Why is AI "a sea of dudes"? Using data science and NLP methods to understand gender imbalance in a scientific community.

Ramona Comanescu

4th Year Project Report
Computer Science
School of Informatics
University of Edinburgh

2018

Abstract

This dissertation carries an in-depth study of gender in the field of Computation Linguistics. Our approach relies heavily on information that we extract directly from the data, using tools that the very field we are investigating promotes.

We perform gender attribution on the authors present in a corpus and investigate new gender classification methods, including character-level LSTMs and face recognition.

We then perform a quantitative analysis the publication patterns of these authors, focusing on career development over time, collaboration through coautorship and conference rankings. Most of our results are statistically significant and help paint the landscape of the field. We find that women are underrepresented in the last author position. What is more, men have a higher number of active years in the field and a higher number of publications per active years. In terms of collaboration, females tend to coauthor more papers with other female authors. Another concerning finding is that women are underrepresented at the highest ranked conferences.

We employ topic modeling to capture how the shift in the field of Computation Linguistics affects the gender gap and contrast this with earlier findings. We report significant differences in the topics that each gender is more likely to choose. Finally, we look at the effect of an online publishing repository (arXiv), as opposed to a traditional corpus(ACL).

Our analysis suggests that there are subtle ways in which gender differences can occur in scholarly authorship and practitioners should be aware of the dangers of any unconscious gender bias.

Acknowledgements

I would like to extend my deepest appreciation to my supervisor, Dr.Adam Lopez for his invaluable time and guidance during my project.
Many thanks to my parents for their immense support and for the unexpected help with manually labelling a list of over 2000 names of unknown gender.
Francisco Vargas for the insightful brainstorming, especially regarding name classification algorithms and for helping me understand details of LDA. Cynthia Yu for insight into Chinese name classification.
Thank you to everyone who took the time to proofread this thesis. (Francisco Vargas, Diana Cremarenco, Clara Vania).
To my friends Elitsa Bankova, Stilyan Ivanov and Andreea Cucu for all their support.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

Artificial intelligence (AI) has been growing rapidly in recent years. However, just 13% of one of 2015's biggest AI conferences (NIPS) were women [Bloomberg, 2016].

Microsoft researcher Margaret Mitchell calls Artificial Intelligence "a sea of dudes" [Bloomberg, 2016] and this situation is definitely concerning. If you train a speech recognition system using British English, it will not generalize well when you test it on American English. Similarly, an unbalanced research environment means that it will lack a diverse pool of ideas and the algorithms that are created will not cover all users.

There are some famous examples of how biased or incomplete datasets create undesirable results: Tay, Microsoft chat-bot learned from users' tweets to be racist [BBC, 2016], while Google mistakenly tagged black people as gorillas[Guardian, 2015]. Machine learning can amplify biases already present in the data and researchers are trying to investigate this. Bolukbasi et al. [2016] show that continuous vector space representation of word embeddings trained on Google News articles learn the hidden sexism in language, resulting in interesting vector space arithmetic:

$$\overrightarrow{man} - \overrightarrow{woman} \approx \overrightarrow{computerprogrammer} - \overrightarrow{housemaker}$$

What is more, bias is not only about training data, machine learning systems also suffer if the designers do not come from diverse backgrounds, specifically if they do not account for gender. We aim to investigate the gender from the perspective of those shaping the future of AI.

Most studies of gender balance in Computer Science are based on reporting statistics on enrolment. Vogel & Jurafsky [2012] go beyond the surface and study the difference in topics that authors of each gender write about. Their findings report that the number of female authors has been increasing between 1980 and 2008. However, due to the rise of deep learning methods, we believe that the shift of computational linguistics has changed and it could have an effect on gender balance. The presidential keynote of the 2015 annual meeting of the Association for Computational Linguistics [Manning, 2015]

outlines ways in which the field has been changing: deep learning giants like Geoffrey Hinton, Yoshua Bengio and Yann LeCun have been increasingly focusing their research towards language, with the rise of large NLP datasets. Given than the gender imbalance in deep learning is worse than other Computer Science fields (according to the gender composition of NIPS, for example), this could have an effect on the Computational Linguistics community.

We start from the idea that despite increasing number of women graduating in Computer Science and Mathematics, there is still a systemic gender imbalance in the distribution of active researchers and professors, where women are underrepresented. Given the recent growth of the Computational Linguistics community, there is need for a comprehensive study on the matter of gender imbalance.

1.2 Outline

The main goal of this thesis is to analyse gender imbalance in the field of Computational Linguistics. We go beyond surface statistics and try to understand how gender affects academia, from the output of research to the career paths of individuals. We formulate different tasks and use data science and NLP methods to understand them.

Description of Corpus and preprocessing steps: Chapter 3

We conduct our main experiments on the ACL Anthology Reference Corpus, presented in more detail in Section 3.1. We also introduce the arXiv corpus, presented in Section 3.2. We introduce methods of detecting ill-formatted documents, including language identification and Optical Character Recognition.

Authorial Gender Attribution: Chapter 4

In order to perform our analysis, we need to attribute gender to all the authors. We present some ethical issues related to treating gender as a variable in our analysis. We then proceed to attribute gender to the authors of the papers in our corpus. We present several approaches that can be used for gender classification, including population database statistics, machine learning methods and manually curated lists. Our methods include face recognition and character level Long Short-Term Memory networks. Finally, we develop our own classification pipeline and we evaluate it on a gold test set to show that it meets the high precision requirements of the task.

