First Women, Second Sex: Gender Bias in Wikipedia

Eduardo Graells-Garrido*[1,2] Mounia Lalmas[3] Filippo Menczer[4,5]

[1]Web Research Group
Universitat Pompeu Fabra
Barcelona, Spain

[2]Telefónica I+D
Santiago, Chile

[3]Yahoo Labs
London, UK

[4]Yahoo Labs
Sunnyvale, USA

[5]Indiana University
Bloomington, USA

ABSTRACT

Contributing to the writing of history has never been as easy as it is today. Anyone with access to the Web is able to play a part on Wikipedia, an open and free encyclopedia, and arguably one of the primary sources of knowledge on the Web. In this paper, we study *gender bias* in Wikipedia in terms of how women and men are characterized in their biographies. To do so, we analyze biographical content in three aspects: meta-data, language, and network structure. Our results show that, indeed, there are differences in characterization and structure. Some of these differences are reflected from the off-line world documented by Wikipedia, but other differences can be attributed to gender bias in Wikipedia content. We contextualize these differences in social theory and discuss their implications for Wikipedia policy.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information networks*

Keywords

Wikipedia; Gender; Gender Bias; Computational Linguistics.

1. INTRODUCTION

Today's Web creates opportunities for global and democratic media, where everyone has a voice. One of the most visible examples is Wikipedia, an open encyclopedia where anyone can contribute content. In contrast to traditional encyclopedias, where a staff of experts in specific areas takes care of writing, editing and validating content, in Wikipedia these tasks are performed by a community of volunteers. Whether or not this *open source* approach provides reliable and accurate content [23, 47], Wikipedia has gained unprecedented reach. An extensive body of research covers Wikipedia [39], with topics like participation, structured data, and analysis of historical figures, among others.

In her book *The Second Sex*, Simone de Beauvoir widely discusses different aspects of women oppression and their historical significance. She wrote in 1949 (in French, originally): *"it is not women's inferiority that has determined their historical insignificance: it is their historical insignificance that has doomed them to inferiority"* [16]. More than 60 years later, almost anyone with access to the Web can contribute to the writing of history, thanks to Wikipedia. In theory, by following its guidelines about verifiability, notability, and neutral point of view, Wikipedia should be an unbiased source of knowledge. In practice, the community of Wikipedians is not diverse in terms of gender, as women represent only 16% of editors [25]. This disparity has been called the *gender gap* in Wikipedia, and has been studied from several perspectives to understand why more women do not join Wikipedia, and what can be done about it. It is a problem because reportedly women are not being treated as equals to men in the community [30], and potentially, in content. For instance, Filipacchi [18] described a controversy where women novelists started to be excluded from the category *"American Novelists"* to be included in the specific category *"American Women Novelists."*

Instead of focusing on the participatory *gender gap*, we focus on how women are characterized in Wikipedia articles, to assess whether gender bias from the off-line world extends to Wikipedia content. The scale of Wikipedia, as well as its openness, allows us to perform a quantitative analysis of how women are characterized in Wikipedia in comparison to men. Encyclopedias characterize men and women in many ways, *e. g.*, in terms of their lives and the events in which they participated or were relevant. We concentrate on *biographies* because they are a good source to study gender bias, given that each article is about a specific person. In this paper, we understand gender bias as the *systematic asymmetry* [8] in the way that three dimensions of analysis favor one gender over the other: *meta-data*, *language*, and *network structure*. Then, the research questions that drive our work are:

> *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?*

We present the following three 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 highlighted 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.

These findings represent a quantification of gender bias in user generated content in Wikipedia. The main contributions of this

*Corresponding author: eduardo.graells@telefonica.com.

HT'15, September 1–4, 2015, Guzelyurt, TRNC, Cyprus.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3395-5/15/09 ...$15.00.
DOI: http://dx.doi.org/10.1145/2700171.2791036.

paper are the aforementioned quantification of gender bias, a first contextualization of differences found in terms of social theory, and a discussion of the implications of our findings for policy design in Wikipedia. Furthermore, even though we focus on the English Wikipedia, our methods are generalizable to other languages.

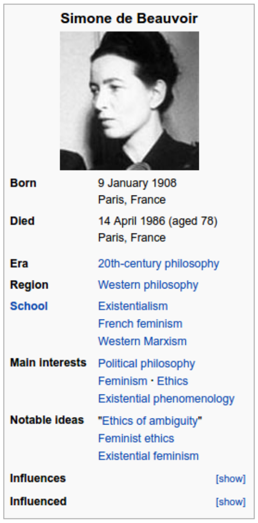
2. BACKGROUND

Research on the community structure and evolution of Wikipedia has been prominent. In its first steps, the focus was on growth [2] and dynamics [45], without attention toward gender. Later, it was found that there is a gender gap, as Wikipedia has fewer contributions from women, and women stop contributing earlier than men [30]. Moreover, men and women communicate differently through the inner communication channels of Wikipedia [31]: they focus on different topics [30] and the level of content revision differs by gender but also by amount of activity [4]. In addition, Lam et al. [30] found that women are more *reverted* than men (*i.e.*, their contributions are discarded), and reportedly women contribute less because of aggressive behavior toward them [15, 52]. Efforts have been made to build a more welcoming community and to encourage participation [36, 14] with initiatives like *WikiWomen's Collaborative*.[1]

Content-wise, the study of biographies in Wikipedia enables the identification of cultural differences in content [42] and coverage [10], as well as the construction of social networks of historical (and current) figures [6]. Lam et al. [30] found that, in terms of interest of contributors, coverage of "female topics" (*i. e.*, topics of interest for women) was inferior to "male topics " when classifying topics as "male" or "female" according to the people who contributed to them. Reagle and Rhue [46] found that in characterization of women, in comparison to commercial encyclopedias like *Britannica*, Wikipedia has better coverage of notable profiles, although this coverage is quite low and is still biased toward men. Bamman and Smith [7] found that women biographies are more likely to include language related to marriage or divorce events. At large-scale, Hecht and Gergle [24] measured *self-focus bias* [24], which studies the relation of the language of a specific Wikipedia and its related cultures.

