differ significantly

Number of biographies in the dataset for the Person class and its most common child classes (in terms of biographies with gender). OutD means Out Degree, and Len means Length.



Ontology	With gender	% Women	OutD. t	Len. t
Person	893,380	15.53	20.77***	-2.65**
Athlete	187,828	8.94	10.64***	-2.83**
Artist	79,690	25.14	12.95***	-0.33
OfficeHolder	38,111	13.04	10.97***	3.77***
Politician	32,398	8.75	1.29	-4.02***
MilitaryPerson	22,769	1.67	4***	1.03
Scientist	15,853	8.79	4.91***	-0.01
SportsManager	11,255	0.62	0.79	-2.79**
Cleric	8,949	6.34	3.23**	0.02
Royalty	7,054	35.24	0.55	1.75
Coach	5,720	2.40	0.27	-2.65**
FictionalCharacter	4,023	26.08	3.03**	0.39
Noble	3,696	23.16	3.16**	2.05*
Criminal	1,976	12.45	1.08	-1.69
Judge	1,949	14.88	3.93***	2.97**

Date of Birth

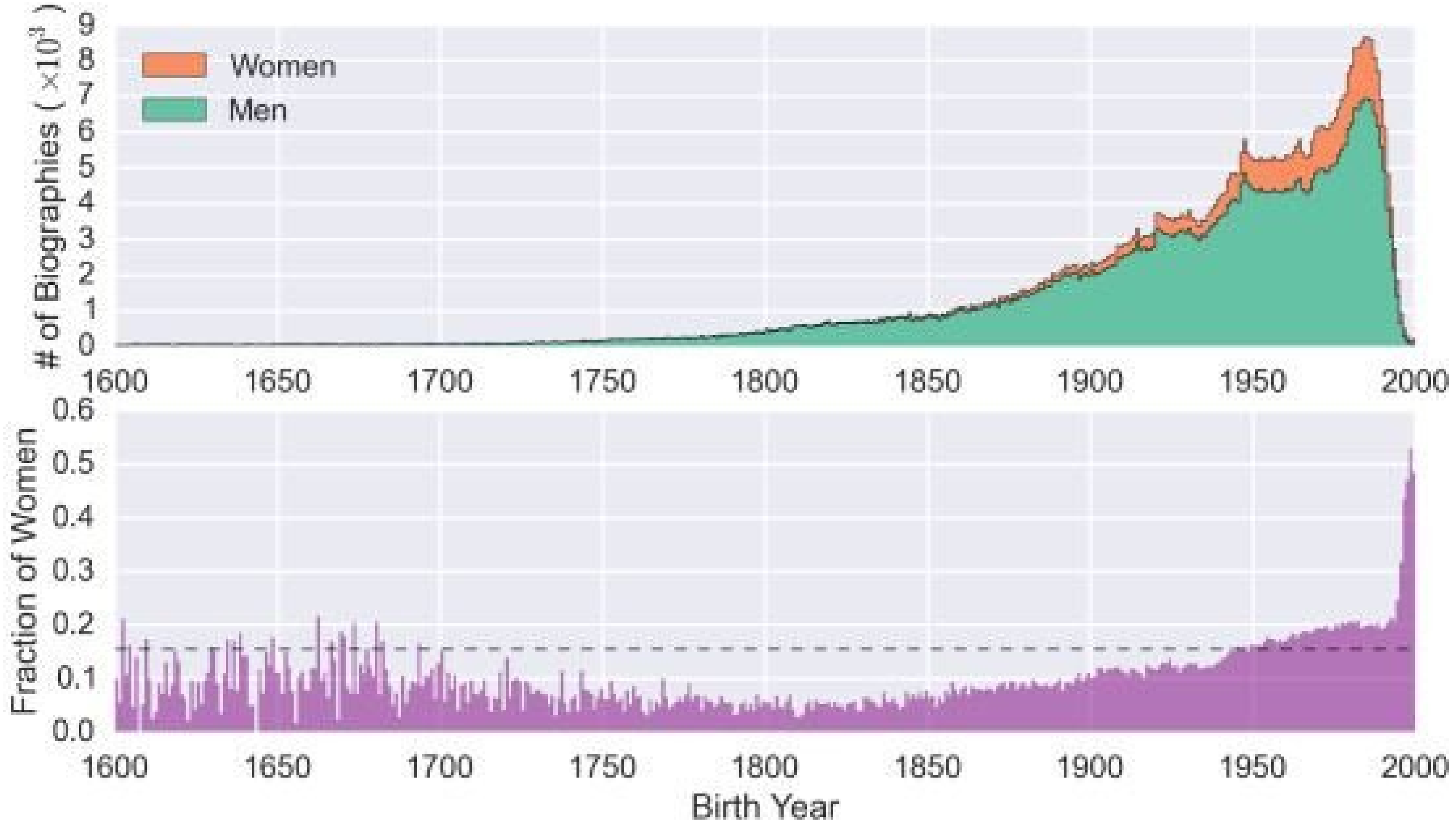


Figure 2: Distribution of biographies according to birth year.

This accounts for 65.48% of biographies with gender (note that 34.07% do not have the date of birth in meta-data).

- Most of the biographies of both genders are about people from modern times
- 53% of Biographies having date of birth in their meta-data are from 1943 until 2000.