Publishing Patterns in the ACL corpus: Chapter 5

The first question we try to answer is: Is there a difference in the publication patterns of men and women? We state that for securing permanent positions, aspiring professors need a good publication record. Analysing publication patterns allows us to understand whether there are any differences between how one's career progresses depending on gender. Do women tend to publish less frequently and with fewer collaborators? How often are they listed first in the list of authors and how long does it take for an author to be listed in the prestigious position of last author? We use the collection of papers released by the ACL and we break down our analysis in the following points:

Overall statistics: In Section 5.2, we look at the time evolution of the number of active authors and the number of publications. We find little progress in closing the gender gap.

Productivity: In Section 5.3, we investigate the differences in productivity between men and women, considering authors with similar counts of publications but different output per year. We find that men publish more papers per active year.

Position in author list: In Section 5.4, we use the position in the list of authors as an indicative of the stage of their career. It is generally accepted that first author represents the person primarily responsible for the paper, while last author indicates the senior scientist who supervised the work. We find that women are underrepresented as last authors, but the percentage of women in first author position has been increasing over the years.

Cohort analysis: In Section 5.5, we consider people who publish at the same time as being part of the same cohort and look at how many are still publishing after 5 or 10 years. Based on this data, we can measure dropout rates and analyse "the productivity gap". We find that women have a higher dropout rate than men. They also publish slightly less at the beginning of their career.

Coauthorship: In Section 5.6, investigate the size of the collaboration networks for men and women and the gender demographics of the people they tend to publish with. We find that each gender tends to have more coauthorship relationships of the same gender. Women also have less single-authored papers.

Publishing at highest ranked venues: In Section 5.7, we investigate the percentage of females publishing at what are considered the highest ranked Computation Linguistics conferences. We find that women have a lower percentage at these conferences than their overall presence in the dataset.

Topic modeling: Chapter 6

The rest of our study is concerned with the actual document content and its relationship with the gender of the authors.

Distribution of research topics: We use topic models to study the difference in the topics that men and women write about. For topic modeling, we employ Latent Dirichlet Allocation(LDA) [Blei et al., 2003]. Section 6.1 gives a background description of the inference process and highlights some key distributions that are useful for the rest of the experiments. The main questions we are trying to answer are related to the distribution of the topics authors publish in. Is there a higher tendency for a particular gender to publish in a particular topic? In Section 6.4.4, we first look at the gender preferences for topics. In Section 6.4.5 we will also consider the distribution of topics at most important venues in Computational Linguistics.

Publishing in an online repository: Chapter 7

The last part of our analysis focuses on a more "non-conventional" corpus: papers on Computational Linguistics included in the online repository arXiv. This analysis yields a very exciting view of the field of Computational Linguistics, as it contains

publications up to 2018, compared with the ACL collection, which contains metadata only up to 2014. The dynamics of this community will be fundamentally different. The arXiv corpus is different to ACL in terms of composition, as publishers do not need to have been accepted to a conference in order to submit their papers, but can instead self-archive. We research other possible differences (such as the absence of double blind reviews) that an online corpus brings. We will also look at the document content of this papers and the topics that emerge, for a more recent view on the trends in computational linguistics, as well as its relation to the authors' genders.

1.3 Summary of contributions

The main contributions of this thesis are:

- Data collection
    1. Cleaning up ACL corpus published by Radev et al. [2013] (removing duplicates, identifying malformed documents through language identificatin, using optical character recognition to extract text form PDF files for cases when automatic conversion has failed, handling escape characters, handling missing first names)
    2. Scraping, parsing and cleaning up the arXiv dataset on Computational Linguistics
- Name classification
    1. Surveying name classification methods
    2. Developing a multi-stage algorithm that combines: available name lists, databases of population statistics for unambiguous cases, handling variations of the same person's name, combining various APIs
    3. Proposing and testing of a new method based on classifying faces returned by search engine results
    4. Training a Long Short-Term Memory network for name classification
    5. Producing an updated list of authors in Computation Linguistics
- Publishing patterns - designing experiments, analysis and visualisations:
    1. Authorship by year: overall statistics
    2. Career development over time: investigate rate of publishing, productivity gap, earning senior positions (single-authored papers, first and last position in the list of authors)
    3. Coauthorship: investigated the number of collaborators of men and women
    4. Conferences: investigated the gender balance at highest ranked conferences

5. Visualisations: Used to illustrate and summarize career patterns of a large population

- Topic modeling:
  1. Preprocessing the data into bag of words, using various tokenizers, lemmatizers, named entity recognition tools
  2. Constructing LDA models for both ACL and arXiv corpus and comparing them based on perplexity, coherence and visualisations
  3. Labelling ACL topics both manually and using results from similar studies
  4. Labelling arXiv topics by finding closest correspondence in the ACL corpus, based on KL divergence
  5. Analysing the evolution of topics over time
  6. Ranking topics that are more likely to be published by males compared to females
  7. Ranking topics at different NLP conferences
- Overall implementation:
  All investigations that we performed are designed as Python modules and are corpus agnostic, once the corpus has the required comma separated values format(title, year, authors, venue, content). This makes it straightforward to carry a similar study on different datasets.