Addressing the gender gap from a content perspective may help to improve the quality and value of the content. Researchers have focused on predicting article quality in Wikipedia [3, 20]. However, focusing on quality without considering readers does not give the whole picture, as Wikipedia readers are not necessarily interested in the same topics as contributors [33] and might have a different concept of quality. Moreover, in our context, Flekova, Ferschke, and Gurevych [20] found that the quality of biographies is assessed differently depending on the gender of the portrayed person. Is it because the raters were biased? Or is it because biographies were written differently? Our hypothesis is that biographies are written differently, an idea inspired by seminal work by Lakoff [29] about how women are characterized by language.

Word frequency is commonly used to study differences in text. Word frequency follows Zipf's law [57, 49], an empirical distribution found in many languages [43]. An interesting property of Zipf distributions in language is that small sets of words that are semantically or categorically related also follow a Zipf distribution [43]. This property implies that, given two subsets of words that are related semantically or categorically, their frequency distributions can be compared. Thus, we compare frequency distributions according to gender for several semantic categories derived from the *Linguistic Inquiry and Word Count* (LIWC) dictionary. LIWC studies *"emotional, cognitive, structural, and process components present in individuals' verbal and written speech samples"* [40]. It has been used to analyze interactions between Wikipedia contributors [26] and article content with respect to emotions [17]. In a context similar to ours, Schmader, Whitehead, and Wysocki [48] used LIWC to quantify differences in characterization of women and men in recommendation letters.



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]

Figure 1: *Infobox* from the biography article of Simone de Beauvoir.

In our work, we quantify gender bias in Wikipedia's characterization of men and women through their biographies. To do so we approach three different dimensions of biographies, which we analyze in different sections on this paper: *meta-data*, provided by the structured version of Wikipedia, DBpedia [34]; *language*, considering how frequent are words and concepts [49]; and *network structure*. In terms of network structure, we build a biography network [6] in which we estimate PageRank, a measure of node centrality based on network connectivity [9, 21]. In similar contexts, PageRank has been used to provide an approximation of historical importance [6, 50] and to study the bias leading to the gender gap [50]. We measure bias in link formation by comparing the importance given by PageRank in the biography network with those of null models, *i. e.*, graphs that are unbiased by construction but that maintain certain properties of the source biography network.

3. DATA SOURCES

To study gender bias in Wikipedia, we consider three freely available data sources:

1. The DBpedia 2014 dataset [34].[2]
2. The Wikipedia English Dump of October 2014.[3]
3. Inferred gender for Wikipedia biographies by Bamman and Smith [7].[4]

In addition, we use the *Linguistic Inquiry and Word Count* dictionary of semantic categories to find if different genders have different characterizations according to those semantic categories.

DBpedia. DBpedia [34] is a structured version of Wikipedia that provides meta-data for articles, normalized article URIs (*Uniform Resource Identifiers*), normalized links between articles (taking care of redirections), and a categorization into a shallow ontology, which includes a *Person* category. To provide the structured meta-data,

[1] http://meta.wikimedia.org/wiki/WikiWomen's_Collaborative

[2] http://wiki.dbpedia.org/Downloads2014

[3] https://dumps.wikimedia.org/enwiki/20141008/

[4] http://www.ark.cs.cmu.edu/bio/

Table 1: Considered semantic categories from LIWC, and their two most frequent words found in biographies from each gender.

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

DBpedia processes content from the infoboxes in Wikipedia articles. Infoboxes are template-based specifications for certain kinds of articles. When DBpedia detects an infobox with a template that matches those of a person, it assigns the article to the *Person* class from the ontology, and to a specific subclass if applicable (*e. g.*, *Artist*). Thus, we consider an article to be a biography if it belongs to the *Person* class. For instance, Figure 1 displays the infobox of Simone de Beauvoir. The infobox contains specific meta-data attributes pertinent to a biography, such as date and place of birth, but it does not include gender (except in certain cases, see "Inferred Gender" next). DBpedia maps infobox properties to specific fields in a person's meta-data.

Wikipedia Biography Text. We consider two versions of the biographies: the overview and the full text. We analyze both in different contexts: in the overview we analyze the full vocabulary employed, while in the full text we analyze only the words pertaining to the LIWC dictionaries. The overview is described by Wikipedia as *"an introduction to the article and a summary of its most important aspects. It should be able to stand alone as a concise overview."* Since those aspects are subjective, the overview content is a good proxy for any potential biases expressed by Wikipedia contributors. At the same time we avoid potential noise included in the full biography text from elements like quotations and the filmography of a given actor/actress. In both cases (overview and full content), template markup is removed from analysis.

Inferred Gender. To obtain gender meta-data for biographies, we match article URIs with the dataset by Bamman and Smith [7], which contains inferred gender for biographies based on the number of grammatically gendered words (*i. e.*, *he*, *she*, *him*, *her*, etc.) present in the article text. Bamman and Smith [7] tested their method in a random set of 500 biographies, providing 100% precision and 97.6% recall. This method has also been used before by Reagle and Rhue [46] and DBpedia itself [34], making DBpedia to include gender meta-data in some cases. Note that the genders considered in these datasets (and thus, in this work) are only *male* and *female*.

Semantic Categories. The LIWC dictionary [40] includes, for each category (and its corresponding subcategories), a list of words and prefixes that match relevant words. We consider the categories (and their subcategories, if applicable): *Social Processes*, *Cognitive Processes*, *Biological Processes*, *Work Concerns*, and *Achievement Concerns*. We believe these categories should be used in the same way when characterizing women and men. Other categories have been left out of analysis, such as *Positivity*, *Negativity*, *Relativity*, *Religion*, and *Death*, as they can be used in different ways in biographies (*e. g.*, it is expected that religious men have completely different characterizations from religious women). To generate the final dictionaries for analysis from the vocabulary, we matched the vocabulary found in biographies with the prefixes in our corpus. Then we performed manual cleaning of noisy keywords, such as places (*e. g.*, *Virginia* matches *virgin** from the *Sexual* category), names (*Victoria* matches *victor** from the *Achievement Concerns* category), and words with unrelated meanings. In total, our cleaned dictionary contained 2,877 words. Table 1 shows the two most frequent words found per gender for each considered category.

Table 2: 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. In this and the following tables, we use this legend for p-values: ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$.

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**

4. META-DATA PROPERTIES

In our first analysis we estimate the proportion of women in Wikipedia. We analyze meta-data by comparing how men and women proportionally have several attributes in the data from DBpedia.