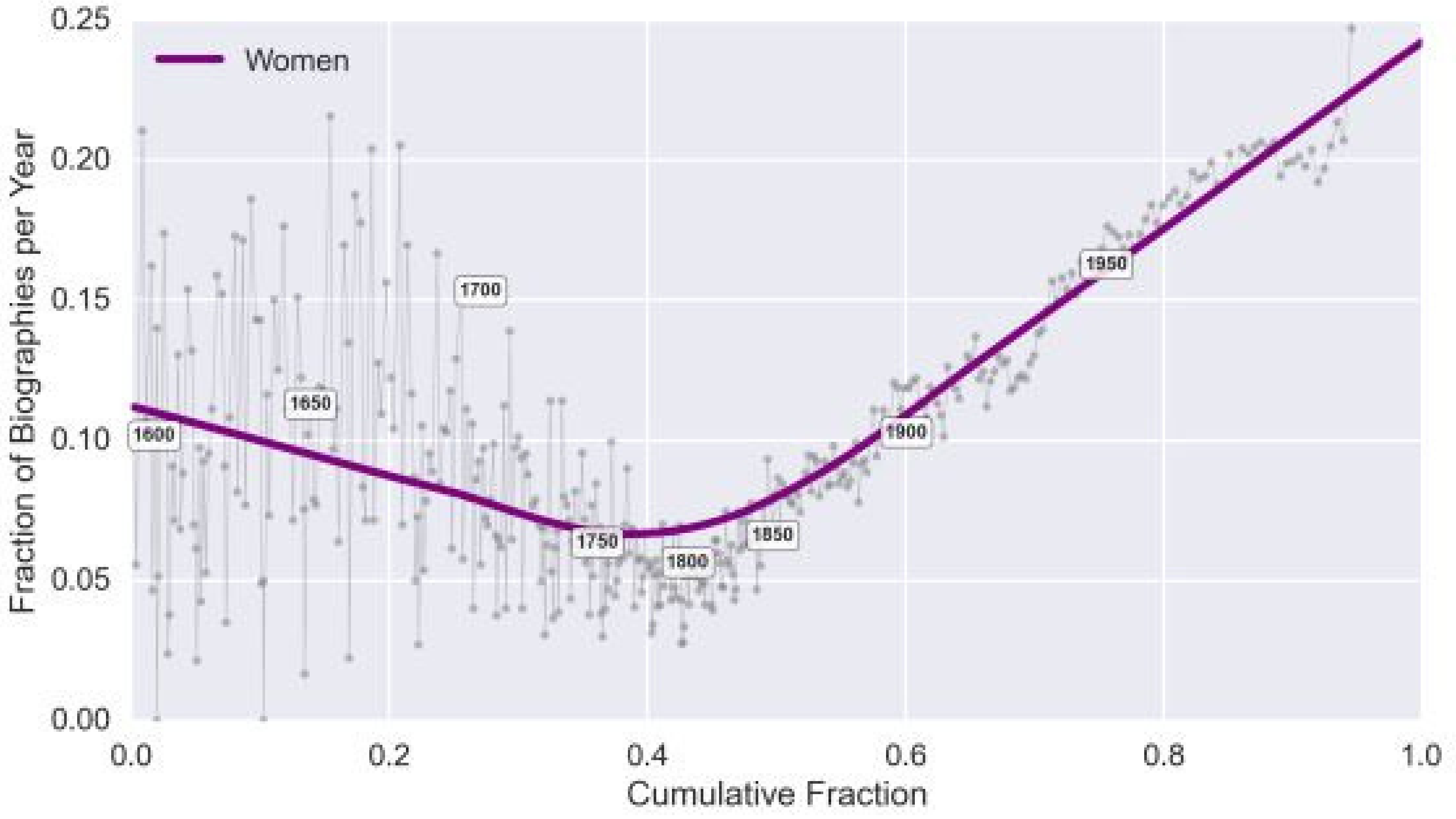


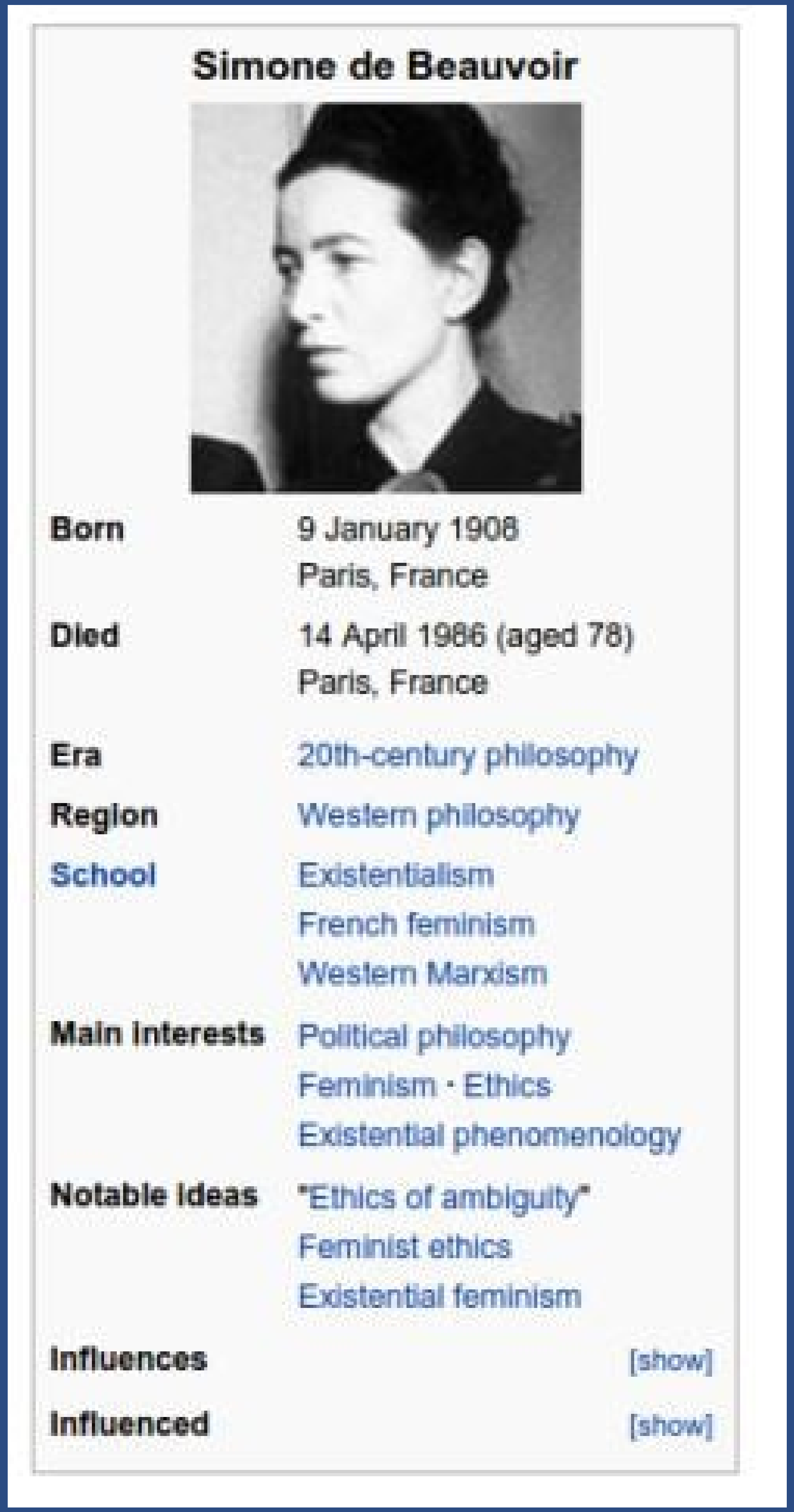
Figure 3: Relation between the cumulative fraction of women in time and the fraction of women per year (dots). The y-axis was truncated to 0.25 for clarity.