Chapter 2

Background on gender studies in academia

In this chapter we are concerned with the overall picture of gender in academia. We present some overall statistics related to enrolment in Computer Science and testimonials of women in the field. We then refer to studies of gender across various fields and their methodologies and findings. We make a clear exposition on how work from previous studies was reused.

2.1 Enrolment statistics

Providing comprehensive statistics of gender presence and employment in academia is not trivial, due to the scattered data by country and institution. There are, however, several surveys that break down the number of undergraduate and postgraduate students.

In the United States, the number of females graduating with a degree in Computer Science has been decreasing, from more than 35% in 1985 down to 18% in 2014, according to a report by the National Science Foundation [2017] . When in comes to the number of awarded PhDs, the Computing Research Association (CRA) Taulbee Surveys [CRA, 2016], report that women were awarded only 18.5% of the PhDs in 2016, compared to 18% in 2002 and 20.5% in 2007.

In the UK, UCAS [UCAS, 2014] statistics report that in 2014, only 13% of new entrants were female. The University of Edinburgh consistently reports percentages above the national average, with 19.5% in the 2014 undergraduate cohort and 21.4% in the postgraduate research cohort, according to an Athena Swan report. [Athena Swan Award, 2016]

When it comes to academic careers in higher education, a report released by the American Association of University Professors [AAUP, 2006] finds that women represent 25% of professors and earn 80% of the salary of men in similar positions.

An even more concerning observation is the situation in academia. The Report prepared

by Laboratory for Computer Science and the Artificial Intelligence Laboratory at MIT [1983] offers an interesting perspective gathered from the daily experiences of women in academia. Feeling intimidated when entering a room full of men, being ignored or interrupted, misplaced expectations of how women should behave or having to hide their family life to fulfill long hour expectations are all outlined in the report. Their main recommendation is that closing the gender gap would also diminish the number of uncomfortable situations that it generates.

There is also the problem of unconscious biases that both women and men exhibit. One empirical example is given by Moss-Racusin et al. [2012]: the same applicant for a lab manager position has been given to science faculty to rate. Their finding was that faculty gave a higher score when the applicant was assigned a male name. This study outlines the role of gender in hiring decisions.

What is clear from all the reports is that there exists an undeniable gender gap in certain fields of academia and this can have an impact on the quality and diversity of research output. The problem starts early in the undergraduate enrolment process and propagates further on.

2.2 Existing studies of gender in academia

There are several studies of gender in academia, across different fields (usually STEM). We list a few of the relevant ones. The methods used across them tend to be similar, analysing aspects such as position in the list of authors, distribution of topics, productivity. A challenge across all studies is the difficulty of attributing gender to authors. Most of the time, the method used is based on statistics of first name popularity by gender.

JSTOR Corpus

West et al. [2013] conduct a large scale study on papers across natural sciences, humanities and social sciences, using over eight million papers from JSTOR [1]. They argue that while metrics such as "grant funding, hiring, productivity" seem to indicate gender inequality will soon be solved, there are still certain aspects where close inspection reveals inequality. Men tend to predominate in prestigious first and last author positions, while women also have less single authored papers. Their findings are strengthened by the fact that in some disciplines, such as Life Sciences or Social Sciences, authors have similar raw publication counts, yet these disparities still arise. In their findings, Computer sciences have the lowest percentage of PhD, between 16 and 21% in the 2000s, compared to the overall rates of 38-40% across all disciplines.

Mathematics

Mihaljević-Brandt et al. [2016] claim that securing permanent positions requires a large number of publications and patents, as well as publishing in reputed journals and having a high citation rate. They analyse mathematical papers over the past four decades and find significant differences between genders, which might hinder the academic career of

[1] https://www.jstor.org/

women in mathematics. One concerning finding is that women abandon their academic careers within 10 years at a higher rate than men (33% female dropout rate compared to 27% for males). They also find that women publish less single-authors papers and have a narrower distribution of research topics. Women also tend to publish in lower impact journals.

Computational Linguistics

Most studies of gender balance in Computer Science are based on reporting statistics on enrolment. Vogel & Jurafsky [2012] go beyond the surface and study the difference in topics that different genders write about. They perform a "mostly-manual annotation" of the gender of each author, which we also use as a starting point for our name classification. They use Latent Dirichlet Allocation(LDA) topic models to study the difference in the topics that genders write about. Their study is conducted on the Association of Computational Linguistics Anthology Network (AAN) corpus [Radev et al., 2013], using papers from 1965 to 2008. They find that the participation of females in the field has been steadily increasing: the percentage for female authorship grew from 13% in 1980 to 25% in 2008. They also find that "women publish more on dialogue, discourse, and sentiment, while men publish more than women in parsing, formal semantics and finite state models". We aim to update this study, by conducting a similar study up to the year of 2014, which is the latest release of the AAN corpus.

2.3 Using topic models to understand a corpus

A topic model is a statistical model for discovering "topics" in a collection of documents, in an unsupervised manner. A commonly used topic model is Latent Dirichlet Allocation, introduced by Blei et al. [2003]. In LDA, each document is assumed to be generated by a generative process. Documents are therefore treated as multinomial mixtures of latent topics, while topics are multinomial distributions over words. The topic mixture for a document is sampled according to a Dirichlet distribution over a set of topics. We describe LDA in more details in Section 6.1.

