Scientists & Women Scientists: Exploring Gender Biases in Institutional Category Systems

Katie Warburton[1,2] **(kwarburton@cs.toronto.edu)**
Charles Kemp[1] **(c.kemp@unimelb.edu.au)**
Lea Frermann[3] **(lea.frermann@unimelb.edu.au)**

[1] School of Psychological Sciences, University of Melbourne
[2] Department of Computer Science, University of Toronto
[3] School of Computing & Information Systems, University of Melbourne

Abstract

For many categories of people, men are perceived as the more default or typical members whereas women are perceived as more atypical. This bias can lead to an asymmetry in the existence and frequency of categories marked by gendered language. Here we explore the extent to which this asymmetry exists in two institutional category systems: the Library of Congress Subject Headings (LCSH) and English Wikipedia. We find that the LCSH exhibits more bias towards women than Wikipedia, and that in the LCSH this bias has not changed in the last 30 years, whereas Wikipedia shows a noticeable increase in gender balanced categories during the early 2010s. These findings suggest that more can be done to reduce gender bias in the LCSH and demonstrate how principles of typicality and categorization play out in real-world settings.

Keywords: gender bias; categorization; institutional category systems; typicality effects

Introduction

Some members of a category are more typical than others (Rosch, 1973), and typicality is influenced by social biases and stereotypes (Lakoff, 1987). For example, a study of human naming choices in English found that photos of male athletes tended to be tagged with the sport they were playing whereas photos of female athletes were more often tagged with their gender (Harrison, Gualdoni, & Boleda, 2023). This pattern reflects the *people = men bias*, which suggests that many cultures have masculine defaults and perceive men as more generic or typical examples of humans than women (Silveira, 1980; Hamilton, 1991; Merritt & Kok, 1995; Van Berkel, Molina, & Mukherjee, 2017; Bailey, LaFrance, & Dovidio, 2019; Cheryan & Markus, 2020).

The perception of women as atypical leads to asymmetries in how people are categorized. While typical category members are often unmarked and can even stand in for the category as a whole (e.g., mankind meaning humankind), atypical members are often explicitly labelled or qualified with the features that make them deviate from the norm (Lakoff, 1987; Brekhus, 1998). So while men in science might generally be categorized as *scientists*, women are more likely to be categorized as *women scientists*.

Here we explore asymmetries in gendered categories provided by two institutional category systems: the Library of Congress Subject Headings (LCSH)[1] and Wikipedia. The LCSH is a controlled vocabulary of terms such as *Linguistic change* or *Odors in the Bible* that can be applied to books. Updates are centrally reviewed by "cataloguing policy specialists" in the Library of Congress (Library of Congress, n.d.). Wikipedia is a free online encyclopedia that can be edited by anyone, and that incorporates a category system for grouping related articles. As such, both systems are designed to organize and help navigate a large body of information. Although institutional category systems can reflect and reinforce the biases of their creators and the items they categorize (Bowker & Star, 2000), systems like the LCSH and Wikipedia are often perceived as objective, which makes it important to document where and how they perpetuate bias.

Both the LCSH (Berman, 1971; Rogers, 1993; Hyde, 2002; Hobart, 2019) and Wikipedia (Wagner, Garcia, Jadidi, & Strohmaier, 2015; Sun & Peng, 2021; Falenska & Çetinoğlu, 2021) have been previously criticized for being biased against women in several respects. Here we focus on just one aspect of these systems, and our first goal is to quantify the extent to which these systems include many categories that mark women (e.g. *women scientists*) in the absence of similar categories for men. If present, this asymmetry demonstrates a bias against women by treating them as the "odd ones out" whose membership in a category (e.g. *scientists*) is not assumed but must instead be explicitly labelled.

Although we expect to find more categories that mark women overall, some categories such as *nurse* are more strongly associated with women than men, and are likely to be marked accordingly (Duffy & Keir, 2004; White & White, 2006; Garg, Schiebinger, Jurafsky, & Zou, 2018). Our second goal is to test whether category base rates (i.e. the proportions of men and women belonging to the category) affect how categories are marked. We hypothesize that base rates will override the *people=men* bias, leading men to be marked in female-dominated categories, while women remain marked in male-dominated categories or categories with near-equal representation of men and women.