Since the year 1943 the fraction of women is consistently above the global fraction of 0.155.

This growth might be explained by the social and cultural changes embraced by people from generations during 1943.

Infobox Attributes

- In total, 340 meta-data attributes were identified from info-boxes that can be included in biographies
- For each one of them, we counted the number of biographies that contained it, and then compared the relative proportions between genders with a chi-square test. Only 3.53% presented statistically significant differences

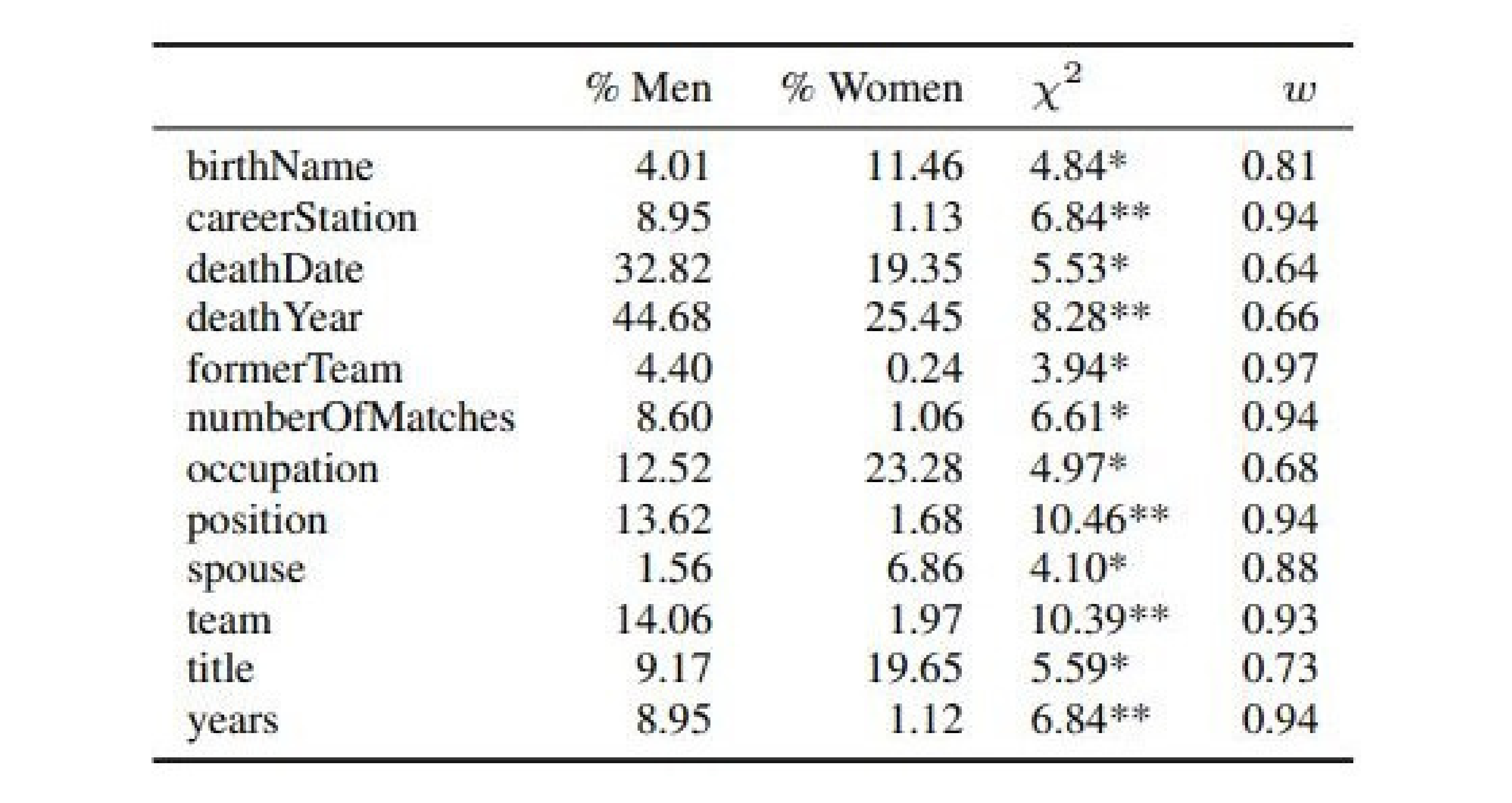


Simone de Beauvoir

Born	9 January 1908 Paris, France
Died	14 April 1986 (aged 78) Paris, France
Era	20th-century philosophy
Region	Western philosophy
School	Existentialism French feminism Western Marxism
Main interests	Political philosophy Feminism · Ethics Existential phenomenology
Notable ideas	"Ethics of ambiguity" Feminist ethics Existential feminism
Influences	[show]
Influenced	[show]

Observations

- Attributes careerStation, formerTeam, numberOfMatches, position, team, and years are more frequent in men.
- Attribute birthName is more frequent in women. Its values refer mostly to the original name of artists, and women have considerable presence in this class
- Attributes occupation and title are more frequent in women, and seem to serve the same purpose but through different mechanisms.
- The spouse attribute is more frequent in women. This attribute indicates whether the portrayed person was married or not, and with whom. In some cases, it contains the resource URI of the spouse, while in other cases, it contains the name



	% Men	% Women	χ^2	w
birthName	4.01	11.46	4.84*	0.81
careerStation	8.95	1.13	6.84**	0.94
deathDate	32.82	19.35	5.53*	0.64
deathYear	44.68	25.45	8.28**	0.66
formerTeam	4.40	0.24	3.94*	0.97
numberOfMatches	8.60	1.06	6.61*	0.94
occupation	12.52	23.28	4.97*	0.68
position	13.62	1.68	10.46**	0.94
spouse	1.56	6.86	4.10*	0.88
team	14.06	1.97	10.39**	0.93
title	9.17	19.65	5.59*	0.73
years	8.95	1.12	6.84**	0.94