There are extensions to this model that approach the problem slightly differently. For example, *Hierarchical Dirichlet Processes* [Teh et al., 2006] treats the number of topics as a random variable generated from a Dirichlet Process, useful when the number of topics is not known.

Topic models allow efficient processing of large collections of data. There have been multiple studies based on research in academia, thanks to initiatives that maintain large collection of papers.

An important contribution is brought by Rosen-Zvi et al. [2012], who introduce the *Author-Topic* model, an extension of LDA to include author information. Here, each author is associated with a multinomial distribution over topics. They experiment with papers from NIPS conferences and find authors with the highest probability conditioned on topics, offering an interesting method for observing individual research careers.

In 2012, the Association of Computational Linguistics held an workshop on "Rediscovering 50 Years of Discoveries". This encouraged a thorough research of the evolution of the field. In particular, Anderson et al. [2012] use topic modeling to put together a history of the field of computation linguistics, from 1980 to 2008, analysing different research epochs and their focus, as well as the effect of government-sponsored initiatives and the flow of authors across topics.

Blei & Lafferty [2006] incorporate gradual evolution of topics over time, with the *Dynamic Topic* model. However, Hall et al. [2008] find that the Dynamic Topic model penalises large changes from year to year and prefer to simply use the probability of each topic given the year. They apply this to the ACL corpus, looking for historical trends in the field. Their results show an unsurprising raise of topics like Probabilistic Models, Classification and Tagging.

Recently, there has also been exciting research on *Topic2Vec* models, that aim to leverage the idea of learning distributed representations along with word representations. In one approach [Niu & Dai, 2015], the authors find that their results in more distinguishable results between similar topics, while the returned topics are also more representative. Their approach still requires running LDA first and incorporating them in the learning objective function. There are many other results that aim to use word embeddings together with LDA. [Das et al., 2015; Liu et al., 2015; Shi et al., 2017] The surge of interest regarding LDA and word embedding is motivated by the fact that they can enhance each other. LDA offers word embeddings a more global view, helping with disambiguation. On the other hand, LDA on its own is prone to select the most frequent words and might miss others. For this reason, Hall et al. [2008] need to manually set seed words to discover all the relevant topics in their study.

2.4 Use of existing work

One of the initial aims for this dissertation is to update the study made by Vogel & Jurafsky [2012], incorporating new ACL data from 2008 to 2014, which is the latest release of the corpus at the time of writing this thesis. Vogel & Jurafsky made some of the outputs of the study available [2]:

- Author names from the ACL, labelled with gender: A number of 11931 of names have been provided. Our extended dataset however contains 18155 authors, so name lists have been updated. We used similar methods to those used by the original authors(census lists of frequency, morphological gender). We were not able to use personal cognisance of the ACL authors or ask other researchers for help. Instead, we came up with additional ways of name classification, as explained in Chapter 4.
- Labelled topic terms: The labelled topic models are also available, so we were able to compare our results for a sanity check. However, the original authors

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

use the Stanford Topic Modeling Toolbox [3] which is not maintained anymore. We explore additional Topic Modeling libraries and their advantages. They also do not mention any of the parameters they used, so we perform a significant amount of experimentation on this matter, including size of the corpus, number of iterations, Dirichlet priors, as detailed in Section 6.4.

[3] https://nlp.stanford.edu/software/tmt/tmt-0.4/

Chapter 3

Corpus

3.1 ACL Anthology Network Corpus

The Association for Computational Linguistics(ACL) is an international society for "people working on computational problems involving human language". [1]. One of their great initiatives is maintaining the ACL Anthology [2], which archives publications in Computational Linguistics from 1974 to present.

However, the anthology in its initial form was just a collection of publications with no metadata, which meant it was hard to analyse. Radev et al. [2013] preprocess the papers in the ACL Anthology so that in contains useful information such as the venue and year of publication, as well as the authors. All the extracted information was put together as a database file. They released this metadata together with the associates documents as the ACL Anthology Network Corpus (AAN). The project involved mainly manual work so that the results are as clean and reliable as possible. As maintaining the corpus is a challenging effort, it is only updated every few years, with the latest release being 2014. The statistics of this release report 24627 publications and 18862 authors.

Given the fact that our project requires both document content and information about the authors involved, the AAN corpus provides a good starting point. The ACL 2012 Contributed Task [Schäfer et al., 2012] outlines some of the problems that make maintaining an up to date and error-free ACL collection difficult. Many of the files in the ACL Anthology were not born digital and had to be scanned, which makes the extraction of document content difficult.

We make our own attempt at cleaning and preprocessing the data, before doing any analysis. First, we intersect the ids in the metadata files with the publications we actually have the text content available. We remove all duplicate ids (papers published in multiple publications) and we are left with 23766 papers out of 24627 and 18158 unique authors.

[1] http://aclweb.org/

[2] http://aclweb.org/anthology/

Preprocessing documents

We discover files for which automatic conversion from PDF format has failed, outputting random characters. We discard any publications with less that 200 characters.

