Gender Gap in Natural Language Processing Research: Disparities in Authorship and Citations

Saif M. Mohammad
National Research Council Canada
Ottawa, Canada
saif.mohammad@nrc-cnrc.gc.ca.

Abstract

Disparities in authorship and citations across genders can have substantial adverse consequences not just on the disadvantaged gender, but also on the field of study as a whole. In this work, we examine female first author percentages and the citations to their papers in Natural Language Processing. We find that only about 29% of first authors are female and only about 25% of last authors are female. Notably, this percentage has not improved since the mid 2000s. We also show that, on average, female first authors are cited less than male first authors, even when controlling for experience and area of research. We hope that recording citation and participation gaps across demographic groups will improve awareness of gender gaps and encourage more inclusiveness and fairness in research.

1 Introduction

Gender gaps are quantitative measures of the disparities in social, political, intellectual, cultural, or economic success due to one's gender or gender identity. They can also refer to disparities in access to resources (such as healthcare, education, economic benefits, and political freedom) or attitudes, which in turn lead to disparities in success. We need to pay attention to gender gaps not only because they are inherently unfair but also because better gender balance leads to higher productivity, better health and well-being, greater economic benefits, better decision making, as well as political and economic stability (Skjelsboek and Smith, 2001; Woetzel et al., 2015; Hakura et al., 2016; Rao and Tilt, 2016; Mehta et al., 2017; Gallego and Gutiérrez, 2018).

The Global Gender Gap Report, a study published by the World Economic Forum every year since 2006, examines data from more than 144 countries to determine the magnitude of gender-based disparities. The 2018 Global Gender Gap Report highlighted the gender gap between men and women in Artificial Intelligence as particularly alarming (WEC, 2018).[1] It indicated that only 22% of the professionals in AI are women and that this low representation in a transformative field requires urgent action—otherwise, the AI gap has the potential to widen other gender gaps. Other studies have identified substantial gender gaps in science (Håkanson, 2005; Larivière et al., 2013; King et al., 2017; Andersen and Nielsen, 2018).

This work examines gender gaps in Natural Language Processing (NLP) research. NLP is a broad interdisciplinary field that includes scholarly work on language and computation with influences from Artificial Intelligence, Computer Science, Linguistics, Psychology, and Social Sciences to name a few. Specifically, we examine NLP literature in the ACL Anthology (AA) for disparities in female authorship. We also conduct experiments to determine whether female first authors are cited more or less than male first authors, based on citation counts extracted from Google Scholar (GS).

The ACL Anthology is a digital repository of public domain, free to access, articles on NLP.[2] It includes papers published in the family of ACL conferences as well as in other NLP conferences such as LREC and RANLP.[3] When it was first launched in 2002, it included 3,100 NLP papers. As of June 2019, at the start of this project, it provided access to the metadata and full text for ∼50K articles published since 1965 (the year of the first ACL conference). It is the largest single source of literature on NLP.

Google Scholar is a free web search engine for academic literature—peer reviewed journals, conferences, preprints, patents, theses, technical re-

[1] http://www3.weforum.org/docs/WEF_GGGR_2018.pdf
[2] https://www.aclweb.org/anthology/
[3] ACL licenses its papers with a Creative Commons Attribution 4.0 International License.

ports, etc.[4] Through it, users can access the metadata associated with an article and often the full text of the article as well. A key aspect of the metadata is the number of citations that an article has received. Google Scholar does not provide information on how many articles are included in its database. However, scientometric researchers have estimated that it included about 389 million documents in January 2018 (Gusenbauer, 2019)—making it the world's largest source of academic information. Thus, it is not surprising that there is growing interest in the use of Google Scholar information to draw inferences about scholarly research (Orduña-Malea et al., 2014; Mingers and Leydesdorff, 2015; Martín-Martín et al., 2018).

We extracted and aligned information from the ACL Anthology (AA) and Google Scholar to create a dataset of tens of thousands of NLP papers and their citations as part of a broader project on analyzing NLP Literature.[5] We refer to this dataset as the *NLP Scholar Dataset*. In this paper, we use the NLP Scholar Dataset to study authorship and citation disparities across males and females in tens of thousands of papers. We do not investigate the reasons for the gender gap. However, we will note that the reasons are often complex, intersectional, and difficult to disentangle. We hope that this work will increase awareness of gender gaps amongst the researchers and inspire concrete steps to improve inclusiveness and fairness in research.

It should also be noted that, even though this paper focuses on two genders (male and female), there are many aspects to demographic diversity including: representation from various gender identities; representation from various nationalities and race; representation by people who speak a diverse set of languages; diversity by income, age, physical abilities, etc. All of these factors impact the breadth of technologies we create, how useful they are, and whether they reach those that need it most.

All of the data and interactive visualizations that are part of this project are freely available through the project homepage.[6]

[4]https://scholar.google.com

[5]Mohammad (2019) presents an overview of the many research directions pursued, using this data. Notably, Mohammad (2020a) explores questions such as: how well cited are papers of different types (journal articles, conference papers, demo papers, etc.)? how well cited are papers published in different time spans? how well cited are papers from different areas of research within NLP? etc. Mohammad (2020c) presents an interactive visualization tool that allows users to search for relevant related work in the ACL Anthology.

[6]http://saifmohammad.com/WebPages/nlpscholar.html

2 Related Work

Gender differences in authorship and citations have been studied in various fields and cross-sections of research. Most of these have found substantial gender disparities in favor of male researchers. They include work on journals of library and information science (Håkanson, 2005), on articles from the Web of Science (for Sociology, Political Science, Economics, Cardiology and Chemistry) (Ghiasi et al., 2016; Andersen and Nielsen, 2018), on articles from PubMed life science and biomedical research (Mishra et al., 2018), on articles from fifty disciplines published in JSTOR (King et al., 2017), and on publications from US research universities (Duch et al., 2012). There also exists some work that shows that in fields such as linguistics (LSA, 2017) and psychology (Willyard, 2011), the gender balance is either close to parity or tilted in favor of women. Our work examines gender gaps in NLP.