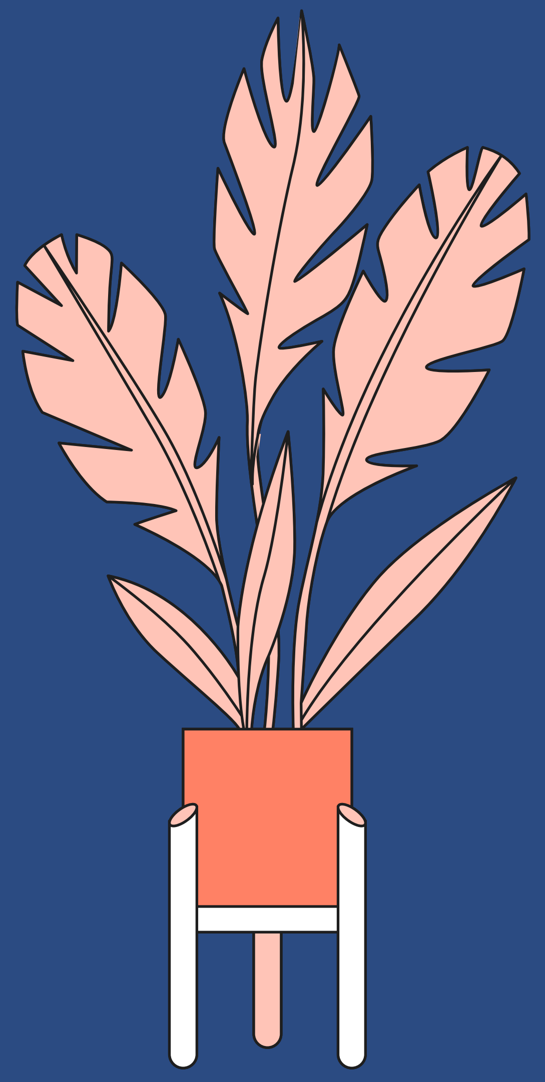
Proportion of men and women who have the specified attributes in their infoboxes

Discussions

It can be found that there are statistically significant differences in biographies of men and women.

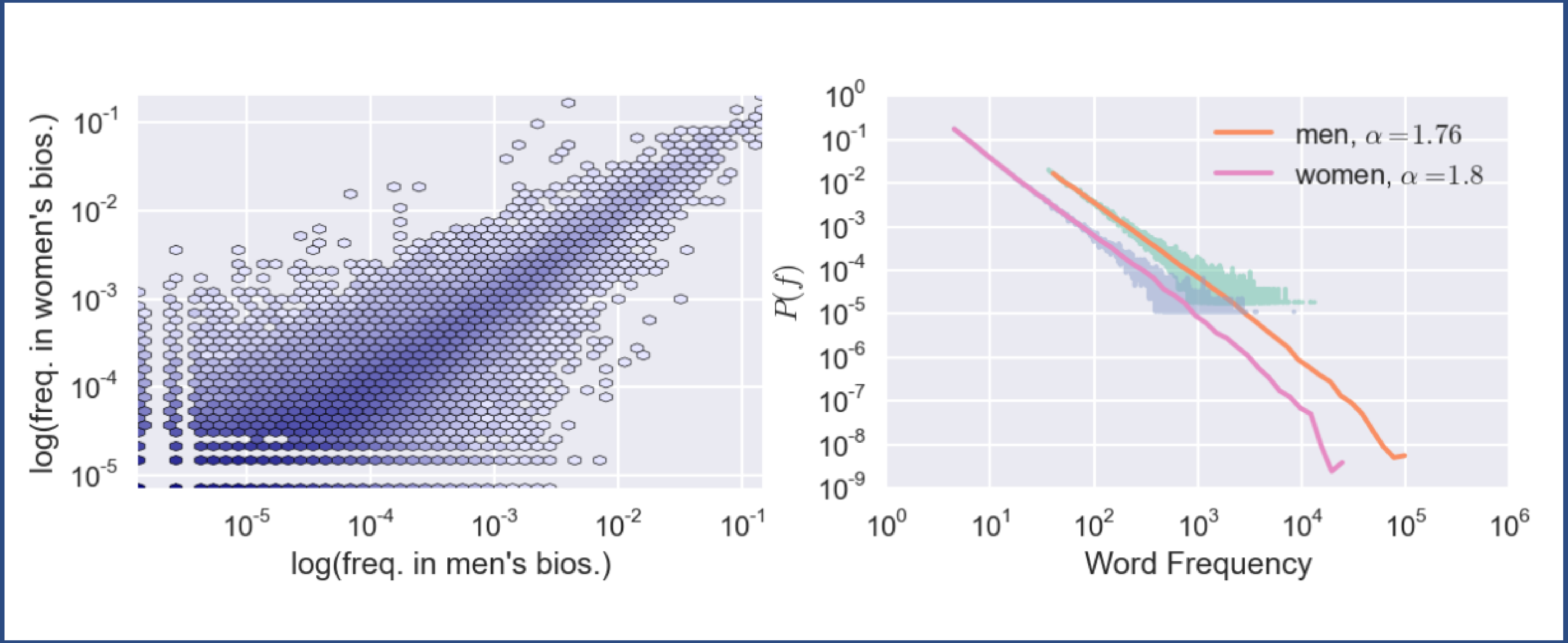
The greater frequency of the spouse attribute in women can be interpreted as specific gender roles attributed to women.

The imbalance found in the Artist (women) and Athlete (men) classes is not a sign of bias from Wikipedians. Instead, it could be a reflection of physical world phenomena under study by the social sciences.