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 meta-data. Of those, only 15.5% are about women.

The mean article length is 5,955 characters for men and 6,013 characters for women (a significant difference according to a t-test for independent samples: $p < 0.01$, Cohen's $d = 0.01$). The mean out-degrees (number of links) of 42.1 for men and 39.4 for women also differ significantly ($p < 0.001$, Cohen's $d = 0.06$). Table 2 displays the number of biographies in the *Person* class, as well as its most common subclasses. From the table, in comparison to the global proportion of women, the following categories over-represent women: *Artist*, *Royalty*, *FictionalCharacter*, *Noble*, *BeautyQueen*, and *Model*. The others over-represent men. The differences in length and degree do not hold for all classes, hinting that a study according to semantic categories of people is needed. However, in this paper we focus on the global differences in *Person*.

Date of Birth. Figure 2 displays the distribution of biographies according to their corresponding *birthYear* property, considering

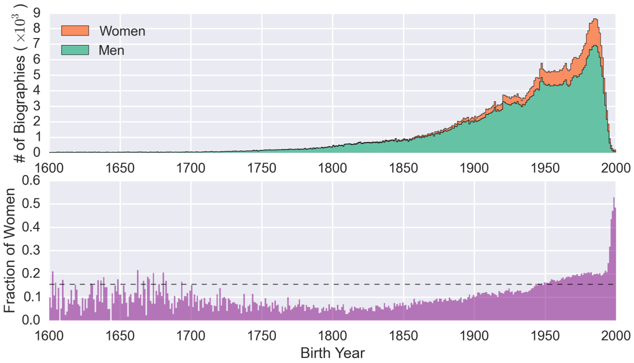


Figure 2: Distribution of biographies according to birth year.

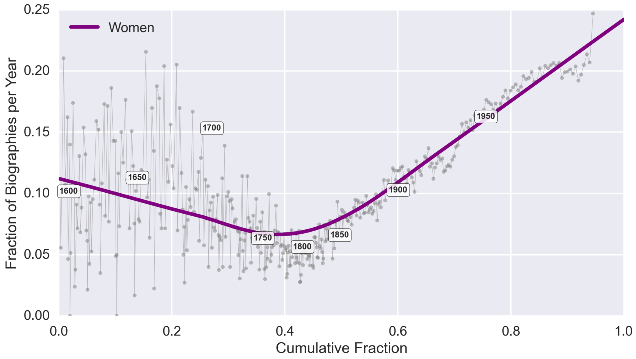


Figure 3: Relation between the cumulative fraction of women in time and the fraction of women per year (dots). The y-axis was truncated to 0.25 for clarity.

only those biographies between years 1600 and 2000 (inclusive). This accounts for 65.48% of biographies with gender (note that 34.07% do not have date of birth in meta-data). The distribution per gender (top chart) shows that most of the biographies of both genders are about people from modern times. The distribution of the fraction of women per year (bottom chart) shows that since the year 1943 the fraction of women is consistently above the global fraction of 0.155. Note that, of the biographies that have date of birth in their meta-data, 53% are from 1943 until 2000. To explore the evolution of growth of women presence, in Figure 3 we display the relationship between the cumulative fraction of biographies in time and the yearly fraction of biographies of women. The chart includes a *LOWESS* (LOcally Weighted Scatterplot Smoothing) fit of the data, to be able to see the tendency of changes in representation. This tendency became positive in the period 1750–1800. These results are discussed in terms of historical significance in Section 7.

Infobox Attributes. Given that there are different classes of infoboxes, there are many different meta-data attributes that can be included in biographies. In total, we identified 340 attributes. 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. Those attributes are displayed in Table 3. All of them have large effect sizes (Cohen's $w > 0.5$). Inspection allows us to make several observations:

Table 3: Proportion of men and women who have the specified attributes in their infoboxes. Proportions were tested with a chi-square test, with effect size estimated using Cohen's w.

	% 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

- Attributes *careerStation*, *formerTeam*, *numberOfMatches*, *position*, *team*, and *years* are more frequent in men. All these attributes are related to sports, and thus, these differences can be explained by the prominence of men in sports-related classes (*e. g.*, *Athlete*, *SportsManager* and *Coach* in Table 2).
- Attributes *deathDate*, *deathYear* are more frequent in men. According to Figure 2, most women are from recent times, and thus they are presumably still alive.
- 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 (see Table 2). In addition, even though it depends on the cultural context, another possible explanation is that married women usually change their surnames to those of their husbands.
- Attributes *occupation* and *title* are more frequent in women, and seem to serve the same purpose but through different mechanisms. On one hand, *title* is a text description of a person's occupation (the most common values found are *Actor* and *Actress*). On the other hand, *occupation* is a DBpedia resource URI (*e.g.*, http://dbpedia.org/resource/Actress). These attributes are present in the infoboxes of art-related biographies. On the contrary, the infoboxes of sport-related biographies do not contain these attributes because their templates are different and contain other attributes (such as the aforementioned *careerStation* and *formerTeam*) which already indicate their occupations. Thus, athletes (which are mostly men) do not contain such attributes.
- 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 (*i. e.*, when the spouse does not have a Wikipedia article), or the resource URI of the article of *"divorced status."* Our manual inspection did not offer a direct explanation for the tendency to include this attribute in women's biographies more than in men's. For instance, the most common class with the spouse attribute is *Person* (without a more specific subclass), with 45% of the instances of the attribute.

5. LANGUAGE PROPERTIES

In this section we explore the characterization of women and men from a lexical perspective. First, we analyzed the vocabulary used in the overview of each biography through word frequency, and we use the estimated frequencies to find which words are associated with each gender. To estimate relative frequencies, words were considered once per biography, and we estimated bi-gram word collocations to identify composite concepts (*e. g.*, *New York*). We obtained a vocabulary of size V_m = 1,013,305 for men, V_w = 376,737 for women, with V = 272,006 common words.