There is some prior related work on the authorship of NLP papers, for example, Schluter (2018) showed, with a mathematical model, that there are barriers in the paths of women researchers, delaying their attainment of mentorship status (as estimated through last author position in papers). Anderson et al. (2012) examine papers from 1980 to 2008 to track the ebb and flow of topics within NLP, and the influence of researchers from outside NLP on NLP. Vogel and Jurafsky (2012) examined about 13,000 papers from 1990 to 2008 to determine basic authorship statistics by gender. Authors are assigned a gender by a combination of automatic and manual means. The automatic method relies on lists of baby names from various languages. They find that female authorship has been steadily increasing from the 1980s to about 27% in 2007. Our work examines a much larger set of NLP papers published from 1965 to 2018, re-examines some of the questions raised in Vogel and Jurafsky (2012), and explores several new questions, especially on first author gender and disparities in citation. We also show results when controlling for various factors such as experience, sub-field within NLP, venue of publication, and paper type.

3 Data

We extracted and aligned information from the ACL Anthology (AA) and Google Scholar to create a dataset of tens of thousands of NLP papers and their citations. We aligned the information across AA and GS using the paper title, year of publi-

cation, and first author last name. Details about the dataset, as well as an analysis of the volume of research in NLP over the years, are available in Mohammad (2020b). We summarize relevant information below, along with additional processing to infer author gender and author experience to facilitate the gender gap analysis.

3.1 ACL Anthology Data

The ACL Anthology is available through its website and a `github` repository.[7] We extracted paper title, names of authors, year of publication, and venue of publication from the repository.[8]

As of June 2019, AA had ∼50K entries; however, this includes some entries that are not truly research publications (for example, forewords, prefaces, programs, schedules, indexes, invited talks, appendices, session information, newsletters, lists of proceedings, etc.). After discarding them, we are left with 44,894 papers.[9]

Inferring Author Gender: Despite possessing rich metadata for each of the papers, the ACL Anthology does not record demographic information of the authors. We made use of three other resources to infer author gender:

1. A list of 11,932 AA authors and their genders provided by Vogel and Jurafsky (2012) (VJ-AA list) (3,359 female and 8,573 male).[10]
2. A list of 55,924 first names that are strongly associated with females and 30,982 first names that are strongly associated with males, that we generated from the US Social Security Administration's (USSA) published database of names and genders of newborns.[11]
3. A list of 26,847 first names that are strongly associated with females and 23,614 first names that are strongly associated with males, that we generated from a list of 9,300,182 PUBMED authors and their genders (Torvik and Smalheiser, 2009; Smith et al., 2013).[12]

	P	R	F
list derived from USSA names	98.4	69.8	81.7
list derived from PUBMED names	98.3	81.4	89.1

Table 1: Precision (P), Recall (R), and F-score (F) of predicting the gender of authors in the VJ-AA list based on their first names (using first name–gender lists).

We determined first name–gender association, by calculating the percentages of first names corresponding to male and female genders as per each of the PUBMED and USSNA fullname–gender lists. We consider a first name to be strongly associated with a gender if the percentage is $\geq 95\%$.[13] We determined the accuracy of this gender prediction approach on AA authors by comparing its predictions to the genders determined by manual curation in the VJ-AA list. Table 1 shows the results.

Given the high precision (over 98%) of the USSNA and PUBMED lists of gender-associated first names, we use them (in addition to the VJ-AA list) to determine the gender of AA authors. We search for AA author names in the various gender-associated lists in the following order until a match is found and the corresponding gender is assigned to the author:

1. Check if the author's full name matches an entry in the VJ-AA list.
2. Check if the first and last name of the author match the first and last name of an author in the VJ-AA list (ignoring middle names).
3. Check if the first name of author matches an entry in the USSNA first name–gender list.
4. Check if the first name of author matches an entry in the PUBMED first name–gender list.

Eventually, we were able to determine the gender for 28,682 (76%) of the 37,733 AA authors; for the first authors of 37,297 (83%) AA papers (we will refer to this subset of papers as AA*), and for the last authors of 39,368 (88%) AA papers (we will refer to this subset of papers as AA**).[14]

[7] https://www.aclweb.org/anthology/ https://github.com/acl-org/acl-anthology

[8] Multiple authors can have the same name and the same authors may use multiple variants of their names in papers. The AA volunteer team handles such ambiguities using both semi-automatic and manual approaches (fixing some instances on a case-by-case basis). Additionally, AA keeps a file that includes canonical forms of author names.

[9] We used simple keyword searches for terms such as *foreword, invited talk, program, appendix* and *session* in the title to pull out entries that were likely to not be research publications. These were then manually examined to verify that they did not contain any false positives.

[10] https://nlp.stanford.edu/projects/gender.shtml

[11] https://www.ssa.gov/oact/babynames/limits.html

[12] https://experts.illinois.edu/en/datasets/genni-ethnea-for-the-author-ity-2009-dataset-2

[13] A choice of other percentages such as 90% or 99% would also have been reasonable.

[14] We acknowledge that individuals may identify with various non-binary gender identities, and they might be facing marked disparities. We also acknowledge that, despite the presence of a large expatriate population in the US, the US census information is not representative of the names of children from around the world. Further, Chinese origin names tend not to be as strongly associated with gender as names from other parts of the world. Thus, Vogel and Jurafsky (2012) made special effort to include a large number of Asian AA authors in their list.