Our third goal is to examine whether gender biases have decreased over time. Bias has been documented in the LCSH since at least 1971 (Berman, 1971) and in Wikipedia since 2011 (Antin, Yee, Cheshire, & Nov, 2011). In the years since there have been dedicated efforts to reduce bias. For example, the Cataloging Lab aims to improve accuracy and reduce bias in the LCSH by facilitating proposals and revisions of headings, and The Women in Red are a group of volunteers who seek to reduce systemic gender bias in Wikipedia. By measuring changes in bias over time we aim to determine how

[1] Note that the LCSH is distinct from the hierarchical Library of Congress Classification that is used by libraries to shelve books.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

much progress has been made and how much work remains to be done.

From one perspective, our work is an exercise in applied cognitive science that aims to understand how principles of typicality and categorization play out in two important real-world settings (the LCSH and Wikipedia). It contributes to a body of research that seeks to deepen our understanding of categorization "in the wild" (Glushko, Maglio, Matlock, & Barsalou, 2008) and our understanding of how social biases are encoded in the datasets, systems and technologies we use (Bailey, Williams, & Cimpian, 2022; Watson, Beekhuizen, & Stevenson, 2023; Warburton, Kemp, Xu, & Frermann, 2024). From another perspective, our project uses institutional category systems to develop large-scale tests of theoretical ideas about categorization, such as the influence of base rates on typicality. While both perspectives may be useful, we mostly emphasize the former.

Gender Bias in Institutional Category Systems

Our first analysis explores the extent to which gender asymmetries exist in the categories of the LCSH and English Wikipedia. In addition to comparing these systems with each other, we compare both with usage frequency data derived from the Google Ngram corpus. Including Ngram data as a baseline helps to determine whether the two institutional systems mirror, mitigate, or exaggerate biases in general English language use.

We start by compiling a broad set of gender categories that aim to be relatively comprehensive, then focus on two subsets of particular interest — categories related to jobs (*Male musicians*) and categories related to ethnicities and nationalities (e.g. *Yoruba men* and *German women*). As discussed later, considering these two subsets allows us to probe the influence of category base rates and of the *people=men* bias on gender asymmetry.

Data & Methods

LCSH and Wikipedia. Our LCSH data are derived from the official LCSH data dump released by the Library of Congress, which contains digital records for headings created as recently as 2024.[2] For Wikipedia we use a preprocessed version of English Wikipedia's 2020 category dump (Lu, 2020). To collect gendered headings we first search for any category that contains the words 'men', 'male', 'women', and 'female'. The singular terms 'man' and 'woman' are excluded because they are not commonly used in Wikipedia (Falenska & Çetinoğlu, 2021) and the LCSH, and when they do appear they are often part of a proper name (e.g. *Isle of Man*, *Spider-Man*), or used to refer to human beings in general (e.g. *Fall of man*).

Next, we remove any category that includes words for men and women (e.g. *European Men's and Women's Team Badminton Championships*). Following Falenska and Çetinoğlu (2021), we then identify pairs such as (*Canadian women*, *Canadian men*) and (*Women scientists*, *Male scientists*) that are identical except for the inclusion of a different gendered noun or adjective.[3] We then group the terms into three classes: W includes unpaired categories that exist only in a female version (e.g. *Female doctors* exists but *Male doctors* does not), M includes categories like *Male caregivers* that exist only in a male version, and WM includes categories that appear in both male and female versions. The total number of gendered categories is $W+M+WM$, and this total equals 2,668 for the LCSH and 15,955 for Wikipedia.

Google Ngram data We derive usage frequencies from version 3 of English Google Ngram (Michel et al., 2011), which includes Part-of-Speech (POS) tags (Lin et al., 2012). These frequencies are based on millions of published books and are therefore roughly representative of the sources reflected by the LCSH and Wikipedia. The LCSH aim to reflect the body of literature that they classify (Library of Congress Policy and Standards Division, 2016), and statements in Wikipedia are supposed to be based on reliable published sources (Wikipedia:Verifiability, 2025).

Using Ngram data we collected bigrams and trigrams such as 'Australian_ADJ men', 'women in science_NOUN', and 'male facial_ADJ attractiveness_NOUN'. Each ngram included exactly one gender word, and for consistency we used the same gender words as for LCSH and Wikipedia. If the gender word is 'female' or 'male', then the gender word must occur before a noun such that it modifies that noun. The words 'women' and 'men' can occur before or after nouns because, as in the LCSH and Wikipedia, 'women' is often used in place of 'female' as a modifier. We applied four additional rules designed to identify the kinds of category labels that frequently appear in Wikipedia and the LCSH,[4] and the final set of ngrams included 273,608 bigrams and 584,488 trigrams. We ignored case when collecting these ngrams, and the frequency of each ngram was defined as its total aggregate frequency between 2010 and 2019.