Language Properties

- Explore characterization of men and women in a lexical perspective
- Use the estimated frequencies to find which words are associated with each gender
- Words considered once per biography along with estimating bi-gram word collocations to identify composite concepts
- Vocabulary of size Vm=1,013,305 for men, Vw=376,737 for women with V=272,006 common words

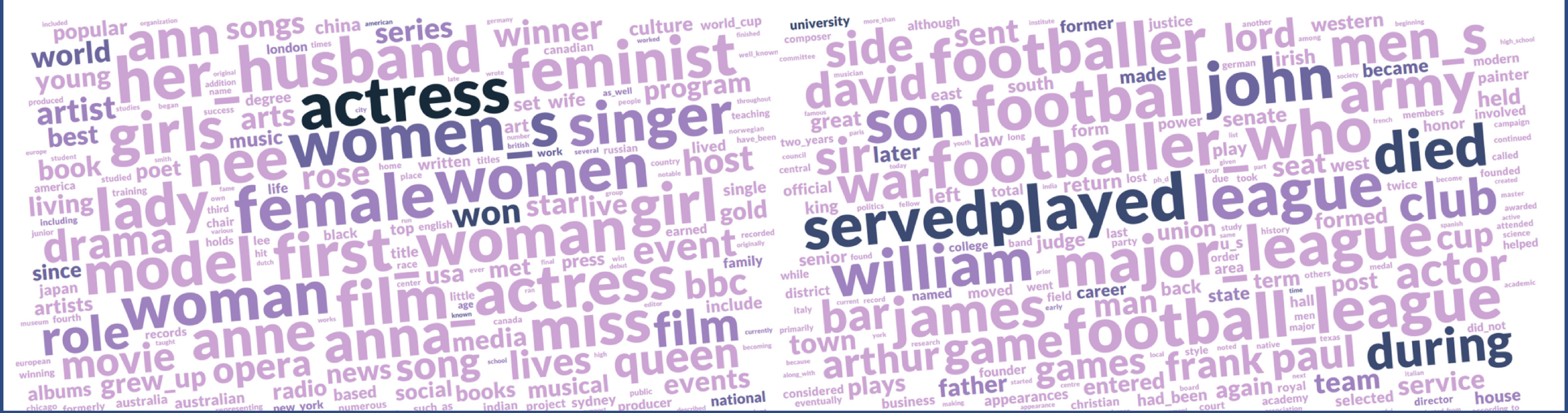


- Similar frequency distribution across genders.
- Word frequencies distribution follow zipf distribution with alpha=1.8
- Frequency with respect to gender presents high correlation value of 0.65
- Interlanguage rank correlation of same meaning of 0.54
- Implies words share meanings when referring to men and women

Associativity of words with gender

- Strength of association through Pointwise Mutual Information over common vocabulary V for both genders

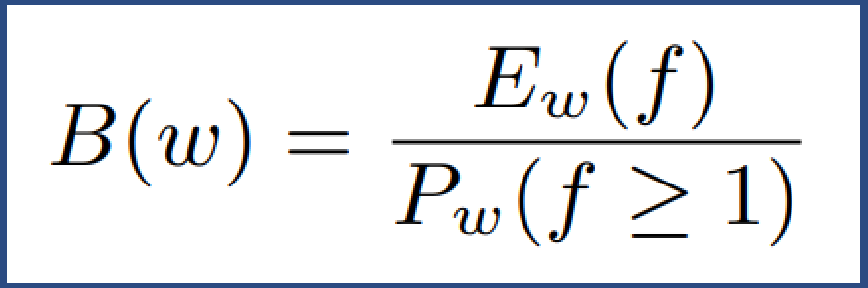
$$\mathrm{PMI}(c, w) = \log \frac{p(c, w)}{p(c)p(w)}$$



Word cloud of PMI and frequency of words for both genders

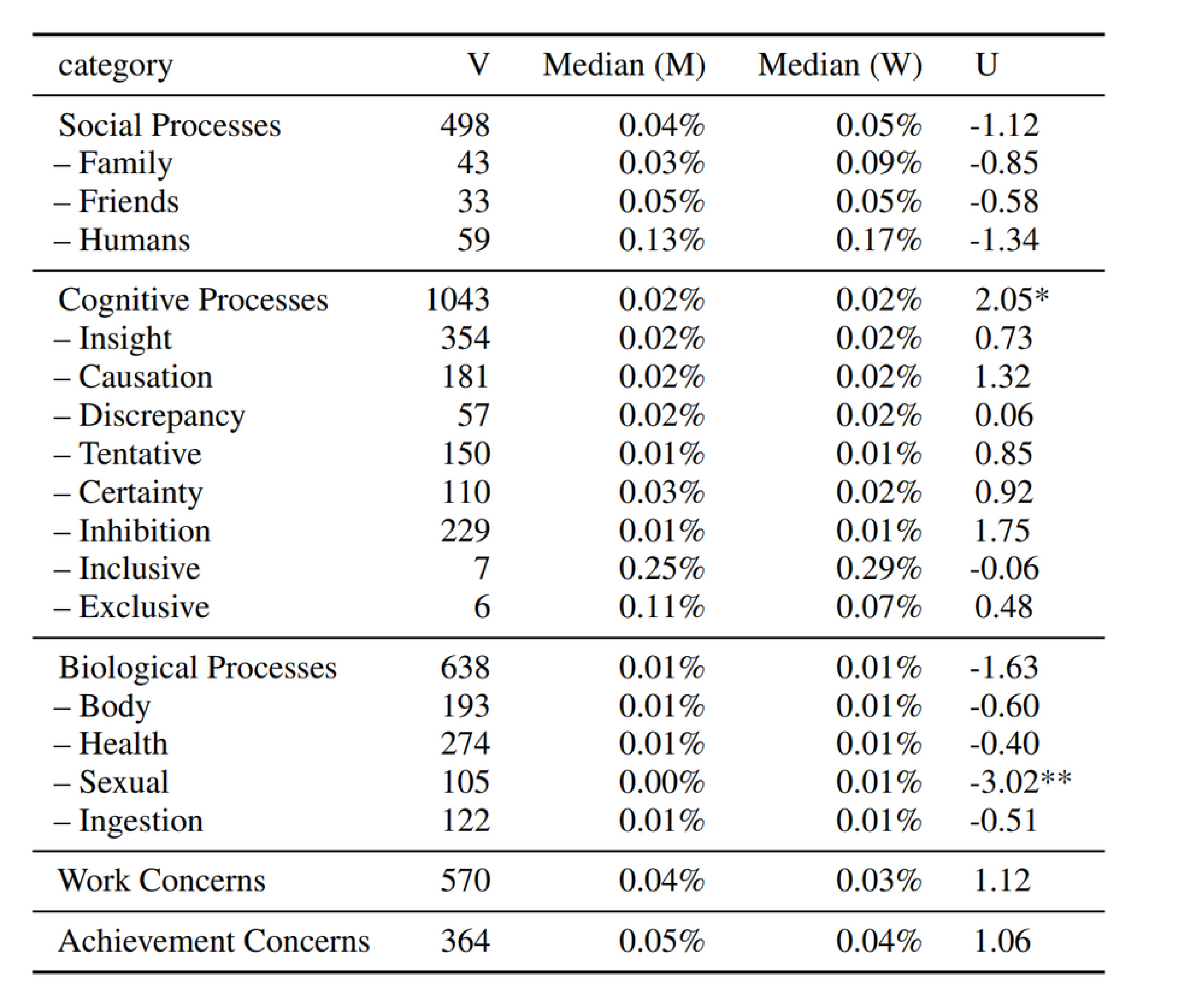
Gender Differences in Semantic Categories of Words