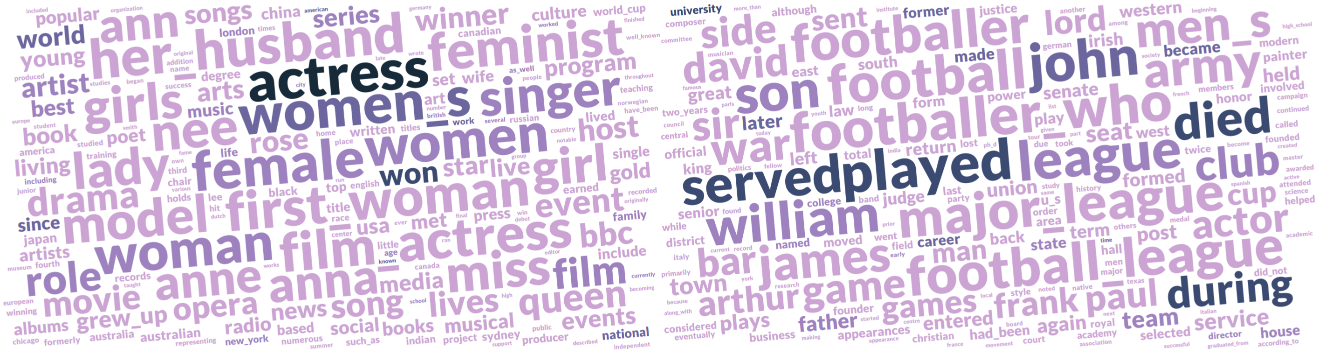


Figure 5: Words most associated with women (left) and men (right), estimated with *Pointwise Mutual Information*. Font size is inversely proportional to PMI rank. Color encodes frequency (the darker, the more frequent).

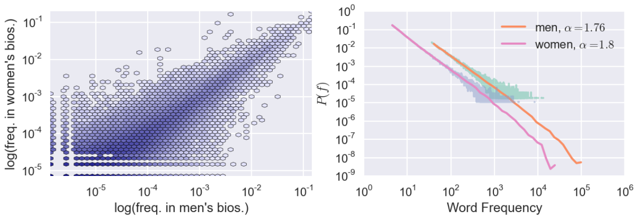


Figure 4: A density hexbin plot of word frequencies in men/women's biographies (left), and the PDF of word frequency distribution according to gender (right).

Figure 4 displays a density plot of word frequency, and the Probability Density Functions (PDFs) for both genders. The frequency distributions are similar across genders. Word frequencies in the common vocabulary for both genders follow a Zipf distribution $P(f) \sim f^{-\alpha}$ with similar exponents $\alpha \approx 1.8$, consistent with the value found by Serrano, Flammini, and Menczer [49]. In addition, frequency with respect to gender presents a high rank-correlation $\rho = 0.65$ (p $<$ 0.001). For reference, consider that the inter-language rank correlation of words with the same meaning across languages is 0.54 [11]. This implies that words share meanings when referring to men and women.

Associativity of Words with Gender. To explore which words are more strongly associated with each gender, we measure *Pointwise Mutual Information* [13] over the common vocabulary V for both genders. PMI is defined as:

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

where c is a class (*men* or *women*), and w is a word. The value of $p(c)$ can be estimated from the proportions of biographies about men and women, and the other probabilities can be estimated from the corresponding proportions of words and bi-grams. Since PMI overweights words with very small frequencies, we consider only words that appear in at least 1% of men or women biographies.

Associativity results are displayed as word clouds in Figure 5. The top-15 words associated to each gender are (relative frequency in parentheses):

- Women: *actress* (15.9%), *women's* (8.8%), *female* (5.6%), *her husband* (4.1%), *women* (5.3%), *first woman* (1.9%), *film actress* (1.6%), *her mother* (1.8%), *woman* (4.4%), *nee* (3.6%), *feminist* (1%), *miss* (1.9%), *model* (3.3%), *girls* (1.5%) and *singer* (6.5%).
- Men: *played* (14.2%), *footballer who* (3%), *football* (4.5%), *league* (5.9%), *john* (7.9%), *major league* (1.8%), *football league* (1.6%), *college football* (1.5%), *son* (7%), *football player* (2.2%), *footballer* (2%), *served* (11.7%), *william* (4.6%), *national football* (2%) and *professional footballer* (1%).

There is an evident difference in the kind of words most associated to each gender. The words most associated with men are related to sports, football in particular, which refers to both popular sports of soccer and American football (recall from Table 2 that *Athlete* is the largest subclass of *Person*). For women, the most associated words are related to arts (recall from Table 2 that *Artist* is the second largest subclass of *Person*), gender (*women's*, *female*, *first woman*, *feminist*), and family roles (*her husband*, *her mother*, and *nee* (adjective used when giving a former name of a woman in some cultures). This is consistent with the results from the meta-data analysis, where women are more likely to have a *spouse* attribute in their infoboxes (see Table 3), and with the results of Bamman and Smith [7].

Gender Differences in Semantic Categories of Words. To compare the distributions of words in the semantic categories, we employed two metrics: relative frequency in overviews, as previously done with PMI, and burstiness in the full text. Word frequencies identify how language is used differently to characterize men and women in terms of semantic categories. However, word frequency alone does not give insights on how those semantic categories portray a given biography, or in other words, the importance that editors give to those categories. Burstiness is a measure of word importance in a single document according to the number of times it appears within the document, under the assumption that important words appear more than once (they appear in *bursts*) when they are relevant in a given document. We use the definition of burstiness from Church and Gale [12]:

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

where $E_w(f)$ is the mean number of occurrences of a given word w per document, and $P_w(f \geq 1)$ is the probability that w appears at least once in a document. The differences in frequency and burstiness are tested using the Mann-Whitney U test, which indicates if one population tends to have larger values than another. It is non-parametric, *i. e.*, it does not assume normality. For all categories under consideration, we report the test value. A positive value indicates that the distribution is biased towards men, and a negative value indicates bias towards women. If the test is significant, we calculate the *common language effect size* (ES) as the percentage of words that had a greater relative frequency for the dominant gender.

Table 4: Word frequency in biography overviews. For each LIWC category we report vocabulary size, median frequencies, and the result of a Mann-Whitney *U* test. *M* and *W* mean men and women, respectively.