In order to detect malfunctioning conversion, we employ automatic language detection, using *langid.py* [Lui & Baldwin, 2011], which is a language identification system trained on over 97 languages. This allows us to discard any publications that do not have a high confidence score for English. This process discards both malformed documents, as well as 10 papers which where written in French. We need to discard these files because we will later on build topic models using the content of these documents. This accounts to around 600 malformed documents.

For empty files, it is usually the case that conversion has failed. We make an attempt at recovering this data using Optical character recognition(OCR) on the input files. OCR is a method of converting images of text into machine-encoded text. Modern software is based on pattern matching techniques and can achieve high quality results. After passing corrupted files through Tesseract [Smith, 2007], we manage to recover 15 publications.

Preprocessing author names

Our analysis relies heavily on the names of authors, which need to be as consistent as possible, in order to identify the relation between an author and their research output correctly. This means care has to be taken to list the names in a consistent manner. One cause of inconsistencies is the presence of HTML escape characters. Since Pierré and Pierr&eacute(which corresponds to the HTML escape character) refer to the same person, these cases are handled programatically so that both names become Pierré.

Its is often the case that an author sometimes uses just an initial instead of his first name, which also introduces the problem of counting the same person twice in our overall analysis. We implement a best match algorithm. We group all names by last names: [Doe: [M.J., Mary Jane, John, Anna]]. We then try to match the initials with a first name. If there are more first names with correspond with the same initial, we check if any of the names match on more than one initial and give priority to those. Most names have more than one initial, so we are confident about those predictions. For example, "Doe, M.J." will match with "Doe, Mary Jane". The grouping by last names also helps with matching Pierré and Pierre. In this way, we match 147 names.

Resulting format

After preprocessing, the input of our analysis contains a collection of files with the following metadata:

id	title	authors	year	venue	content

3.2 ArXiv Corpus

In Chapter 7 we will analyse the content of a different corpus: arXiv [3], which is an online repository of scientific publications in different fields. The main difference to other datasets, such as ACL, is that the publications can be self-archived and published after moderation, with no need to be accepted at a conference. Since the repository was stared in 1991, all papers present are born-digital, which means the document content is readily available.
Given the online nature of the dataset, authors are required to provide metadata upon submission. We therefore have easy access to the year of publication, authors, title and abstract, available in a clean format, with no further preprocessing being required. The only step that we undertake is matching any initials with first names, where possible, using the same approach as for the ACL corpus. We build a file with the same fields as for ACL, except there is no venue field.

Acquiring the metadata

We aim to obtain all metadata in the Computational Linguistics subsection of arXiv. Since arXiv is a registered Open Archives Initiative Protocol for Metadata Harvesting[4] dataprovider, we can obtain the metadata programatically through HTTP requests. We harvest all files from 1991 to February 2018. (7934 publications)

[3] https://arxiv.org/
[4] http://www.openarchives.org/pmh/

Chapter 4

Name classification

Gender information is often used for novel sociological research, allowing a better understanding of the interactions present in society. We will use gender as a variable for discovering patterns in the data.

4.1 Task

We aim to label all author names in the dataset as either male, female or unknown. We classify authors both in the ACL and arXiv corpus. The rest of this chapter deals with various issues regarding name classification and our approach to solve them.

Disclaimer

Using gender as a variable in research can raise ethical concerns. We are only assigning binary genders for the purpose of conducting a study on the demography of the researchers. In the following section, we aim to present the way in which we assign gender very clearly. In cases where it is not possible to assign gender, we label the respective name as "unknown". Our main approach is to assign gender based on names, as supported by census statistics, based on the assumption that certain names are generally associated with a certain gender. When dealing with unisex names, we had to resort to either automatic face identification or manually inspecting search engine results and looking for hints such as pronouns (he, she) or images. DeFrancisco et al. [2013] sees gender as performative, consisting of "the behaviours and appearances society dictates a body of a particular sex should perform" and this is the main interpretation of the word that we use.

Challenges

Assigning gender to names is a non-trivial task. Most literature on name-gender attribution uses large databases of first name statistics. However, in the case of researchers, we are dealing with an international collection of names which introduces some additional issues. Karimi et al. [2016] find that most automatic name inference methods are biased towards countries of origin. For example, we would not be able to just use a

USA database of first names. *Jan* is a common American female nickname, while it is a typical male name in parts of Europe. The name *Andrea* will be used for men in Italy, while it is usually a girl's name in USA. It becomes clear that the gender most commonly attributed to a first name sometimes depends greatly on the country of origin. A crucial issue that we face is that some names are unisex (*Alex, Jessie, Jamie*) - in this case, any statistical method will fail, as long as the only available information we have is the name.
As we only have the Latin version of the names, this generates problems for names in Chinese and Japanese. Gender could be easily identified by a native speaker in their original version. However, some of the names could be generated by different characters combinations, each giving it a different meaning and a different gender. For example, the Chinese name ZiXuan has a female version 紫萱(floral, elegant) , while the male version is 子轩(knowledgeable, tolerant) [1] . This makes it impossible to correctly classify based on the name form only.

We tackle this challenges by coming up with methods that use more than the first name, relying on search engines results in either a manual or algorithmic manner, which will be described in detail in this chapter.