- Two metrics i.e. frequency in overview and burstiness in full text



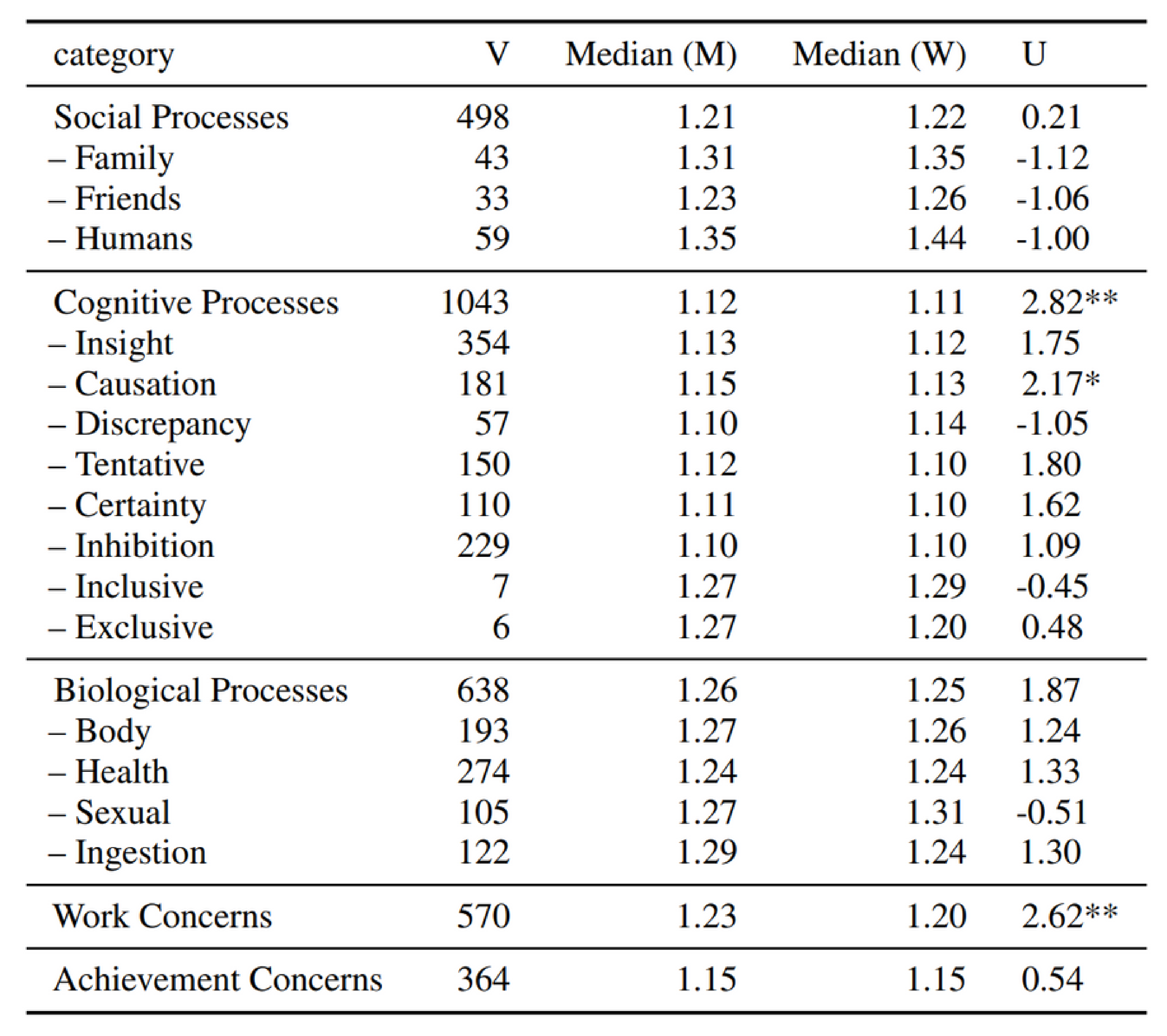
$$B(w) = \frac{E_w(f)}{P_w(f \geq 1)}$$

- Difference measured using Mann-Whitney Test
- Positive value indicates bias towards men and negative value indicates bias towards women
- If significant, calculate common language effect size(ES) as percent of words that had greater relative frequency for dominant gender



category	V	Median (M)	Median (W)	U
Social Processes	498	0.04%	0.05%	-1.12
– Family	43	0.03%	0.09%	-0.85
– Friends	33	0.05%	0.05%	-0.58
– Humans	59	0.13%	0.17%	-1.34
Cognitive Processes	1043	0.02%	0.02%	2.05*
– Insight	354	0.02%	0.02%	0.73
– Causation	181	0.02%	0.02%	1.32
– Discrepancy	57	0.02%	0.02%	0.06
– Tentative	150	0.01%	0.01%	0.85
– Certainty	110	0.03%	0.02%	0.92
– Inhibition	229	0.01%	0.01%	1.75
– Inclusive	7	0.25%	0.29%	-0.06
– Exclusive	6	0.11%	0.07%	0.48
Biological Processes	638	0.01%	0.01%	-1.63
– Body	193	0.01%	0.01%	-0.60
– Health	274	0.01%	0.01%	-0.40
– Sexual	105	0.00%	0.01%	-3.02**
– Ingestion	122	0.01%	0.01%	-0.51
Work Concerns	570	0.04%	0.03%	1.12
Achievement Concerns	364	0.05%	0.04%	1.06