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

Table 4 shows the result of the test applied to word frequency in biography overviews. Note that, although the medians are very similar for each category, the *U* test compares differences in the distribution instead of differences in means or medians. Of the 20 categories under consideration, two of them (one top-level) showed significant differences between genders: *Cognitive Processes* ($U = 2.04$, ES $= 63\%$) is dominated by men, while *Sexual* (subcategory of *Biological Processes*, $U = -3.02$, ES $= 85\%$) is dominated by women. Burstiness distributions in full biographies per semantic category are displayed in Table 5. There are three (two top-level) categories with significant differences, both dominated by men: *Cognitive Processes* ($U = 2.85$, ES $= 60\%$), its subcategory *Causation* ($U = 2.17$, ES $= 71\%$), and *Work Concerns* ($U = 2.62$, ES $= 64\%$).

In this section, we have analyzed generic semantic categories with words that should be used in the same way to describe women and men. Although the results found imply more similarities than differences, in the discussion section we elaborate over the importance of such differences and the implications of these findings.

6. NETWORK PROPERTIES

To study structural properties of biographies, we first built a directed network of biographies from the links between articles in the *Person* DBpedia class. This empirical network was compared with several null graphs that, by construction, preserve different known properties of the original network. This allows us to attribute observed structural differences between genders either to empirical fluctuations in such properties, such as the heterogeneous importance of historical figures, or to gender bias. To do so, we consider PageRank, a measure of node centrality based on network connectivity [9, 21].

Empirical Network and Null Models. We study the properties of the directed network constructed from the links between 893,380 biographical articles in the *Person* class. After removing 192,674 singleton nodes, the resulting graph has 700,706 nodes and 4,153,978 edges. We use this graph to construct the following null models:

Table 5: Word burstiness in full biographies for LIWC categories. Columns are analog to Table 4.

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

- *Random.* We shuffle the edges in the original network. For each edge (u,v), we select two random nodes (i,j) and replace (u,v) by (i,j). The resulting network is a random graph with neither the heterogeneous degree distribution nor the clustered structure that the Wikipedia graph is known to have [58].
- *In-Degree Sequence.* We generate a graph that preserves the in-degree sequence (and therefore the heterogeneous in-degree distribution) of the original network by shuffling the sources of the edges. For each edge (u,v), we select a random node (i) and rewire (u,v) to (i,v). Each node has the same in-degree, or popularity, as the corresponding biography.
- *Out-Degree Sequence.* We generate a graph that preserves the out-degree sequence (and therefore the out-degree distribution) of the original network by shuffling the targets of the edges. For each edge (u,v) select a random node (j) and rewire (u,v) to (u,j).
- *Full Degree Sequence.* We generate a graph that preserves both in-degree and out-degree sequences (and therefore both distributions) by shuffling the structure in the original network. For a random pair of edges ((u,v), (i,j)) rewire to ((u,j), (i,v)). We repeat this shuffling as many times as there are edges. Note that although the in- and out-degree of each node is unchanged, the degree correlations and the clustering are lost.
- *Small World.* We generate a undirected small world graph using the model by Watts and Strogatz [55]. This model interpolates a random graph and a lattice in a way that preserves two properties of small world networks: average path length and clustering coefficient.

All null models have the same number of nodes $n = 700{,}706$ and approximately the same mean degree $k \approx 4$ as the empirical network.

Gender, Link Proportions and Self-Focus Ratio. For each graph, we estimated the proportion of links from gender to gender, and we tested those proportions against the expected proportions of men and women present in the dataset using a chi-square test. Table 6 shows the results. None of the null models show any bias in link proportions. The observed graph, on the other hand, shows a significant difference in the proportion of links from women biographies. In particular, articles about women tend to link to other women

Table 6: Comparison of the empirical biography network and the null models. *M* and *W* mean men and women, respectively.

	Nodes	Edges	Clust.	Edges	Edges	χ^2	Edges	Edges	χ^2	SFR
			Coeff.	(M to M)	(M to W)	(M to W)	(W to M)	(W to W)	(W to W)	
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

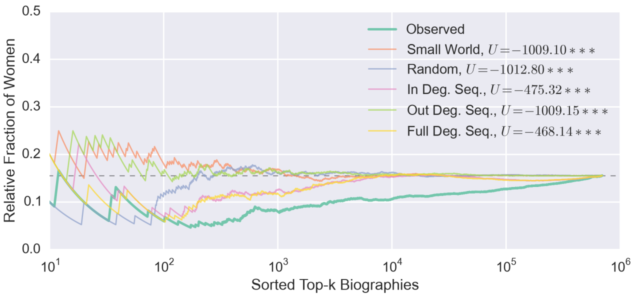


Figure 6: Women fraction in top biographies sorted by PageRank.

biographies more than expected ($\chi^2 = 40.54, p < 0.001$, Cohen's $w = 0.76$). Men biographies show a 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* [24]. In our context, this ratio is defined as the relation between the sum of PageRank for men and the sum of PageRank for women. A SFR above 1 confirms the presence of self-focus, which, given the proportions of men and women in the dataset, is expected. In fact, given those proportions, the expected SFR is 5.41. Note that the null models have similar SFRs to the expected value, in contrast with the observed model with SFR of 6.55.

Biography Centrality. As an approximation for importance in our biography network we considered the ranking of biographies based on their PageRank values. To compare the observed distribution of PageRank by gender to those of the null models, we analyzed the fraction of women biographies among the top-r articles by PageRank, having $r \geq 10$. In the absence of any kinds of bias, whether endogenous to Wikipedia or exogenous, one would expect the fraction of women to be around 15% (the overall proportion of women biographies) irrespective of r. 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.

The results are shown in Figure 6. While the null models stabilize around the expected value by $r \leq 10^4$, the proportion of women in the observed network reaches 15% only when the entire dataset is considered. This systematic under-representation of women among central biographies is not mirrored in the null models. We tested the differences between observed and null models using a Mann-Whitney *U* test, and found that the observed model is significantly different with all null models (*U* values shown in Figure 6, p < 0.001 for all pairwise comparisons with the observed model, Holm-Sidak corrected). This implies an asymmetry that cannot be explained by any of the heterogeneities in the structure of the network preserved by the null models. For instance, even if men biographies tended to have more incoming links (as they do), or to be more densely clustered, those factors would not explain the lower centrality observed in women biographies.

7. DISCUSSION

Even though we found more similarities than differences in characterization, in this section we contextualize those differences in social theory and history. We do this to understand why such differences exist, and whether they can be attributed to bias in the English Wikipedia or to a reflection of the society documented in it.