To simplify the dataset we replace all occurrences of the word 'men' and 'male' with the token $\langle M \rangle$ and 'female' and 'women' with the token $\langle W \rangle$. We then lemmatize the remaining words in each ngram based on their POS tag and aggregate frequencies across ngrams that now have the same form. Finally we pair ngrams (e.g. ('Australian_ADJ men', 'Australian_ADJ women')) as we did for the LCSH and Wikipedia. For each pair, we computed the percentage of gendered occurrences that mention women (%W). We removed all pairs where the frequency of either member of the

[2] See https://id.loc.gov/authorities/subjects.html

[3] *Women* is paired with *male* because in the LCSH, Wikipedia, and Google Ngram *women* often appears as a modifier (instead of female) but the same is rarely true for *men*. In our data sets the distinction between female/male and women/men does not reliably indicate a difference between sex and gender.

[4] (1) An ngram may not include punctuation. (2) Each ngram must have a complete set of POS tags. (3) 'in', 'of', 'for', 'and', or 'by' may not appear as the first or last word of an ngram. (4) All other words must have more than two letters and must be tagged as a noun, proper noun, or adjective except for the last word which cannot be tagged as an adjective.

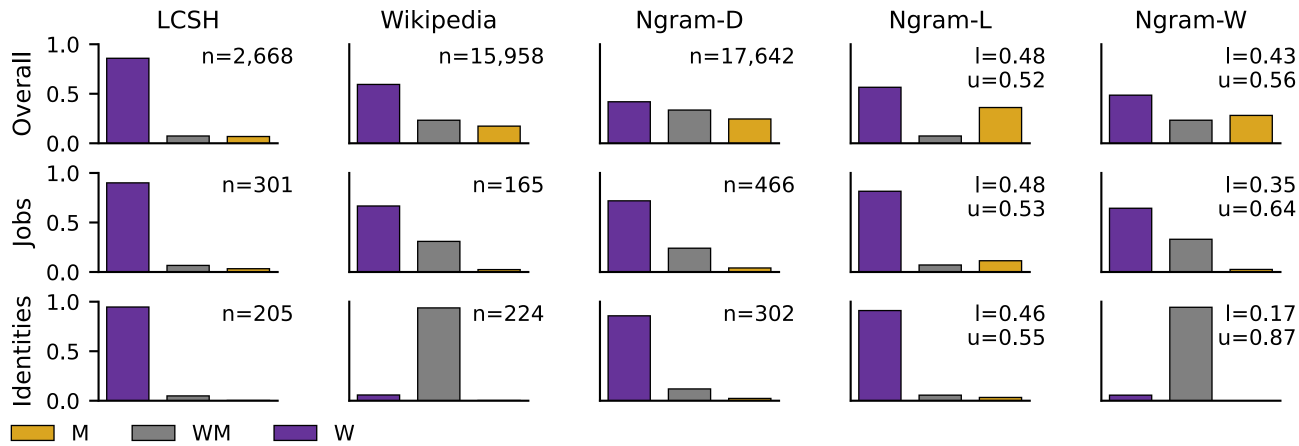

Figure 1: Classification of gendered categories in the LCSH, Wikipedia, and Google Ngram. The top row shows overall results, and the middle and bottom rows show results for jobs and identities. As described in the text, the thresholds for the Ngram-L and Ngram-W classifications are chosen so that the WM proportions match those for LCSH and Wikipedia respectively.

Domain	Datasets
Identities	List of contemporary ethnic groups; List of adjectival and demonymic forms for countries and nations
Jobs	Detailed occupation by sex education age earnings 2019; Gender occupation percentages (Garg et al., 2018); Thesaurus of job titles; Table 3. Gender perception of role nouns (Misersky et al., 2014).

Table 1: Datasets used to compile a list of jobs and identities.

pair was less than 100. The total number of gendered pairs was 17,642.

To make our Ngram data comparable to our LCSH and Wikipedia data we classify each ngram pair as W, M, or WM based on the value of $\%W$. Given lower and upper thresholds l and u, pairs with $\%W < l$ are classified as M, pairs with $l \leq \%W \leq u$ are classified as WM, and pairs with $\%u < W$ are classified as M. We create three classifications, each of which uses different values of l and u. Ngram-D is a default classification that sets $l = 0.4$ and $u = 0.6$. Ngram-L and Ngram-W preserve the property that l and u are equidistant (or nearly so) from 0.5, but adjust the distance between l and u to allow the proportion of WM pairs to match the LCSH and Wikipedia classifications respectively. The values of the thresholds for Ngram-L and Ngram-W are reported in Figure 1.