Test applied to word frequency in overviews



category	V	Median (M)	Median (W)	U
Social Processes	498	1.21	1.22	0.21
– Family	43	1.31	1.35	-1.12
– Friends	33	1.23	1.26	-1.06
– Humans	59	1.35	1.44	-1.00
Cognitive Processes	1043	1.12	1.11	2.82**
– Insight	354	1.13	1.12	1.75
– Causation	181	1.15	1.13	2.17*
– Discrepancy	57	1.10	1.14	-1.05
– Tentative	150	1.12	1.10	1.80
– Certainty	110	1.11	1.10	1.62
– Inhibition	229	1.10	1.10	1.09
– Inclusive	7	1.27	1.29	-0.45
– Exclusive	6	1.27	1.20	0.48
Biological Processes	638	1.26	1.25	1.87
– Body	193	1.27	1.26	1.24
– Health	274	1.24	1.24	1.33
– Sexual	105	1.27	1.31	-0.51
– Ingestion	122	1.29	1.24	1.30
Work Concerns	570	1.23	1.20	2.62**
Achievement Concerns	364	1.15	1.15	0.54

Test applied to burstiness of full biography

Discussions

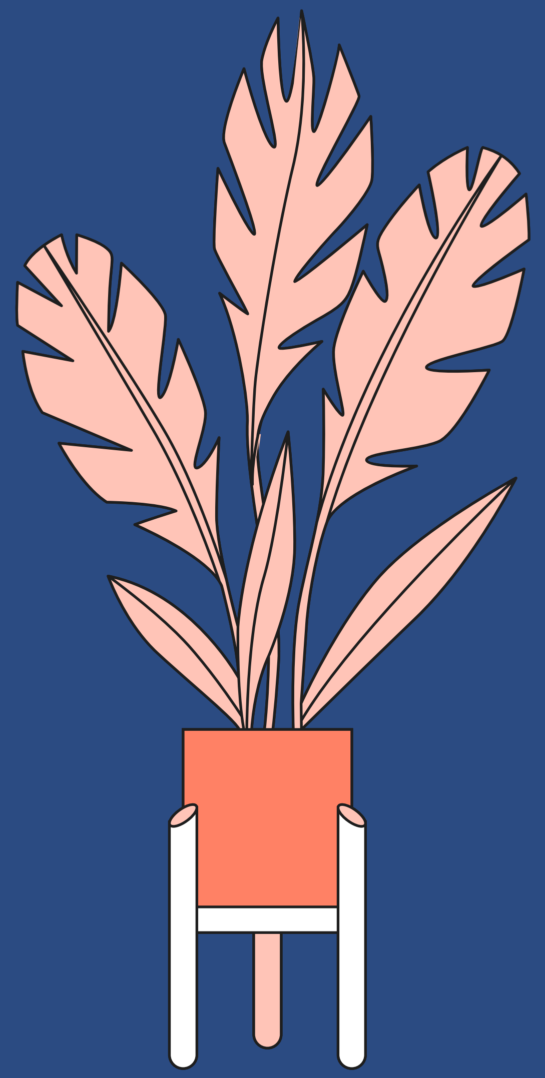
It was found that the words most associated with men are mostly about sports, while the words most associated with women are about arts, gender and family.

Two concepts of particular interest namely: her husband and first woman.

Possible indicator of objectification is the denial of subjectivity i.e. treating the "object" whose feelings and experience need not to be taken into account.

This is supported in the U test values indicating men are more frequently described to their Cognitive Processes while women are frequently described with words related to their sexuality.

Study of specific sub-classes should be approached in future work like analysing the Athlete class through a sociology of sports perspective.



Network Properties

HOW TO COMPARE THE NETWORK PROPERITES?

- Directed graph of biographies from links between articles in PERSON DBpedia class
- Compare with several Null graph.
- Comparison among the structural differences between genders either to empirical fluctuations or gender bias.

Null Models

DIRECTED NETWORK HAS LINKS B/W 893,380 BIOGRAPHICAL ARTICLES FROM PERSON CLASS.
REMOVING SINGLETON NODES FINAL GRAPH HAS 700,706 NODES AND 4153978 EDGES

All graphs have same no of node n=700706 and same mean degree k=4

Random

Shuffle the edges of network.
A random graph with no heterogenous degree distribution and no clustered structure.

In-Degree Sequence

Graph that preserves in-degree distribution or popularity as corresponding biography.

Out-Degree Sequence

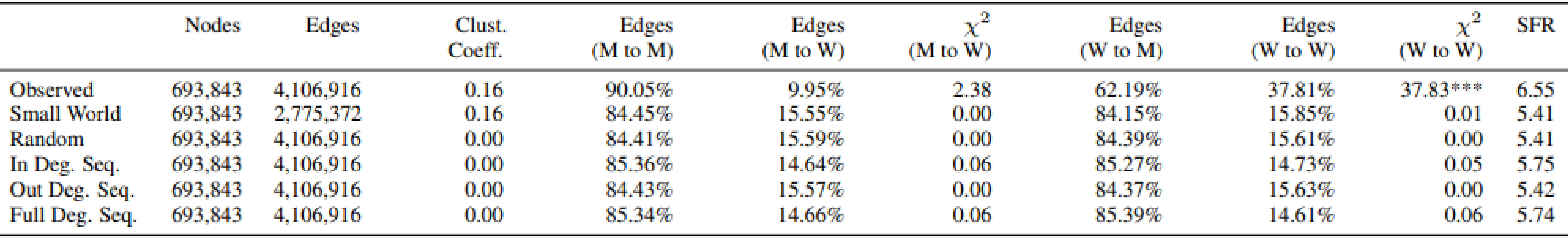
Graph that preserves out-degree distribution.

Full-Degree Sequence

Graph that preserves both in and out-degree distribution . Degree correlation and clustering are lost.