Meta-data. We found that there are statistically significant differences in biographies of men and women. Most of them can be explained because of the different areas to which men and women belong (mostly *sports* and *arts*, respectively), as well as the recency of women profiles available on Wikipedia. Other differences, like article length and article out-degree, although significant, have very small effect sizes, and depend on the person class being analyzed.

The greater frequency of the *spouse* attribute in women can be interpreted as specific gender roles attributed to women. Regarding this *Implicit Association*, Nosek, Banaji, and Greenwald [37] found that Internet visitors tended to associate women and language related to family and arts. Arguably, an alternative explanation is that people in the arts could be more likely to marry a notable spouse than people in sports. Yet, we found that the most common specific class was the generic one, not assigned any of its sub-classes.

In terms of time, we found that the year 1943 marked a hit on the growth of women presence. According to Strauss and Howe [53], the post-war *Baby Boomers* generation started in 1943. The following generations are *Generation X* (1961–1981) and *Millenials* (1982–2004). The social and cultural changes embraced by people from those generations, plus the increased availability of secondary sources, might explain this growth. We also observed that the cumulative growth of women presence started in dates nearby the French Revolution (1789–1799), where women had an important role, although they were oppressed after it [1]. During these years seminal works about feminist philosophy and women's rights were published, like the works of *Mary Wollstonecraft* (1792) and *Olympe de Gouges* (1791). It is reasonable to assume that these historical events paved the way for women to become more notable.

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. For instance, according to Lauter [32], in the 20th century women became *mythmakers* through arts, an hypothesis that is supported by our results.

Language. We 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. Of particular interest are two concepts strongly associated with women: *her husband* and *first woman*. These results are arguably indicative of systemic bias: the usage of *her husband* was found in concordance with our meta-data results and previous work by Bamman and Smith [7], and the aforementioned work on *Implicit Association* [37]. These results can be contextualized in terms of *stereotyping theory* [44], as they categorize women, either as norm breaking (being the first is an exception to the norm) or as with predefined roles (being wives). As Fiske and Neuberg [19] indicate in their *continuum model of impression formation*, such categorization makes individuals more prone to stereotyping than those who are not categorized. The usage of *first woman* might indicate notability, but it also has been seen as an indicator of gender bias, as indicated by the Bechdel-inspired *Finkbeiner-test*[5] about scientific women, where it is explicitly mentioned that an article about a woman does not pass the test if it mentions *"How she's the 'first woman to . . . '"* Despite being informal, the Finkbeiner-test raises awareness on how gender becomes more important than the actual achievements of a person.

According to Nussbaum [38], one possible indicator of *objectification* is the *"denial of subjectivity: the objectifier treats the object as something whose experience and feelings (if any) need not be taken into account."* This idea is supported as, in the overviews, men are more frequently described with words related to their *Cognitive Processes*, while women are more frequently described with words related to *sexuality*. In the full biography text, the *Cognitive Processes* and *Work Concerns* categories are more bursty in men biographies, meaning that those aspects of men's lives are more important than others at the individual level.

It could be thought that, instead of gender bias, such use of language could be a consequence of the imbalance in *Person* sub-classes like *Artist* and *Athlete* (recall Table 2). However, the semantic word categories under study are neutral in that aspect, as each word should be used equally (in terms of meaning) when portraying a person, regardless of her/his gender. Recall that we omitted categories that could be prone to be different for each gender (*e. g.*, *Religion*). We acknowledge that the study of specific sub-classes should be approached in future work, specially from a social science point of view. For instance, the *Athlete* class can be analyzed from a *sociology of sports* framework [35].

Presence and Centrality of Women. Women biographies tend to link more to other women than to men, a disproportion that might be related with women editing women biographies in Wikipedia, one of the reported interests of women editors [52]. Since we are considering notable people, it is known that men and women's networks evolve differently through their careers [27], not to mention the set of life-events that influence those changes like child-bearing and marriage (see an in-depth discussion by Smith-Lovin and McPherson [51]). Thus, link proportion between women cannot be attributed to bias in Wikipedia, as it seems to be more a reflection of what happens in the physical world.

We found that network structure is biased in a way that gives more importance to men than expected, by comparing the distribution of PageRank across genders. The articles with highest centrality, or historical importance [6], tend to be predominantly about men, beyond what one could expect from the structure of the network. As shown in Figure 6, there are women biographies with high centrality, but their presence is not a sign of an unbiased network: *"the successes of some few privileged women neither compensate for nor excuse the systematic degrading of the collective level; and the very fact that these successes are so rare and limited is proof of their unfavorable circumstances"* [16].

7.1 Implications

At this point, considering the *gender gap* that affects Wikipedia [25], it is pertinent to recall the concept of *feminine mystique* by Friedan [22], developed from the analysis of women's magazines from the 50s in the United States, which were edited by men only. Fortunately, as discussed earlier, we have found women in different fields, mostly *arts*, in contrast to the *"Occupation: Housewife"* identified by Friedan [22], as well as more similarities in characterization than differences. Moreover, the presence of women is increasing steadily and most of the differences found are not from an inherent bias in Wikipedia. Nevertheless, the identified language differences objectify women and the network structure diminishes their findability and centrality. Hence, the gender bias in Wikipedia is not just a matter of women participation in the community, because content and characterization of women are also affected. This is important, for example, because Wikipedia is used as an educational tool [28], and *"children learn which behaviors are appropriate to each sex by observing differences in the frequencies with which male and female models as groups perform various responses in given situations"* [41].

Editing Wikipedia and NPOV. Critics may rightly say that by relying on secondary sources, Wikipedia just reflects the biases found in them. However, editors are expected to write in their own words, *"while substantially retaining the meaning of the source material,"*[6] and thus, the differences found in terms of language that objectify women are caused explicitly by them. In this aspect, Wikipedia should provide tools that help editors reduce sexism in language, for instance, by considering already existing manuals like [5]. Furthermore, their neutral point of view guidelines should be updated to explicitly include gender bias, because biased language is a clear violation of their guidelines.