Jobs and Identities. For simplicity, we refer to both ethnicities and nationalities as identities. We compiled a lists of jobs and identities from sources listed in Table 1. Given these lists, we used string matching to find job categories of the form *(male|female|women|men) JOB* and identity categories of the form *IDENTITY (women |men)* in the LCSH, Wikipedia, and Ngram. For the LCSH, 11.28% of the gendered categories correspond to jobs and 7.68% to identities. For Wikipedia, these percentages are 1.03% and 1.40% respectively, and for Ngram data 2.64% and 1.72%. Because the ngrams are lemmatized, we used lemmatized versions of the jobs and identities for Ngram data only.

Results

Figure 1 shows the relative proportions of W, WM, and M classes for the three data sets. The top row shows that the LCSH has a very high proportion of W categories and very small proportions of WM and M categories overall. This trend holds for both jobs and identities, showing that the LCSH has a strong gender bias against women. Wikipedia also shows this gender bias overall and for jobs, however in both these cases the proportion of WM categories is higher than for the LCSH, indicating less bias than the LCSH. Wikipedia shows almost no gender bias for identities, because almost all identity categories are classified as WM.

To compare the LCSH and Wikipedia to Ngram-D, we performed a proportional odds ordinal regression with classification (W, WM, or M) as the dependent variable and data set (LCSH, Wikipedia, or Ngram) as a predictor. Compared to Ngram-D, the LCSH was 2.05 times more likely to have W classes (95% CI [1.94, 2.16], $p < 0.001$). For Wikipedia, we find a difference of 0.63 (95% CI [0.59, 0.67], $p < 0.001$). Overall both Wikipedia and the LCSH have more W classes and fewer M and WM classes than Ngram-D.

We used the same ordinal regression approach to compare Ngram-D to Wikipedia and the LCSH in the domains of jobs and identities. For jobs, LCSH is 1.24 (95% CI [-0.58, 0.17], $p < 0.001$) times more likely to have W classes than Ngram-D whereas Wikipedia is not significantly different from it (log odds = 0.21, 95% CI [0.81, 1.66], $p = 0.28$). Wikipedia is more similar to Ngram-D than to the LCSH as both have more WM classes. Finally, for identities the LCSH is 1.10 times more likely to have W classes than Ngram-D (95% CI [0.408, 1.784], $p = 0.002$) and Wikipedia is 4.09 times less likely (95% CI [-4.641, -3.535], $p < 0.001$). The LCSH is much more similar to Ngram-D than Wikipedia, largely because

Wikipedia has such a high proportion of WM categories.

When comparing the LCSH to Ngram-L we find that overall as well as for jobs and identities, the LCSH has more W categories and fewer M categories than Ngram. This finding is supported by an ordinal regression indicating that the LCSH is 1.59 times more likely to have W categories (95% CI [1.479, 1.703], $p < 0.001$). The corresponding figure for jobs is 0.75 (95% CI [0.308, 1.194], $p = 0.001$), and for identities the proportions of W are not significantly different (log odds = 0.57, 95% CI [-0.157, 1.292], $p = 0.125$). These results suggest that gender asymmetries are more pronounced overall for the LCSH than for language use, and that this conclusion is robust for jobs but not identities.

A similar comparison between Wikipedia and Ngram-W suggests that Wikipedia is more biased overall, and is 0.49 times more likely to have W categories (95% CI [0.45, 0.53], $p < 0.001$). For jobs we find no significant difference between Wikipedia and Ngram-W, suggesting that Wikipedia's bias for jobs reflects asymmetries in language use (log odds = 0.10, 95% CI [-0.27, 0.47], $p = 0.60$). For identities, Wikipedia is again not significantly different from Ngram-W (log odds = -0.05, 95% CI [-0.78,0.69], $p = 0.91$), but extreme thresholds $l = 0.17$ and $u = 0.87$ are needed to match the WM proportion across Wikipedia and Ngram-W, suggesting that Wikipedia gives balanced treatment to pairs of categories that are skewed in usage.

Our results confirm that the LCSH and Wikipedia both display gender bias, and that this bias is more extreme for LCSH than for Wikipedia. The top row of Figure 1 also suggests that the global bias in both systems is more extreme than the bias found in usage data, but for identities Wikipedia seems less biased than usage data. The following section considers factors that may contribute to these biases.