NLP Academic Age as a Proxy for Experience in NLP: First author percentage may vary due to many factors including: experience, area of research within NLP, and venue of publication. To gauge experience, we use the number of years one has been publishing in AA; we will refer to to this as the *NLP Academic Age*. So if this is the first year one has published in AA, then their NLP academic age is 1. If one published their first AA paper in 2001 and their latest AA paper in 2018, then their academic age is 18.

Note that NLP academic age is not always an accurate reflection of one's experience in publishing NLP papers because it is possible to publish NLP papers strictly outside of AA for many years before publishing one's first paper in AA. However, we expect the number of such instances to be small.

3.2 Google Scholar Data

Google Scholar was launched in November 2004 and has undergone several rounds of refinements since. Notably, since 2012, it allowed scholars/researchers to create and edit public author profiles called *Google Scholar Profiles*. Authors can include their papers (along with their citation information) on this page.

We extracted citation information from Google Scholar profiles of authors who published at least three papers in the ACL Anthology.[15] This yielded citation information for 1.1 million papers in total. We will refer to this dataset as the *NLP Subset of the Google Scholar Dataset*, or *GScholar-NLP* for short. Note that GScholar-NLP includes citation counts not just for NLP papers, but also for non-NLP papers published by authors who have at least three papers in AA.

GScholar-NLP includes 32,985 of the 44,894 papers in AA (about 75%). We will refer to this subset of the ACL Anthology papers as AA′. The citation analyses presented in this paper are on AA′.

4 Gender Gap in Authorship

We use the dataset of papers with known gender information about their authors (AA* and AA** described in §3.1) to answer a series of questions on female authorship in AA. First author is a privileged position in the author list that is usually reserved for the researcher that has done the most work and writing. In NLP, first authors are also often students. Thus we are especially interested in investigating gender gaps that effect them. The last author position is often reserved for the most senior or mentoring researcher. Due to space constraints, we explore last author disparities only briefly in this paper (in Q1), but will explore more in future work.

Q1. What percentage of the authors in AA are female? What percentage of the AA papers have female first authors (FFA)? What percentage of the AA papers have female last authors (FLA)? How have these percentages changed since 1965?

A. Overall, about 29.7% of the 28,682 authors (whose gender we were able to infer) are female; about 29.2% of the first authors in 37,297 AA* papers are female; and about 25.5% of the last authors in 39,368 AA** papers are female. Figure 1 shows how these percentages have changed over the years.

Discussion: Across the years, the percentage of female authors overall is close to the percentage of papers with female first authors. (These percentages are around 28% and 29%, respectively, in 2018.) However, the percentage of female last authors is markedly lower. (Hovering at about 25% in 2018.) These numbers indicate that, as a community, we are far from obtaining male–female parity. A further striking (and concerning) observation is that the female author percentages have not improved since the year 2006.

To put these numbers in context, the percentage of female scientists worldwide (considering all areas of research) has been estimated to be around 30%. The reported percentages for many computer science sub-fields are much lower.[16] The percentages are much higher for certain other fields such as psychology (Willyard, 2011) and linguistics (LSA, 2017).

Q2. How does FFA vary by paper type and venue?

A. Figure 2 shows FFA percentages by paper type and venue.

Discussion: Observe that FFA percentages are lowest for CoNLL, EMNLP, IJCNLP, and system demonstration papers (21% to 24%). FFA percentages for journals, other top-tier conferences, SemEval, shared task papers, and tutorials are the next lowest (24% to 28%). The percentages are markedly higher for LREC, *Sem, and RANLP (33% to 36%), as well as for workshops (31.7%).

[15] This is explicitly allowed by GS's robots exclusion standard. This is also how past work has studied Google Scholar (Khabsa and Giles, 2014; Orduña-Malea et al., 2014; Martín-Martín et al., 2018).