Affirmative Action for Women in Notability Guidelines. The current notability guidelines for biographies in Wikipedia state: *"1. The person has received a well-known and significant award or honor, or has been nominated for one several times. 2. The person has made a widely recognized contribution that is part of the enduring historical record in his or her specific field."*[7] However, the boundary between not being notable according to sources and exclusion from history is blurred when evaluating the notability of women. For instance, consider a discussion about women in philosophy: *"Feminist historians of philosophy have argued that the historical record is incomplete because it omits women philosophers, and it is biased because it devalues any women philosophers it forgot to omit. In addition, feminist philosophers have argued that the philosophical tradition is conceptually flawed because of the way that its fundamental norms like reason and objectivity are gendered male"* [56]. Women, specially in historical contexts before 1943, should be targeted by affirmative actions that would allow them to appear in the content if they are not there, and be linked from other articles. We acknowledge that this is not easy, because relaxing notability guidelines can open the door to original research, which is not allowed. However, a correctly defined affirmative strategy would allow to grow the proportion of women in Wikipedia, make women easier to find, both through search (as it increases relevance) and exploratory browsing.

8. CONCLUSIONS

We studied gender bias in Wikipedia biographies. Our results indicate significant differences in meta-data, language, and network structure that can be attributed not only to the mirroring of the offline world, but also to gender bias endogenous to content generation in

[5] http://www.doublexscience.org/the-finkbeiner-test/

[6] https://en.wikipedia.org/wiki/Wikipedia:No_original_research

[7] https://en.wikipedia.org/wiki/Wikipedia:Notability_(people)#Any_biography

Wikipedia. Our contribution is a quantification of *systematic asymmetries* [8], which we define as gender bias with respect to content and structure, as well as a contextualization of the differences found in terms of social theory. In concluding remarks, we proposed that Wikipedia may wish to consider revising its guidelines, both to account for the non-findability of women and to encourage a less biased use of language, as such a bias is a violation of the neutral point of view guideline.

Limitations. Our study has two main limitations. First, our focus is on the English Wikipedia, which is biased towards western cultures [24]. However, a parallel work to ours by Wagner et al. [54] focused on hyperlingual quantitative analysis, and obtained similar results for other languages. Our methods can be applied in other contexts given the appropriate dictionaries with semantic categories. The second limitation is a binary gendered view, but we believe this is a first step towards analyzing the gender dimension in content from a wider perspective, given the social theory discussion we have made.

Future Work. At least three areas are ripe for further work. The first is the construction of editing tools for Wikipedia that would help editors detect bias in content, and suggest appropriate actions. The second is a study of individual differences among contributors, as our work analyzed user generated content without considering *who* published and edited it. This aspect can be explored by analyzing how contributors discuss and edit content based on their gender and other individual factors. The last area is a further exploration of bias considering more fine-grained ontology classes and metadata attributes. For instance, it may be possible that gender bias is stronger or weaker for different ontology classes (e.g., *Scientist* vs. *Artist*) or in biographies of people from different regions and religions. Finally it would be helpful to study whether gender bias depends on the quality of an article: does bias decrease with increasing number of edits or other measures of article maturity?

Acknowledgments. We thank Daniela Alarcón Sánchez for fruitful discussion and Luca Chiarandini for tool code, as well as the anonymous reviewers for their advice to improve the paper. This work was partially funded by Grant TIN2012-38741 (Understanding Social Media: An Integrated Data Mining Approach) of the Ministry of Economy and Competitiveness of Spain.

References

[1] Jane Abray. “Feminism in the French Revolution”. In: *The American Historical Review* (1975), pp. 43–62.

[2] Rodrigo Almeida, Barzan Mozafari, and Junghoo Cho. “On the Evolution of Wikipedia.” In: *International Conference on Weblogs and Social Media*. 2007.

[3] Maik Anderka, Benno Stein, and Nedim Lipka. “Predicting quality flaws in user-generated content: the case of Wikipedia”. In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2012, pp. 981–990.

[4] Judd Antin et al. “Gender differences in Wikipedia editing”. In: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*. ACM. 2011, pp. 11–14.

[5] APA. “Publication Manual of the American Psychological Association”. In: Sixth. American Psychological Association, 2000. Chap. General Guidelines for Reducing Bias.

[6] Pablo Aragón et al. “Biographical social networks on Wikipedia: a cross-cultural study of links that made history”. In: *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*. ACM. 2012, p. 19.

[7] David Bamman and Noah A Smith. “Unsupervised Discovery of Biographical Structure from Text”. In: *Transactions of the Association for Computational Linguistics* 2 (2014), pp. 363–376.

[8] CJ Beukeboom. “Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies”. In: *Social Cognition and Communication* (2014), pp. 313–330.

[9] Sergey Brin and Lawrence Page. “The anatomy of a large-scale hypertextual Web search engine”. In: *Computer networks and ISDN systems* 30.1 (1998), pp. 107–117.

[10] Ewa S Callahan and Susan C Herring. “Cultural bias in Wikipedia content on famous persons”. In: *Journal of the American society for information science and technology* 62.10 (2011), pp. 1899–1915.

[11] Andreea S Calude and Mark Pagel. “How do we use language? Shared patterns in the frequency of word use across 17 world languages”. In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 366.1567 (2011), pp. 1101–1107.

[12] Kenneth W Church and William A Gale. “Poisson mixtures”. In: *Natural Language Engineering* 1.02 (1995), pp. 163–190.

[13] Kenneth W Church and Patrick Hanks. “Word association norms, mutual information, and lexicography”. In: *Computational linguistics* 16.1 (1990), pp. 22–29.

[14] Giovanni Luca Ciampaglia and Dario Taraborelli. “MoodBar: Increasing new user retention in Wikipedia through lightweight socialization”. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM. 2015, pp. 734–742.

[15] Benjamin Collier and Julia Bear. “Conflict, criticism, or confidence: an empirical examination of the gender gap in Wikipedia contributions”. In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM. 2012, pp. 383–392.

[16] Simone De Beauvoir. *The second sex*. Random House LLC, 2012.

[17] Michela Ferron and Paolo Massa. “Psychological processes underlying Wikipedia representations of natural and manmade disasters”. In: *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*. ACM. 2012, p. 2.

[18] Amanda Filipacchi. “Wikipedia’s sexism toward female novelists”. In: *The New York Times, April 28th, 2013* (2013).

[19] Susan T Fiske and Steven L Neuberg. “A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation”. In: *Advances in experimental social psychology* 23 (1990), pp. 1–74.

[20] Lucie Flekova, Oliver Ferschke, and Iryna Gurevych. “What makes a good biography?: multidimensional quality analysis based on Wikipedia article feedback data”. In: *Proceedings of the 23rd international conference on World wide web*. International World Wide Web Conferences Steering Committee. 2014, pp. 855–866.