Base rates and gender bias

The LCSH bias for identities cannot be explained by base rates, because the number of Canadian women (say) is almost identical to the number of Canadian men. Instead, the asymmetry suggests that the LCSH incorporates the default *people=men* bias, which means that categories involving women need to be explicitly marked.

For jobs, however, the base rates of men and women are often different, and our second analysis asks if these differences shape the bias we see in job categories. As mentioned earlier, we hypothesize that the *people=men* bias is overridden in the case of jobs with a higher proportion of women. We consider both actual and perceived base rates for the jobs in our data to see which is the better predictor of bias.

Data & Methods

We derived base rates for our list of jobs by combining US 2019 census data[5] with 2015 US jobs base rates from Garg et al. (2018). Perceived base rates were drawn from Table 2 of Misersky et al. (2014), which includes results from an experiment in which people were asked to "estimate [...] the extent to which the presented social and occupational groups actually consisted of women and men." For LCSH, Wikipedia, and Ngram-D the number of categories for which we have actual and perceived base rates is shown in Figure 2.

[5]Available at: https://www.census.gov/data/tables/2022/demo/acs-2019.html

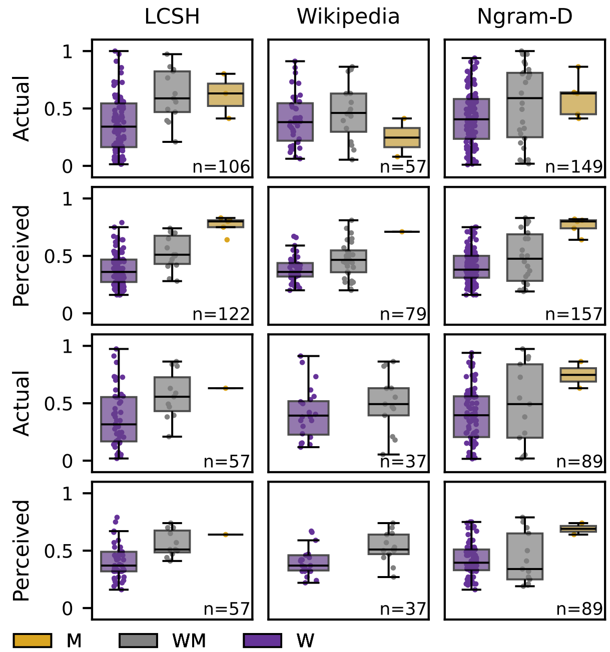

Figure 2: Boxplot showing the distribution of actual and perceived gender base rates across cases W, WM, and M in the LCSH, Wikipedia, and Ngram-D. The top two rows include categories for which we have actual (top row) or perceived (second row) base rates. The bottom rows include only categories for which we have both actual and perceived base rates.

Results

The top two rows of Figure 2 show the distributions of actual and perceived base rates for W, WM, and M categories in each classification. Across all three data sets the purple boxes (W categories) are lower than the grey boxes (WM categories), indicating that women are marked in male-dominated categories. The yellow boxes (M categories) are consistently higher than the grey boxes for the second row but not the first row, suggesting that perceived base rates account better than actual base rates for gender bias across the three data sets.

To confirm this impression we again used proportional odds ordinal regression models with classification (W, WM or M) as the dependent variable. We compared three models for each data set: the first was a null model with an intercept only, and the second and third included actual and perceived base rates as predictors. To enable these models to be compared we included only categories for which we have both actual and perceived base rates, and the bottom two rows of Figure 2 show distributions for these categories.

BIC scores for the regressions are reported in Table 2. For

	LCSH	Wiki	Ngram-D
Null	7.92	4.51	**0**
Actual	5.85	6.68	1.95
Perceived	**0**	**0**	3.45

Table 2: BIC scores for each proportional odds ordinal regression model. Rows correspond to the model type and columns to the dataset. BICs are reported as offsets from the smallest entry in each column, which means that 0 indicates the best model in each column.

both LCSH and Wikipedia, the model based on perceived base rates accounts best for the data. Following the conventions for interpreting BIC scores proposed by (Raftery, 1995), the magnitudes of the BIC scores provide positive but not strong evidence that the perceived base rates model is the best of the three. For Ngram-D, however the null model scores slightly better than the two models that incorporate base rates.