[16] https://unesdoc.unesco.org/ark:/48223/pf0000235155

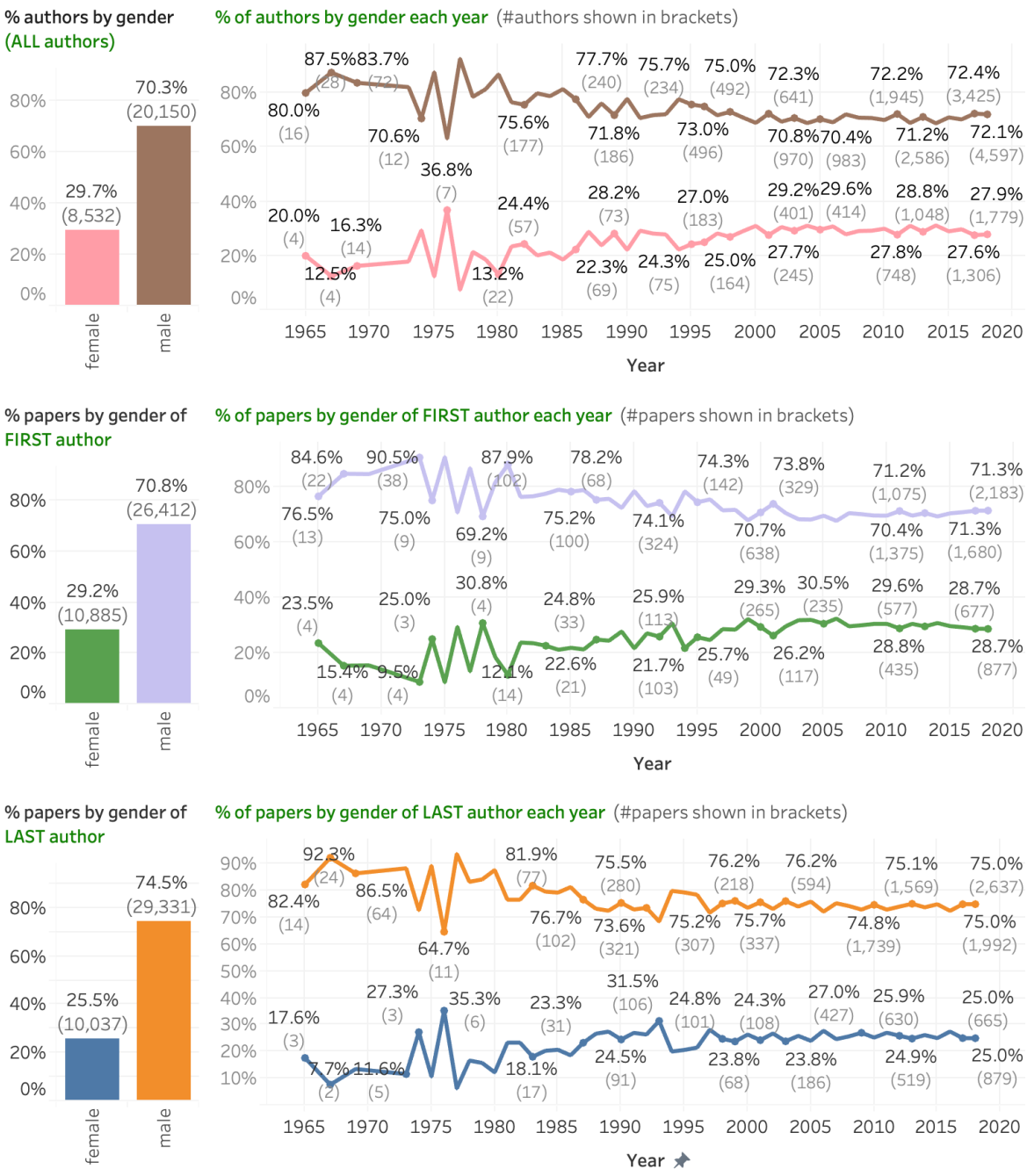

Figure 1: Female authorship percentages in AA over the years: overall, as first author, and as last author.

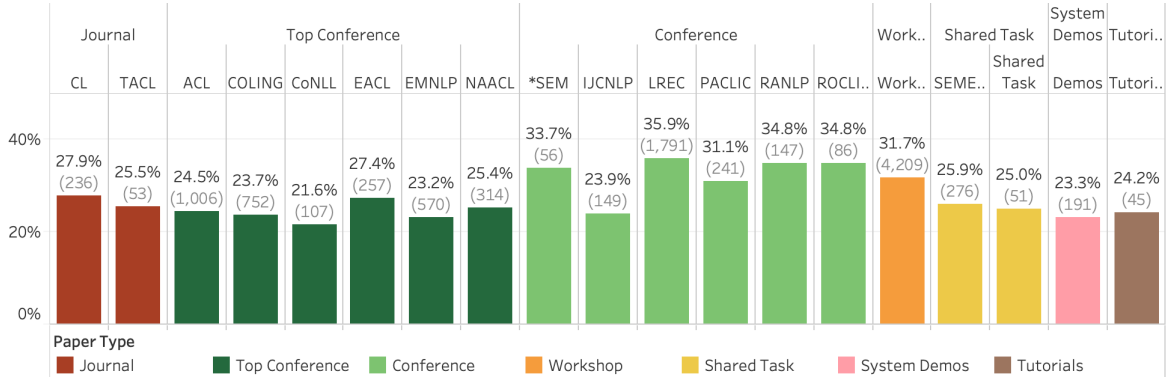

Figure 2: FFA percentage by venue and paper type. The number of FFA papers is shown in parenthesis.

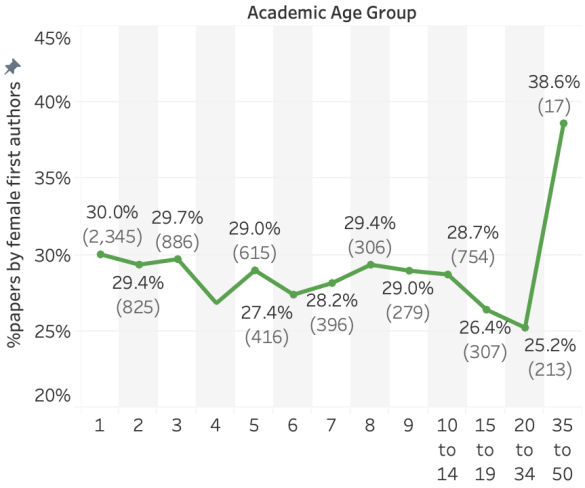

Figure 3: FFA percentage by academic age. The number of FFA papers is shown in parenthesis.

Q3. How does female first author percentage change with NLP academic age?

A. In order to determine these numbers, every paper in AA* was placed in a bin corresponding to NLP academic age: if the paper's first author had an academic age of 1 in the year when the paper was published, then the paper is placed in bin 1; if the paper's first author had an academic age of 2 in the year when the paper was published, then the paper is placed in bin 2; and so on. The bins for later years contained fewer papers. This is expected as senior authors in NLP often work with students, and students are encouraged to be first authors. Thus, we combine some of the bins in later years: one bin for academic ages between 10 and 14; one for 15 to 19; one for 20 to 34; and one for 35 to 50. Once the papers are assigned to the bins, we calculate the percentage of papers in each bin that have a female first author. Figure 3 shows the results.