[21] S. Fortunato et al. “On local estimations of PageRank: A mean field approach”. In: *Internet Mathematics* 4.2–3 (2007), pp. 245–266.

[22] Betty Friedan. *The feminine mystique*. WW Norton & Company, 2010.

[23] Jim Giles. "Internet encyclopaedias go head to head". In: *Nature* 438.7070 (2005), pp. 900–901.

[24] Brent Hecht and Darren Gergle. "Measuring self-focus bias in community-maintained knowledge repositories". In: *Proceedings of the fourth international conference on Communities and technologies*. ACM. 2009, pp. 11–20.

[25] Benjamin Mako Hill and Aaron Shaw. "The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation". In: *PloS ONE* 8.6 (2013), e65782.

[26] Daniela Iosub et al. "Emotions under discussion: Gender, status and communication in online collaboration". In: *PloS ONE* 9.8 (2014), e104880.

[27] Jerry A Jacobs. *Revolving doors: Sex segregation and women's careers*. Stanford University Press, 1989.

[28] Piotr Konieczny. "Teaching with Wikipedia and other Wikimedia foundation wikis". In: *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*. ACM. 2010, p. 29.

[29] Robin Tolmach Lakoff. "Language and woman's place". In: *Language in Society* 2, No. 1, Apr. (1973), pp. 45–80.

[30] Shyong Tony K Lam et al. "WP: clubhouse?: an exploration of Wikipedia's gender imbalance". In: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*. ACM. 2011, pp. 1–10.

[31] David Laniado et al. "Emotions and dialogue in a peer-production community: the case of Wikipedia". In: *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*. ACM. 2012, p. 9.

[32] Estella Lauter. *Women as mythmakers: poetry and visual art by twentieth-century women*. Indiana University Press, 1984.

[33] Janette Lehmann et al. "Reader preferences and behavior on Wikipedia". In: *Proceedings of the 25th ACM conference on Hypertext and Social Media*. ACM. 2014, pp. 88–97.

[34] Jens Lehmann et al. "DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia". In: *Semantic Web Journal* (2014).

[35] Michael A Messner. "Sports and male domination: The female athlete as contested ideological terrain". In: *Sociology of sport journal* 5.3 (1988), pp. 197–211.

[36] Jonathan T Morgan et al. "Tea and sympathy: crafting positive new user experiences on Wikipedia". In: *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM. 2013, pp. 839–848.

[37] Brian A Nosek, Mahzarin Banaji, and Anthony G Greenwald. "Harvesting implicit group attitudes and beliefs from a demonstration web site." In: *Group Dynamics: Theory, Research, and Practice* 6.1 (2002), p. 101.

[38] Martha C Nussbaum. "Objectification". In: *Philosophy & Public Affairs* 24.4 (1995), pp. 249–291.

[39] Chitu Okoli et al. "Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership". In: *Journal of the American Society for Information Science and Technology* (2014).

[40] James W Pennebaker, Martha E Francis, and Roger J Booth. "Linguistic inquiry and word count: LIWC 2001". In: *Mahway: Lawrence Erlbaum Associates* 71 (2001), p. 2001.

[41] David G Perry and Kay Bussey. "The social learning theory of sex differences: Imitation is alive and well." In: *Journal of Personality and Social Psychology* 37.10 (1979), p. 1699.

[42] Ulrike Pfeil, Panayiotis Zaphiris, and Chee Siang Ang. "Cultural differences in collaborative authoring of Wikipedia". In: *Journal of Computer-Mediated Communication* 12.1 (2006), pp. 88–113.

[43] Steven T Piantadosi. "Zipf's word frequency law in natural language: A critical review and future directions". In: *Psychonomic bulletin & review* (2014), pp. 1–19.

[44] Felicia Pratto, Peter J Hegarty, and Josephine D Korchmaros. "How communication practices and category norms lead people to stereotype particular people and groups". In: *Stereotype dynamics: Language based approaches to the formation, maintenance, and transformation of stereotypes* (), pp. 293–313.

[45] Jacob Ratkiewicz et al. "Characterizing and modeling the dynamics of online popularity". In: *Physical review letters* 105.15 (2010), p. 158701.

[46] Joseph Reagle and Lauren Rhue. "Gender bias in Wikipedia and Britannica". In: *International Journal of Communication* 5 (2011), p. 21.

[47] Roy Rosenzweig. "Can history be open source? Wikipedia and the future of the past". In: *The Journal of American History* 93.1 (2006), pp. 117–146.

[48] Toni Schmader, Jessica Whitehead, and Vicki H Wysocki. "A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants". In: *Sex Roles* 57.7-8 (2007), pp. 509–514.

[49] M Ángeles Serrano, Alessandro Flammini, and Filippo Menczer. "Modeling statistical properties of written text". In: *PloS ONE* 4.4 (2009), e5372.

[50] Steven S. Skiena and Charles B. Ward. *Who's Bigger?: Where Historical Figures Really Rank*. Cambridge Univ. Press, 2014.

[51] Lynn Smith-Lovin and J Miller McPherson. "You are who you know: A network approach to gender". In: *Theory on gender/feminism on theory* (1993), pp. 223–51.

[52] Sarah Stierch. *Women and Wikimedia Survey 2011*. https://meta.wikimedia.org/wiki/Women_and_Wikimedia_Survey_2011. [Online; accessed April 2015]. 2013.

[53] William Strauss and Neil Howe. *Generations: The history of America's future, 1584 to 2069*. Morrow New York, NY: 1991.

[54] Claudia Wagner et al. "It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia". In: *Ninth International AAAI Conference on Web and Social Media* (2015).

[55] Duncan J Watts and Steven H Strogatz. "Collective dynamics of 'small-world' networks". In: *Nature* 393.6684 (1998), pp. 440–442.

[56] Charlotte Witt and Lisa Shapiro. "Feminist History of Philosophy". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2014. 2014.

[57] George Kingsley Zipf. *Human behavior and the principle of least effort*. Addison-Wesley Press, 1949.

[58] Vinko Zlatić et al. "Wikipedias: Collaborative web-based encyclopedias as complex networks". In: *Physical Review E* 74.1 (2006), p. 016115.