The results just described suggest that gender asymmetries in LCSH and Wikipedia are influenced by base rates, and that perceived base rates account for these asymmetries better than do actual base rates. Overall our results for identities and jobs show asymmetry in the LCSH is shaped by both base rates and the *people=men* bias, whereas for Wikipedia we have evidence only of the role of base rates.

Gender Bias Across Time

Gender bias has become less extreme in Western society over time (Bhatia & Bhatia, 2021), and this section explores whether the asymmetries we have documented in the LCSH, Wikipedia and Ngram have become weaker over the past few decades.

Data & Methods

We use the date a LCSH digital record was created for the heading as a proxy for its date of addition to the system. Because the system was not digitized until 1986, any heading added before has a date of addition of 1986. Digitizing all existing headings may have taken some time, and we therefore only consider changes in gendered subject headings between 1990 and 2019 (the most recent Ngram year).

The Wikipedia dataset does not contain the date a category was added to the system. Most Wikipedia categories, however, have an associated web page, and we approximate a category's date of addition by extracting the date on which its associated page was first revised. This approach yielded addition dates for 94% of the gendered Wikipedia categories. We consider the period from 2004 to 2019 because Wikipedia's category system did not exist prior to 2004.

For Ngram we create gendered phrase pairs using the same method as study 1 except for two key differences. First, the phrase pairs are created for each year between 1990 and 2019. Second, the minimum frequency for each phrase in a gendered phrase pair is 20 as opposed to 100. A lower threshold is appropriate because the counts for each pair are now based on the frequency of gendered phrases in a single year as opposed to an aggregate frequency across ten years.

For each of the three data sets, we compute the proportion of W, WM and M categories for each year. We use the default frequency boundaries of 0.40 and 0.60 when classifying ngrams as W, WM, or M.

Results

Figure 3 shows how the proportions of W, WM, and M categories have changed over time. The LCSH shows no overall change despite the addition of 1,960 gendered headings between 1990 and 2019. For jobs and identities there appears to be a gradual increase in the proportion of WM, although this change is very small.

For Wikipedia the picture is very different. The overall plot reveals an increase in WM categories since 2004, and this pattern is even clearer for jobs and especially dramatic for identities. In 2004, all categories were classified as W, but in 2012 there was a sharp increase in the proportion of WM categories, and by 2019, the categories added over the previous seven years left the complete set of identities relatively balanced.

For Ngram-D there are no discernible trends for jobs and identities, and the proportions of W, WM, and M appear stable over time. For the full set of gendered ngrams, however, the proportion of M categories gradually increases between 1990 and 2019, although this increase is not very large.

Overall, our results suggest that the gender asymmetries in the LCSH and Ngram have been relatively stable over the past three decades. In contrast, Wikipedia asymmetries for jobs and identities have decreased since the early 2010s, and this decrease is especially striking in the case of identities.

Why has Wikipedia changed but not LCSH?

Wikipedia has policies specifying how the intersection between gender and ethnicity should be approached, and the page outlining these policies (Wikipedia:Categorizing articles about people, 2025) was created in 2005, before the 2012 change we saw for identities in Figure 3. The 2012 change may reflect the publication in the early 2010s of papers highlighting gender bias in Wikipedia (Antin et al., 2011; Reagle & Rhue, 2011) and the emergence of groups such as the Women in Red (Wikipedia:WikiProject Women in Red, 2025) dedicated to reducing these biases.

For the LCSH, gender biases have been criticized since 1971 (Berman, 1971) if not before, and although groups have emerged with the goal of reducing these biases (e.g. the Cataloging Lab), our data suggest that gender asymmetries have remained stable since 1990. One reason why Wikipedia alone has changed in recent decades could be the effort required before a new LCSH heading is approved. Proposals for new headings must be approved by the Library of Congress, and must be accompanied by evidence that they meet the requirement of literary warrant (Svenonius, 2000). In contrast,

Figure 3: Cumulative proportions of gendered categories in the LCSH, Wikipedia, and Google Ngram over time. The top row shows overall results, and the middle and bottom rows show results for jobs and identities.

Wikipedia follows a collaborative editing model that allows for very rapid changes.

Discussion

We explored gender asymmetries in two institutional category systems and found three key results. First, although both the LCSH and Wikipedia contain gender bias, the LCSH tends to exaggerate bias found in language use whereas Wikipedia mirrors it (jobs) or mitigates it (identities). Second, while the gender asymmetries in both the LCSH and Wikipedia are influenced by perceived base rate, only the LCSH displays evidence of marking behaviour consistent with the *people=men* bias. Third, gender asymmetries have remained relatively unchanged in the LCSH and Ngram over the past three decades, however, asymmetries in Wikipedia's jobs and identities have decreased with time, most notably in the case of identities.