Discussion: Observe that, with the exception of the 35 to 50 academic age bin, FFA% is highest (30%) at age 1 (first year of publication). There is a period of decline in FFA% until year 6 (27.4%)—this difference is statistically significant (t-test, $p < 0.01$). This might be a potential indicator that graduate school has a progressively greater negative impact on the productivity of women than of men. (Academic age 1 to 6 often correspond to the period when the first author is in graduate school or in a temporary post-doctoral position.) After year 6, we see a recovery back to 29.4% by year 8, followed by a period of decline once again.

Q4. How does female first author percentage vary by area of research (within NLP)? Which areas have higher-than-average FFA%? Which areas have lower-than-average FFA%? How does FFA% correlate with popularity of an area—that is, does FFA% tend to be higher- or lower-than-average in areas where lots of authors are publishing?

A. We use word bigrams in the titles of papers to sample papers from various areas.[17] The title has a privileged position in a paper. Primarily, it conveys what the paper is about. For example, a paper with *machine translation* in the title is likely about machine translation. Figure 4 shows the list of top 66 bigrams that occur in the titles of more than 100 AA* papers (in decreasing order of the bigram frequency). For each bigram, the figure also shows the percentage of papers with a female first author. In order to determine whether there is a correlation between the number of papers corresponding to a bigram and FFA%, we calculated the Spearman's rank correlation between the rank of a bigram by number of papers and the rank of a bigram by FFA%. This correlation was found to be only 0.16. This correlation is not statistically significant at $p < 0.01$ (two-sided p-value = 0.2). Other experiments with a lower threshold for number of papers per title bigram (174 bigrams occurring in 50 or more papers and 1408 bigrams occurring in 10 or more papers) also resulted in very low and non-significant correlation numbers (0.11 and 0.03, respectively).

Discussion: Observe that FFA% varies substantially depending on the bigram. It is particularly low for title bigrams such as *dependency parsing, language models, finite state, context free,* and *neural models*; and markedly higher than average for *domain specific, semantic relations, dialogue system, spoken dialogue, document summarization,* and *language resources*. However, the rank correlation experiments show that there is *no* correlation between the popularity of an area (number of papers that have a bigram in the title) and the percentage of female first authors.

To obtain further insights, we also repeat some of the experiments described above for unigrams in paper titles. We found that FFA rates are relatively high in non-English European language research such as papers on Russian, Portuguese, French, and Italian. FFA rates are also relatively high for

[17] Other approaches such as clustering are also reasonable; however, results with those might not be easily reproducible.

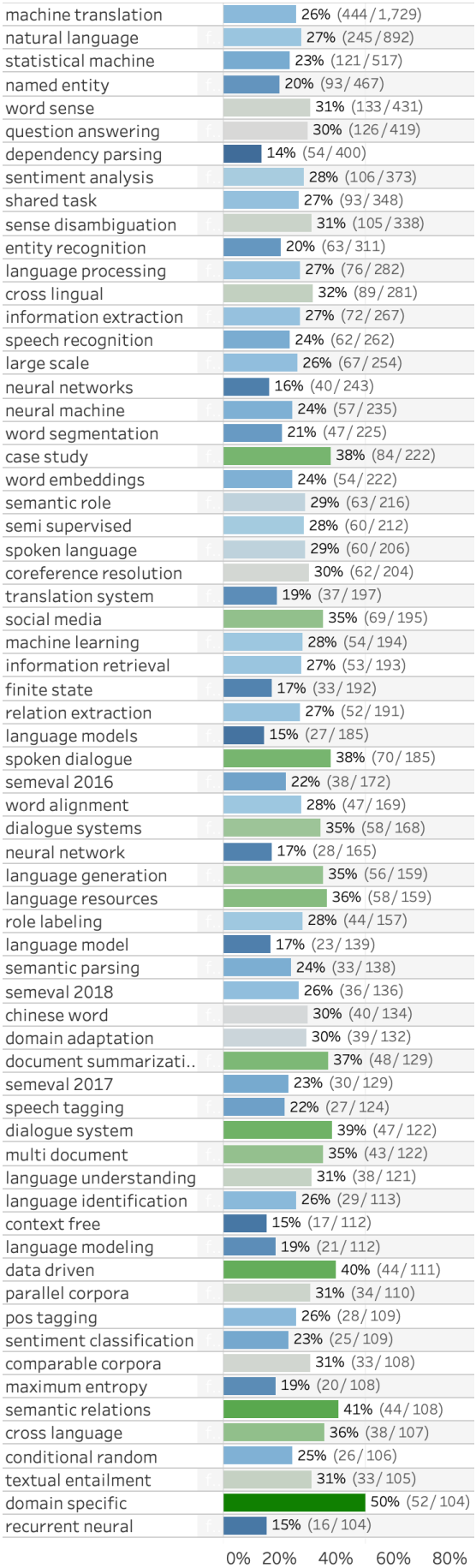

Figure 4: Top 66 bigrams in AA* titles and FFA%.

work on prosody, readability, discourse, dialogue, paraphrasing, and individual parts of speech such as adjectives and verbs. FFA rates are particularly low for papers on theoretical aspects of statistical modelling, and for terms such as *WMT, parsing, markov, recurrent*, and *discriminative*.

5 Gender Gap in Citations

Research articles can have impact in a number of ways—pushing the state of the art, answering crucial questions, finding practical solutions that directly help people, etc. However, individual measures of research impact are limited in scope—they measure only some kinds of contributions. The most commonly used metrics of research impact are derived from citations including: number of citations, average citations, h-index, and impact factor (Bornmann and Daniel, 2009). Despite their limitations, citation metrics have substantial impact on a researcher's scientific career; often through a combination of funding, the ability to attract talented students and collaborators, job prospects, and other opportunities in the wider research community. Thus, disparities in citations (citation gaps) across demographic attributes such as gender, race, and location have direct real-world adverse implications. This often also results in the demoralization of researchers and marginalization of their work—thus negatively impacting the whole field.