4.2 Basic approaches

Database lookup

A common practice among researchers using gender in their research is to assign labels based on first names and their frequencies in various databases.(Tang et al. [2011] collect a list of Facebook names from the New York area and they assign genders based on frequency in these name lists. It is also common to crawl profiles on social media websites and crowdsource annotation based on profile pictures: Liu & Ruths [2013] do this for Twitter.

First name based classification can be successfully performed on unambiguous names. All names in the dataset are in the form "Last name, First name", therefore it is straightforward to obtain the first names. We lookup first names in the following databases that we identified as being released by trustworthy bodies:

- US Census from 1990 [2]
- Jörg Michael manually curated database, with names from different countries [3]
- List of Indian names [4]

If a first name is found in one of the database with only one gender assigned to it, we consider it unambiguous and assign it that gender. About 70% of the ACL names were classified in this way.

[1] https://baike.baidu.com/
[2] https://www2.census.gov/topics/genealogy/
[3] Source: File "nam dict.txt" from https://heise.de/ct, softlink 071718
[4] https://gist.github.com/mbejda/9b93c7545c9dd93060bd

Morphological Gender

For languages like Czech, Bulgarian and Russian, gender is marked on last names. For example, the female version of the name *Anisimov* would be *Anisimova*. We classify 261 ACL authors with a last name ending in "ova" or "ov" accordingly.

APIs

GPeters[5] baby name classifier is used by many name classification tasks (Vogel & Jurafsky [2012], Mihaljević-Brandt et al. [2016]) . It works by crawling internet results and reporting how frequent a name is associated with each gender. We have set a threshold of 2 for the male/female ratio in order to trust a result.

NamSor API [6] analyses both ethnicity and gender based on social media profiles over the Internet. It returns a probability score, therefore we chose to only handle those cases that have a score of over 80%. The main advantage is that it uses both first and last names, as it accounts for ethnicity. It therefore manages to classify well cases such as Jan or Andrea, which depend on the country of origin.

Manual annotation

When none of the above methods can classify a name, we resort to manual classification. If the gender of a name is not obvious to human inspection, we use search engine results such as the personal page of the corresponding name. If there are any pictures or useful pronouns ("*she* is known for.."), we can assign a gender. We try to limit this method by improving algorithmic approaches as much as possible, as it it very time consuming and it can introduce bias.

4.3 Advanced approaches

4.3.1 Deep learning approach: Character Level LSTM

The methods outlined above require a lot of effort for gathering and combining different databases, deciding on thresholds for trusting the results. They also rely on external APIs and libraries. Therefore, we would like to be able to build a system with less dependencies. An obvious answer is building a machine learning system that would automatically learn from a large collection of names.

We propose a deep learning approach to the name classification task to remove the need for hand crafted features and external APIs. One of the main arguments in favour of this approach is that we would be able to classify based on both and first and last name, which will bring ethnic information able to handle ambiguous cases where gender depends on the country of origin.

[Lee et al., 2017] propose a Recurrent Neural Network(RNN) approach to nationality classification based on names. We believe a similar approach would work for gender

[5] https://www.gpeters.com/names
[6] http://www.namsor.com/

classification. Since this task involves learning from sequences (of characters), RNNs are a good fit.

Recurrent Neural Networks and variants have been proven very successful for many tasks including text classification and are known for their ability to model invariance across time. Recent advancements propose variants of RNNs that do not suffer from the problem of vanishing gradients and are able to keep track of longer dependencies: Long Short-Term Memory(LSTM)[Hochreiter & Schmidhuber, 1997] networks are now often used.

The main challenge of this approach is finding a representative corpus, with international names (containing both first and last name), labelled by gender. Due to the sensitivity of gender information in data, such information is often not readily available. For our experiments, we use a database retrieved from US public inmate records. [7] The nature of the dataset could introduce a potential nationality bias, but since we test on the ACL dataset, which has an international component, we believe the testing methodology alleviates the problems of the training dataset. After balancing the dataset to have equal training data for each gender, our training data consists of 14720 names. Since there are less females in the training set, we deal with class imbalance by performing undersampling, each category matching the size of the smallest one.

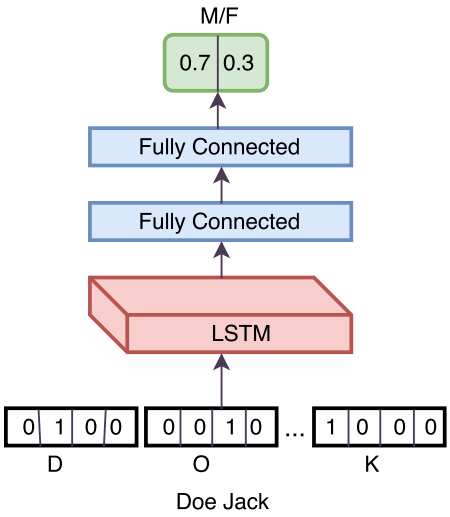

Figure 4.1: LSTM architecture for gender classification

To transform our task into a sequence classification task, we interpret each name as a sequence of characters. We include an out-of-vocabulary character encoding for any unseen characters at test time. The boundary between last and first name is encoded as the special character ",". We assume a fixed vocabulary of K characters and use K-dimensional vectors of 1-to-K encoding as our features. Our K value is 31, accounting for the English alphabet and a few special characters. The architecture uses one character-level LSTM layer, 2 hidden layers with 200 units and a sigmoid output

[7] https://mbejda.github.io/

layer. (Figure 4.1) For training, we use the Adam optimizer, 64 LSTM units and 200 units in each dense layer.