Our results demonstrate how institutional category systems can encode people's biases through unequal labelling of various groups, and that ideas of category typicality, which have mostly been investigated in the context of cultural categories that we carry around in our heads, extend to these systems. In most cases, the intent behind marking female categories is probably not to cause harm, and explicit labelling can address historical exclusion (e.g. there was an explosion of LCSH categories for women in the 1970s (Hyde, 2002)). Categories, however, still shape how we perceive the world (Goldstone, Lippa, & Shiffrin, 2001), and biased systems can reinforce a picture of women as the "odd ones out" whose belonging is not assumed but specified (Brekhus, 1998).

Our results demonstrate that these biases are not always inevitable. Work can be done to alleviate and remove them from institutional systems as in the case of identities in Wikipedia. Although we speculate that this is at least in part due to Wikipedia's more collaborative and accessible editing model, future work is needed to better understand the factors affecting change and provide a better understanding of when debiasing efforts succeed.

Our analysis is limited in several respects. First, we did not focus on the subset of categories shared by the LCSH, Wikipedia, and Ngram, and some observed differences may reflect variation in the concepts represented in each system. Nonetheless, we believe the differences we find are meaningful, as they illustrate how gender is marked within each system as a whole. Second, our automatic text processing may have influenced our results as our use of string matching to align jobs and identities with categories prioritized precision over recall. Moreover, we excluded morphologically gendered terms (e.g., actress), which led to a few unpaired cases (e.g., male actor). These are rare in our data and unlikely to affect our overall assessment of asymmetry. Third, our conclusions are based on asymmetries in English ngrams and Western category systems which may not generalize to other cultural or linguistic contexts. Finally, we base our analysis on an over-simplified binary view of gender due to the very small fraction of categories pertaining to non-binary or gender-diverse people in all systems included in our work. Our methods could be used in future studies that take a more nuanced view of gender.

In summary, we demonstrate how gender asymmetries manifest in institutional category systems and how principles of categorization and typically can manifest out in the real world. The stability of LCSH asymmetries versus the decline of some of Wikipedia's suggests that structural factors —such as editorial accessibility and collaborative editing—may play a critical role in bias reduction. Understanding these mechanisms is essential for developing more balanced classification systems. Although the LCSH has not changed much in the past three decades, we hope that the next three will lead to a more balanced presentation of gender.

Acknowledgements

We would like to thank anonymous reviewers for constructive comments on an earlier version of the paper. KW is funded by the U of T–UoM IRTG program. This work was supported by ARC Future Fellowship FT19010020.

References

Antin, J., Yee, R., Cheshire, C., & Nov, O. (2011). Gender differences in Wikipedia editing. In *Proceedings of the 7th international symposium on Wikis and open collaboration* (pp. 11–14).

Bailey, A. H., LaFrance, M., & Dovidio, J. F. (2019). Is man the measure of all things? a social cognitive account of androcentrism. *Personality and Social Psychology Review*, *23*(4), 307–331.

Bailey, A. H., Williams, A., & Cimpian, A. (2022). Based on billions of words on the internet, PEOPLE = MEN. *Science Advances*, *8*(13). doi: 10.1126/sciadv.abm2463

Berman, S. (1971). *Prejudice and antipathies: A tract on the LC subject heads concerning people*. Metuchen, NJ: Scarecrow Press.

Bhatia, N., & Bhatia, S. (2021). Changes in gender stereotypes over time: A computational analysis. *Psychology of Women Quarterly*, *45*(1), 106–125.

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT Press.

Brekhus, W. (1998). A sociology of the unmarked: Redirecting our focus. *Sociological Theory*, *16*(1), 34–51.

Cheryan, S., & Markus, H. R. (2020). Masculine defaults: Identifying and mitigating hidden cultural biases. *Psychological Review*, *127*(6), 1022-1052.

Duffy, S. A., & Keir, J. A. (2004). Violating stereotypes: Eye movements and comprehension processes when text conflicts with world knowledge. *Memory & Cognition*, *32*(4), 551–559.