Therefore, we examine gender disparities in citations in NLP. We use a subset of the 32,985 AA′ papers (§3.2) that were published from 1965 to 2016 for the analysis (to allow for at least 2.5 years for the papers to collect citations). There are 26,949 such papers.

Q5. How well cited are women and men?

A. For all three classes (females, males, and gender unknown), Figure 5 shows: a bar graph of number of papers, a bar graph of total citations received, and box and whisker plots for citations received by individuals. The whiskers are at a distance of 1.5 times the inter-quartile length. Number of citations pertaining to key points such as 25th percentile, median, and 75th percentile are indicated on the left of the corresponding horizontal bars. The average number of citations are indicated with orange dashed lines.

Discussion: On average, female first author papers have received markedly fewer citations than male first author papers (37.6 compared to 50.4). The difference in median is smaller (11 compared

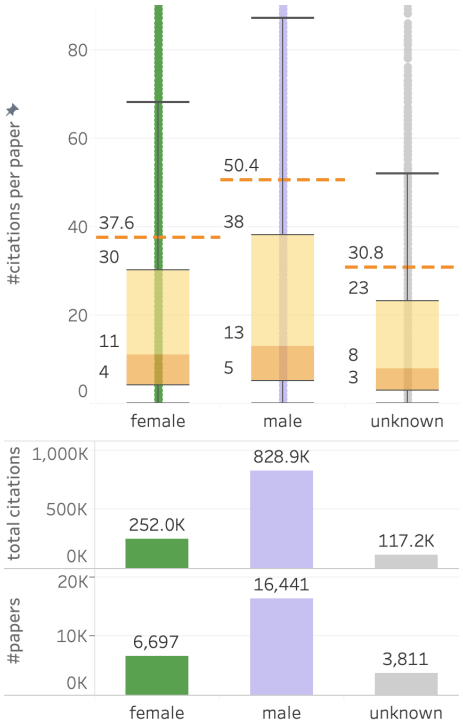

Figure 5: #papers, total citations, box plot of citations per paper: for female, male, gender-unknown first authors. The orange dashed lines mark averages.

to 13). The difference in the distributions of males and females is statistically significant (Kolmogorov–Smirnov test, $p < 0.01$).[18] The large difference in averages and smaller difference in medians suggests that there are markedly more very heavily cited male first-author papers than female first-author papers.

The differences in citations, or *citation gap*, across genders may itself vary: (1) by period of time; (2) due to confounding factors such as academic age and areas of research. We explore these next.

Q6. How has the citation gap across genders changed over the years?

A. Figure 6 (left side) shows the citation statistics across four time periods.

Discussion: Observe that female first authors have always been a minority in the history of ACL; however, on average, their papers from the early years (1965 to 1989) received a markedly higher number of citations than those of male first authors from the same period. We can see from the graph that this changed in the 1990s when male first-author papers obtained markedly more citations on average. The citation gap reduced considerably in the 2000s, and the 2010–2016 period saw a further reduction. It remains to be seen whether the citation gap for these 2010–2016 papers widens in the coming years.

It is also interesting to note that the gender-unknown category has almost bridged the gap with the male category in terms of average citations. Further, the proportion of the gender-unknown authors has steadily increased over the years—arguably, an indication of better representation of authors from around the world in recent years.[19]

Q7. How have citations varied by gender and academic age? Is the citation gap a side effect of a greater proportion of new-to-NLP female first authors than new-to-NLP male first authors?

A. Figure 6 (right side) shows citation statistics broken down by gender and academic age.

Discussion: The graphs show that female first authors consistently receive fewer citations than male first authors across the spans of their academic age. (The gap is highest at academic age 4 and lowest at academic age 7.) Thus, the citation gap is likely due to factors beyond differences in average academic age between men and women.

Q8. How prevalent is the citation gap across areas of research within NLP? Is the gap simply because more women work in areas that receive low numbers of citations (regardless of gender)?

A. On average, male first authors are cited more than female first authors in 54 of the 66 areas (82% of the areas) discussed earlier in Q4 and Figure 4.[20] Female first authors are cited more in the sets of papers whose titles have: *word sense, sentiment analysis, information extraction, neural networks, neural network, semeval 2016, language model, document summarization, multi document, spoken dialogue, dialogue systems,* and *speech tagging*.

If women chose to work in areas that happen to attract less citations by virtue of the area, then we would not expect to see citation gaps in so many areas. Recall also that we already showed that FFA% is not correlated with rank of popularity of an area (Q4). Thus it is much more likely that systemic biases and inequities, rather than the choice of area of research, are behind the gender gap.

[18] Kolmogorov–Smirnov (KS) test is a non-parametric test that can be applied to compare any two distributions without making assumptions about the nature of the distributions.

[19] The first-name based gender prediction method is expected to have a lower coverage of names from outside North America and Europe because USSNA and PUBMED databases historically have fewer names from there.

[20] The percentage is roughly the same even if one collapses morphologically related bigrams such as *neural network* and *neural networks* into one canonical form of the bigram.