As mentioned, we test our approach on a manually labelled dataset with ACL authors, which is potentially more difficult than US names, since it contains names for a more international pool, including Chinese and Japanese names. We use the dataset released in Vogel & Jurafsky [2012], removing those names that the author has indicated as classified with unreliable online APIs. While the validation accuracy on the US inmate dataset reaches 94.3%, the ACL test set accuracy is 76.7% (with 78.3% for males and 75.2% for females, on a dataset with 5718 names).

With more hyperparameter tuning, we could probably see a slight improvement in accuracy. We are also aware that the dataset is not rich enough and an alternative source should be found. A better dataset with a higher percentage of international names should improve generalization performance.

We believe that these results are encouraging for the task of gender classification based on names. However, an accuracy of 76.7% is too low for the scope of this project, as we want to reliably predict the gender of authors, in order to make any conclusions concerning the gender gap. We therefore choose not to switch to a deep network approach and instead use the basic database and API based methods outlined in Section 4.2.

4.3.2 Face classification

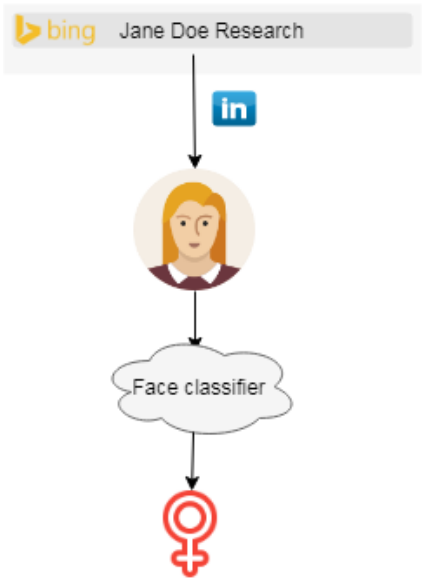

Figure 4.2: Face classification

We address the problem of certain cases where the name is not enough for classification. For unisex names, we needed additional information about the author, apart from

name. One possible solution is based on image classification, as it brings in additional information.

At an implementation level, we choose to scrape images returned by search engines (Bing [8]), by appending the word "research" after the author's name and searching for this term. We only trust results from professional profiles (LinkedIn.com, scholar.google.com) or academic websites (in the .ed, .edu subdomains) and we look at the first five results, trusting the highest one ranked that comes from a reliable website. Anything beyond the first five results is probably not related to our search.

We use Microsoft Azure Face API [9], a service for face detection which returns various face attributes, including gender. The images returned by the search engine reslts are passed to this service for classification (Figure 4.2). Unfortunately, there is no confidence score for the prediction, just a binary answer. We are being very conservative and would rather return "unknown" than an unreliable classification. On a gold test set of 300 names (the small sized is due to the costs of the API), we misclassify 18 names. (6%) The errors happen either because we are scraping the wrong image (when researchers of both genders exist with the same name), or the Face API gender attribution is mistaken.

Due to the error proneness of the results, we use this method only as an aid to human manual annotation. Due to the nature of our research, we want very precise name classification results. Therefore, we ultimately use the face detection in combination with human supervision. This speeds up the manual annotation task, as it directly presents possible images scraped by the search engine, without needing to manually look up a name page.

It is worth noting that commercial systems for image classification systems can also suffer from bias in their training sets. Buolamwini & Gebru [2018] assess such systems and find that darker-skinned females have error rates as high as 34.7%. This raises further awareness of the biases introduced by unbalanced datasets in machine learning systems.

4.4 Final system and evaluation

From our experiments, it is clear that there is no perfect approach for this task, as even human annotators fail to classify all instances. We were, however, able to build a multi-stage system that leaves only 8% of the dataset to manual classification, which we perform with the aid of our Face detection search approach.

The stages of the final system are:

1. Look up the first name in all available databases (US census, Indian names, and assign the gender that has a non-zero count. Pass if the name has non-zero counts for both genders.

[8] www.bing.com

[9] https://azure.microsoft.com/en-gb/services/cognitive-services/face/

2. Check the morphological gender of the last name (ova, ov)
3. Classify first name with Namsor API, only if the confidence score is above 0.8
4. Look up baby name statistics with GPeters and trust any ratio above 2
5. Anything else is manually classified using search engine results

To test our end system, we select a sample of the dataset released by Vogel & Jurafsky [10]. A smaller sample was used due to the cost charged by some of the APIs we used, which depends on the number of requests. We use a random stratified sample of 1064 names, with 27.25% females. Names come from diverse ethnic backgrounds.

The results of gender classification system are summarised below:

Total	Correct	Incorrect	Unknown
1064	997	8	59

Table 4.1: Name classification on test set

This translates to 99.2% precision and 94.5% recall. Out of the incorrectly classified names, 50% where females. Out of the unknown names, the true distribution was 27.1% females. The mistakes are introduced by the database retrieval method (37.5%) and by Namsor API.

In terms of algorithms used, the following breakdown results:

- retrieved from a database: 70%
- morphological gender: 0.03%
- Namsor API: 14.7%
- GPeters: 9.8%

and 5.5% of the names are unknown.