Falenska, A., & Çetinoğlu, Ö. (2021). Assessing gender bias in Wikipedia: Inequalities in article titles. In M. Costa-jussa, H. Gonen, C. Hardmeier, & K. Webster (Eds.), *Proceedings of the 3rd workshop on gender bias in natural language processing* (pp. 75–85). Association for Computational Linguistics. doi: 10.18653/v1/2021.gebnlp-1.9

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Glushko, R. J., Maglio, P. P., Matlock, T., & Barsalou, L. W. (2008). Categorization in the wild. *Trends in cognitive sciences*, *12*(4), 129–135.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27–43.

Hamilton, M. C. (1991). Masculine bias in the attribution of personhood: People= male, male= people. *Psychology of Women Quarterly*, *15*(3), 393–402.

Harrison, S., Gualdoni, E., & Boleda, G. (2023). Run like a girl! sport-related gender bias in language and vision. In *Findings of the association for computational linguistics: ACL 2023* (pp. 14093–14103).

Hobart, E. (2019). Reflecting diversity in the library of congress subject headings. *Social justice and activism in libraries: Essays on diversity and change*.

Hyde, K. (2002). *From suffrage to postfeminism: an evolution of the library of congress subject treatment of women's issues*. Unpublished doctoral dissertation, University of North Carolina at Chapel Hill.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press.

Library of Congress. (n.d.). *Process for adding and revising Library of Congress Subject Headings.* Retrieved 2025-02-02, from https://www.loc.gov/aba/cataloging/subject/lcsh-process.html

Library of Congress Policy and Standards Division. (2016, June). "Module 1.5: Introduction to LCSH". In *Library of Congress Subject Headings: Online Training*. Retrieved from https://www.loc.gov/catworkshop/lcsh/Module-1-5/index.html

Lin, Y., Michel, J.-B., Lieberman, E. A., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the Google Books NGram corpus. In *Proceedings of the ACL 2012 system demonstrations* (pp. 169–174).

Lu, K. (2020, September). *Wikipedia category tree.* Retrieved from https://www.kaggle.com/datasets/kevinlu1248/wikipedia-category-tree

Merritt, R. D., & Kok, C. J. (1995). Attribution of gender to a gender-unspecified individual: An evaluation of the people= male hypothesis. *Sex Roles*, *33*, 145–157.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, G. B., … others (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182.

Misersky, J., Gygax, P. M., Canal, P., Gabriel, U., Garnham, A., Braun, F., … others (2014). Norms on the gender perception of role nouns in Czech, English, French, German, Italian, Norwegian, and Slovak. *Behavior research methods*, *46*, 841–871.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 111–163.

Reagle, J., & Rhue, L. (2011). Gender bias in Wikipedia and Britannica. *International Journal of Communication*, *5*, 21.

Rogers, M. N. (1993). Are we on equal terms yet? Subject headings concerning women in LCSH, 1975-1991. *Library Resources and Technical Services*, *37*(2), 181–96.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328-350. doi: https://doi.org/10.1016/0010-0285(73)90017-0

Silveira, J. (1980). Generic masculine words and thinking. *Women's Studies International Quarterly*, *3*(2-3), 165–178.

Sun, J., & Peng, N. (2021). Men are elected, women are married: Events gender bias on Wikipedia. *arXiv preprint arXiv:2106.01601*.

Svenonius, E. (2000). Subject Languages: Introduction, Vocabulary Selection, and Classification. In *The Intellectual Foundation of Information Organization.* Cambridge, MA: MIT Press.

Van Berkel, L., Molina, L. E., & Mukherjee, S. (2017). Gender asymmetry in the construction of american national identity. *Psychology of Women Quarterly*, *41*(3), 352–367.

Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a man's Wikipedia? assessing gender inequality in an online encyclopedia. In *Proceedings of the international aaai conference on web and social media* (Vol. 9, pp. 454–463).

Warburton, K., Kemp, C., Xu, Y., & Frermann, L. (2024). Quantifying bias in hierarchical category systems. *Open Mind*, *8*, 102–130.

Watson, J., Beekhuizen, B., & Stevenson, S. (2023). What social attitudes about gender does BERT encode? leveraging insights from psycholinguistics. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 6790–6809).

White, M. J., & White, G. B. (2006). Implicit and explicit occupational gender stereotypes. *Sex roles*, *55*, 259–266.

Wikipedia:Categorizing articles about people. (2025, January 23). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Wikipedia:Categorizing_articles_about_people&oldid=1271312953

Wikipedia:Verifiability. (2025, January 26). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Wikipedia:Verifiability&oldid=1272031077

Wikipedia:WikiProject Women in Red. (2025, January 30). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Wikipedia:WikiProject_Women_in_Red&oldid=1272889530