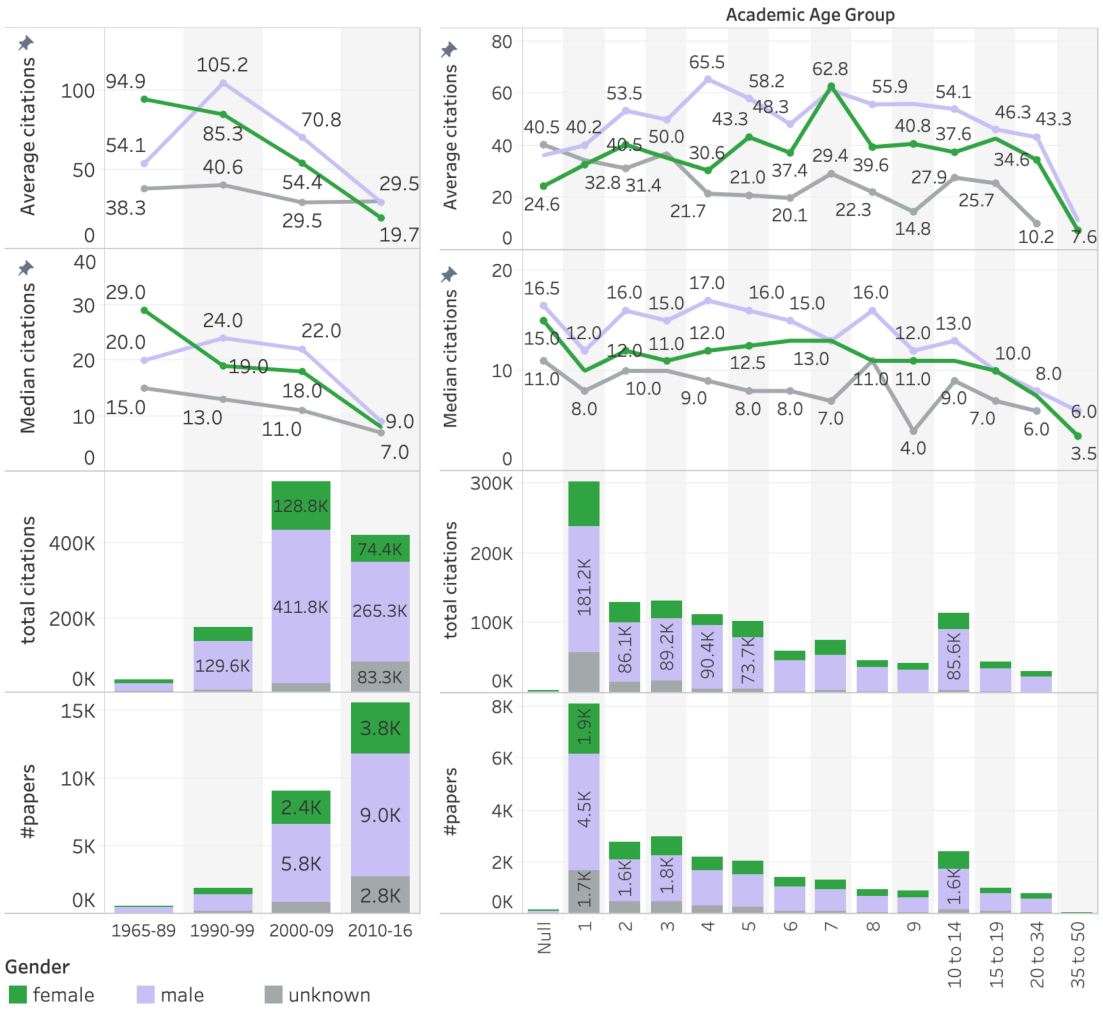

Figure 6: Citation gap across genders for papers: published in different time spans (left); by academic age (right).

6 Conclusions

We analyzed the ACL Anthology to show that only ∼30% have female authors, ∼29% have female first authors, and ∼25% have female last authors. Strikingly, even though some gains were made in the early years of NLP, overall FFA% has not improved since the mid 2000s. Even though there are some areas where FFA% is close to parity with male first authorship, most areas have a substantial gap in the numbers of male and female authorship. We found no correlation between popularity of research area and FFA%. We also showed how FFA% varied by paper type, venue, academic age, and area of research. We used citation counts extracted from Google Scholar to show that, on average, male first authors are cited markedly more than female first authors, even when controlling for experience and area of work. (Albeit, this citation gap is smaller for papers published in recent years.) Thus, in NLP, gender gaps exist both in authorship and citations.

This paper did not explore the reasons behind the gender gaps. However, the inequities that impact the number of women pursuing scientific research (Roos, 2008; Foschi, 2004; Buchmann, 2009) and biases that impact citation patterns unfairly (Brouns, 2007; Feller, 2004; Gupta et al., 2005) are well-documented. These factors play a substantial role in creating the gender gap, as opposed to differences in innate ability or differences in quality of work produced by the two genders. If anything, past research has shown that self-selection in the face of inequities and adversity leads to more competitive, capable, and confident cohorts (Nekby et al., 2008; Hardies et al., 2013).

Acknowledgments

Many thanks to Rebecca Knowles, Ellen Riloff, Tara Small, Isar Nejadgholi, Dan Jurafsky, Rada Mihalcea, Isabelle Augenstein, Eric Joanis, Michael Strube, Shubhanshu Mishra and Cyril Goutte for the tremendously helpful discussions.

References

Jens Peter Andersen and Mathias Wullum Nielsen. 2018. Google scholar and web of science: Examining gender differences in citation coverage across five scientific disciplines. *Journal of Informetrics*, 12(3):950–959.

Ashton Anderson, Dan McFarland, and Dan Jurafsky. 2012. Towards a computational history of the acl: 1980-2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 13–21.

Lutz Bornmann and Hans-Dieter Daniel. 2009. The state of h index research. *EMBO reports*, 10(1):2–6.

Margo Brouns. 2007. The making of excellence–gender bias in academia. *Wissenschaftsrat (Hrsg.)*.

Claudia Buchmann. 2009. Gender inequalities in the transition to college. *Teachers College Record*.