Small World

Undirected small world graph that preserves avg path length and clustering. coeff

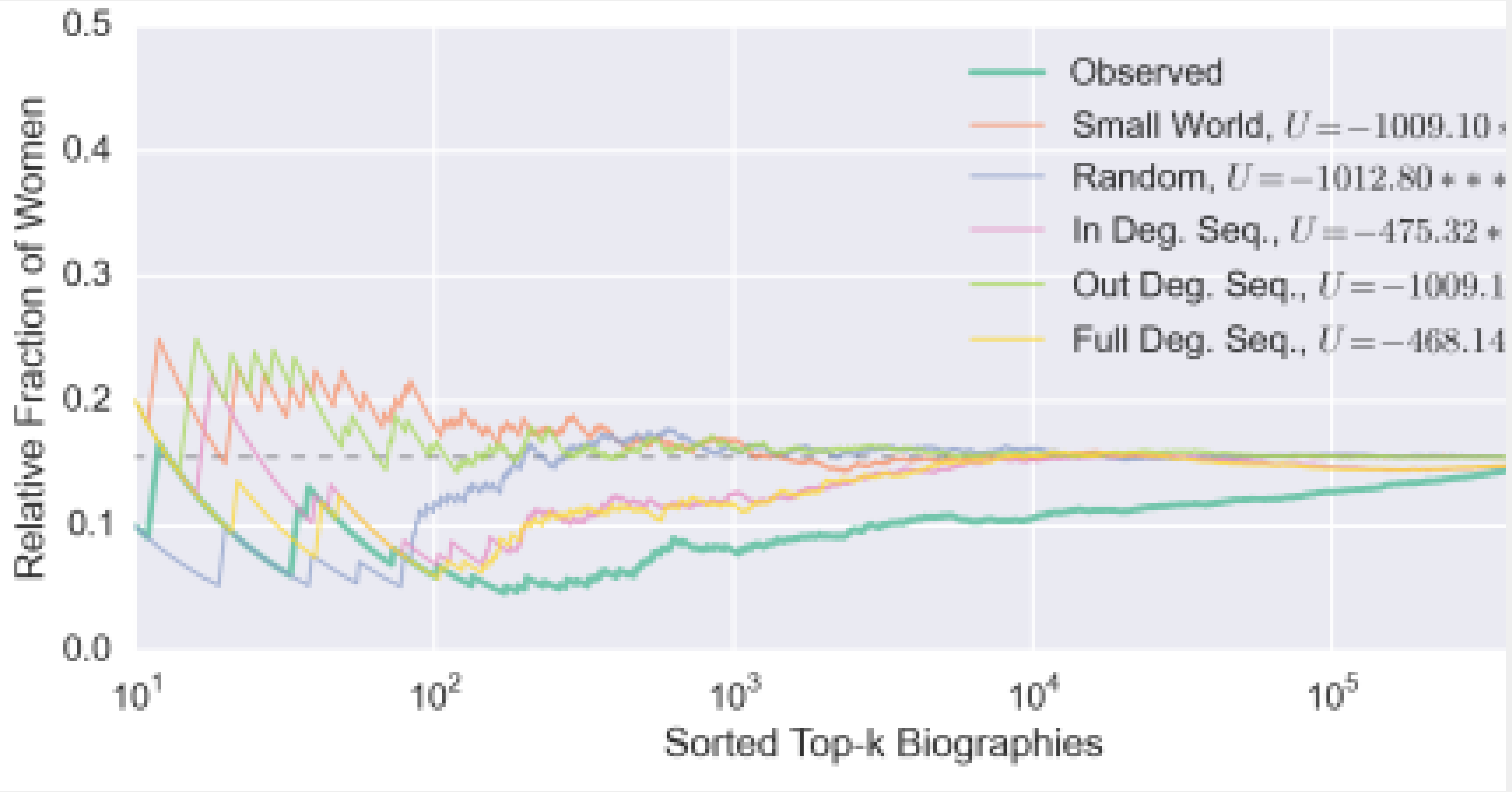


	Nodes	Edges	Clust. Coeff.	Edges (M to M)	Edges (M to W)	χ^2 (M to W)	Edges (W to M)	Edges (W to W)	χ^2 (W to W)	SFR
Observed	693,843	4,106,916	0.16	90.05%	9.95%	2.38	62.19%	37.81%	37.83***	6.55
Small World	693,843	2,775,372	0.16	84.45%	15.55%	0.00	84.15%	15.85%	0.01	5.41
Random	693,843	4,106,916	0.00	84.41%	15.59%	0.00	84.39%	15.61%	0.00	5.41
In Deg. Seq.	693,843	4,106,916	0.00	85.36%	14.64%	0.06	85.27%	14.73%	0.05	5.75
Out Deg. Seq.	693,843	4,106,916	0.00	84.43%	15.57%	0.00	84.37%	15.63%	0.00	5.42
Full Deg. Seq.	693,843	4,106,916	0.00	85.34%	14.66%	0.06	85.39%	14.61%	0.06	5.74

Gender, Link Proportion and Self-Focus Ratio:

- Compared proportion of .links from gender to gender by chi-Squared test among estimated and expected connections.
- No Biasness in null model but significant difference in the proportion of links from women biographies.
- Men biographies have greater proportion of links to men and a lesser proportion to women than expected, but the difference is not statistically significant, although it has an impact on the estimated Self-Focus Ratio(SFR)
- A SFR above 1 confirms the presence of self-focus, which, given the proportions of men and women in the dataset, is expected
- The null models have similar SFR(5.74) to the expected value(5.41), in contrast with the observed model with SFR(6.55)

Biography Centrality



- Ranking of biographies baesd on pagerank values
- analysed the fraction of women in top-r articles.
- 15% is the expected value because of the proportion of women biographies.
- Null models stabalises around 10^4 but observed network stabalises for complete dataset.
- In the presence of correlations between popularity or historical importance and gender, we expect the ratio to fluctuate. But such fluctuations would also be observed in the null models.

Conclusion of Presence and centrality of Women

- Women biographies tend to link more women than men might be related with women editing women's biographies.
- Men and Women's life evolve differently through their career.
- Link proportion can't be attributed to biasness in Wikipedia rather what happens in Physical world.
- Network str is biased in a way to give more importance to men.
- The articles with high centrality or historical importance tend to be about men.
- There are high centrality women biographies but presence is not a sign of unbiased network

Major findings

1. Differences in meta-data are coherent with results in previous work, where women biographies were found to contain more content related to marriage than men's.
2. Sex-related content is more frequent in women biographies than men's, while cognition-related content is more high lighted in men biographies than women's.
3. A strong bias in the linking patterns results in a network structure in which articles about men are disproportionately more central than articles about women.

Conclusions

- Limitations
  - Focus on English Wikipedia which is biased towards Western Cultures
  - Binary gendered view
- Future work
  - Construct editing tools for Wikipedia to help editors detect bias in content and suggest appropriate actions
  - Study individual differences of contributors as analysis on the content was done without considering who published it
  - Further exploration of bias considering more fine-grained ontology classes and attributes for eg differences in biographies based on region and religions etc
  - Whether bias depends on quality of article i.e. number of edits, references etc

Thank You!