Jordi Duch, Xiao Han T Zeng, Marta Sales-Pardo, Filippo Radicchi, Shayna Otis, Teresa K Woodruff, and Luís A Nunes Amaral. 2012. The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PloS one*, 7(12):e51332.

Irwin Feller. 2004. Measurement of scientific performance and gender bias. *Gender and Excellence in the Making*, 35.

Marta Foschi. 2004. Blocking the use of gender-based double standards for competence. *Gender and Excellence in the Making*, pages 51–56.

Juan Miguel Gallego and Luis H Gutiérrez. 2018. An integrated analysis of the impact of gender diversity on innovation and productivity in manufacturing firms. Technical report, Inter-American Development Bank.

Gita Ghiasi, Vincent Larivière, and Cassidy Sugimoto. 2016. Gender differences in synchronous and diachronous self-citations. In *21st International Conference on Science and Technology Indicators-STI 2016. Book of Proceedings*.

Namrata Gupta, Carol Kemelgor, Stefan Fuchs, and Henry Etzkowitz. 2005. Triple burden on women in science: A cross-cultural analysis. *Current science*, pages 1382–1386.

Michael Gusenbauer. 2019. Google scholar to overshadow them all? comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 118(1):177–214.

Malin Håkanson. 2005. The impact of gender on citations: An analysis of college & research libraries, journal of academic librarianship, and library quarterly. *College & Research Libraries*, 66(4):312–323.

Dalia S Hakura, Mumtaz Hussain, Monique Newiak, Vimal Thakoor, and Fan Yang. 2016. *Inequality, gender gaps and economic growth: Comparative evidence for sub-Saharan Africa*. International Monetary Fund.

Kris Hardies, Diane Breesch, and Joël Branson. 2013. Gender differences in overconfidence and risk taking: Do self-selection and socialization matter? *Economics Letters*, 118(3):442–444.

Madian Khabsa and C Lee Giles. 2014. The number of scholarly documents on the public web. *PloS one*, 9(5):e93949.

Molly M King, Carl T Bergstrom, Shelley J Correll, Jennifer Jacquet, and Jevin D West. 2017. Men set their own cites high: Gender and self-citation across fields and over time. *Socius*, 3:2378023117738903.

Vincent Larivière, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R Sugimoto. 2013. Bibliometrics: Global gender disparities in science. *Nature News*, 504(7479):211.

The Linguistic Society of America LSA. 2017. The state of linguistics in higher education annual report 2017.

Alberto Martín-Martín, Enrique Orduna-Malea, Mike Thelwall, and Emilio Delgado López-Cózar. 2018. Google scholar, web of science, and scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12(4):1160–1177.

Sangeeta Mehta, Karen EA Burns, Flavia R Machado, Alison E Fox-Robichaud, Deborah J Cook, Carolyn S Calfee, Lorraine B Ware, Ellen L Burnham, Niranjan Kissoon, John C Marshall, et al. 2017. Gender parity in critical care medicine. *American journal of respiratory and critical care medicine*, 196(4):425–429.

John Mingers and Loet Leydesdorff. 2015. A review of theory and practice in scientometrics. *European journal of operational research*, 246(1):1–19.

Shubhanshu Mishra, Brent D Fegley, Jana Diesner, and Vetle I Torvik. 2018. Self-citation is the hallmark of productive authors, of any gender. *PloS one*, 13(9):e0195773.

Saif M. Mohammad. 2019. The state of nlp literature: A diachronic analysis of the acl anthology. *arXiv preprint arXiv:1911.03562*.

Saif M. Mohammad. 2020a. Examining citations of natural language processing literature. In *Proceedings of the 2020 Annual Conference of the Association for Computational Linguistics*, Seattle, USA.

Saif M. Mohammad. 2020b. Nlp scholar: A dataset for examining the state of nlp research. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC-2020)*, Marseille, France.

Saif M. Mohammad. 2020c. Nlp scholar: An interactive visual explorer for natural language processing literature. In *Proceedings of the 2020 Annual Conference of the Association for Computational Linguistics*, Seattle, USA.

Lena Nekby, Peter Skogman Thoursie, and Lars Vahtrik. 2008. Gender and self-selection into a competitive environment: Are women more overconfident than men? *Economics Letters*, 100(3):405–407.

Enrique Orduña-Malea, Juan Manuel Ayllón, Alberto Martín-Martín, and Emilio Delgado López-Cózar. 2014. About the size of google scholar: playing the numbers. *arXiv preprint arXiv:1407.6239*.

Kathyayini Rao and Carol Tilt. 2016. Board composition and corporate social responsibility: The role of diversity, gender, strategy and decision making. *Journal of Business Ethics*, 138(2):327–347.

Patricia A Roos. 2008. Together but unequal: Combating gender inequity in the academy. *Journal of Workplace Rights*, 13(2).

Natalie Schluter. 2018. The glass ceiling in NLP. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2793–2798.

Inger Skjelsboek and Dan Smith. 2001. *Gender, peace and conflict*. Sage.

Brittany N Smith, Mamta Singh, and Vetle I Torvik. 2013. A search engine approach to estimating temporal changes in gender orientation of first names. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 199–208. ACM.

Vetle I Torvik and Neil R Smalheiser. 2009. Author name disambiguation in medline. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(3):1–29.

Adam Vogel and Dan Jurafsky. 2012. He said, she said: Gender in the acl anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 33–41.

World Economic Forum WEC. 2018. The global gender gap report 2018.

Cassandra Willyard. 2011. Men: A growing minority. *GradPSYCH Magazine*, 9(1):40.

Jonathan Woetzel et al. 2015. The power of parity: How advancing women's equality can add $12 trillion to global growth. Technical report.