Given the high precision of the system, we use this approach to classify all names in ACL and arXiv. In the ACL corpus, out of 18128 distinct authors, we were not able to identify the gender of 973. (5.4%). We consider this a good result, as similar work carried by Vogel & Jurafsky [2012] does not classify 702 authors out of 12692 (5.5%). In the arXiv corpus, we were not able to classify 3% of 11918 the authors. The task was easier in this case, as many authors have social media profiles, or where already present in the ACL corpus.

Summary

We look up first names in reliable name lists. If a first name consistently only appears with one possible gender, we assign it. We then use reliable APIs and perform manual checks on the results. We use search engine retrieved images and face detection

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

algorithms for aiding human labelling. Our methods show comparable results for both classes and therefore no systematic gender-bias can be asserted. Classification on Chinese names achieves poor results so we resort to manual labelling and images found on the Internet seem to be the best method foe dealing with this issue, when classification based on first names will not be able to disambiguate.

Chapter 5

Publishing Patterns in the ACL corpus

In Chapter 2, we presented overall statistics of PhD and undergraduate enrolment across the field, with a growing presence of women at all academic levels. However, we are far from closing the gender gap, with inequalities persisting especially towards higher academic ranks. While the effort of initiatives focusing on mitigating gender disparities do solve some of the problems, close analysis might reveal differences in terms of author position, acceptance at scholarly journals, productivity and longevity in the field. Being the author of a published paper indicates active involvement in research and can represent an important factor for getting tenure. Therefore, analysing the publishing patterns of authors is extremely relevant.

The main aim of this chapter is to go beyond simple statistics of enrolment and actually analyse factors involved in career progress. Our ultimate goal is to see how the increasing number of women graduating in the field translates into an increase of women in academia. A successful career track encourages women to stay in the field and write quality research, from a more diverse perspective than a man dominated field could achieve.

5.1 Challenges

It is very easy to introduce preconceptions in our interpretation and specifically look for bias, but considerable care was taken in order to tackle this challenge. Our study is comprehensive and aims to present a multifaceted view of gender's presence in academia and the way they interact with the field over time. Any claims that we make are backed up by data. We look at various components of academic careers, including productivity, dropout, collaborations.

Another challenge of this chapter was capturing the right patterns in the data and displaying them. Given the size of the population, we can not look at individuals, but instead need to consider summary statistics and perform statistical significance tests.

We often produce side by side visualisation to compare patterns in the female and male populations.

5.2 Overall statistics

We first discuss some overall statistics for the ACL Anthology.

Authors by gender

There are 23766 publications in our dataset and 18128 unique authors.

Figure 5.1 shows gender publication statistics over time. A given author is included in the analysis for a year if they published at least one paper in that year. The number of authors is generally increasing every year, for both women and men, showing Computational Linguistics is a growing field.

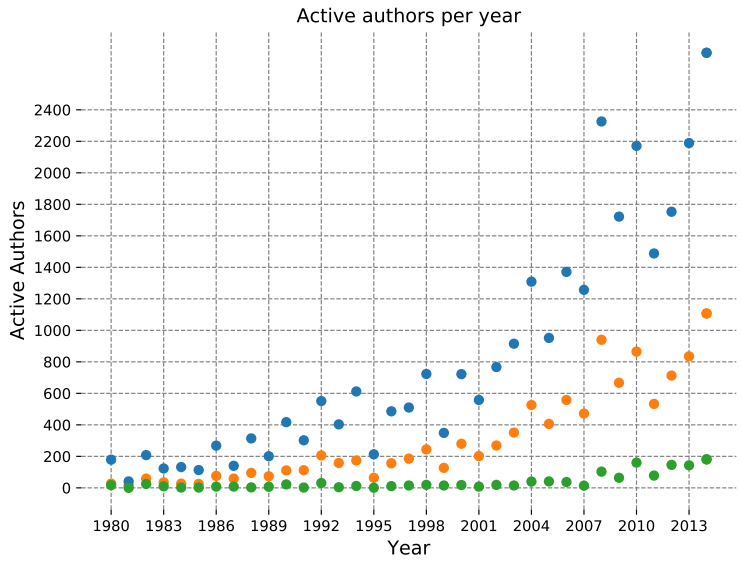

Figure 5.1: The number of authors of a given gender publishing at least one paper in a given year

Figure 5.2 shows the percentage of authors of a given gender overtime rather that count values. We only display years after 1980, as the data was too sparse in earlier years. We fit a regression line for both genders. We find:

$$y = -0.244x + 560.45, \text{ male best fit line}$$

$$y = 0.248x - 461.05, \text{ female best fit line}$$

The R^2 (squared correlation coefficient) for the male best fit line is $R^2 = 0.448$, with p-value (with the null hypothesis being that the slope is zero) $p = 1.09 \times 10^{-5}$. The female best fit line has $R^2 = 0.446$, $p = 1.14 \times 10^{-5}$. The regression line gives us a robust estimate of the gender specific authorship percentage over time.

The steepness of the line suggest a rather slow rate of change. In 1980, 12.1% (27 as a raw count) of the authors publishing that year were females, compared to 27.3% (1107 raw count) publishing in 2014. However, the percentage was already 27.9% in 2008, with improvement slowing down in more recent years.

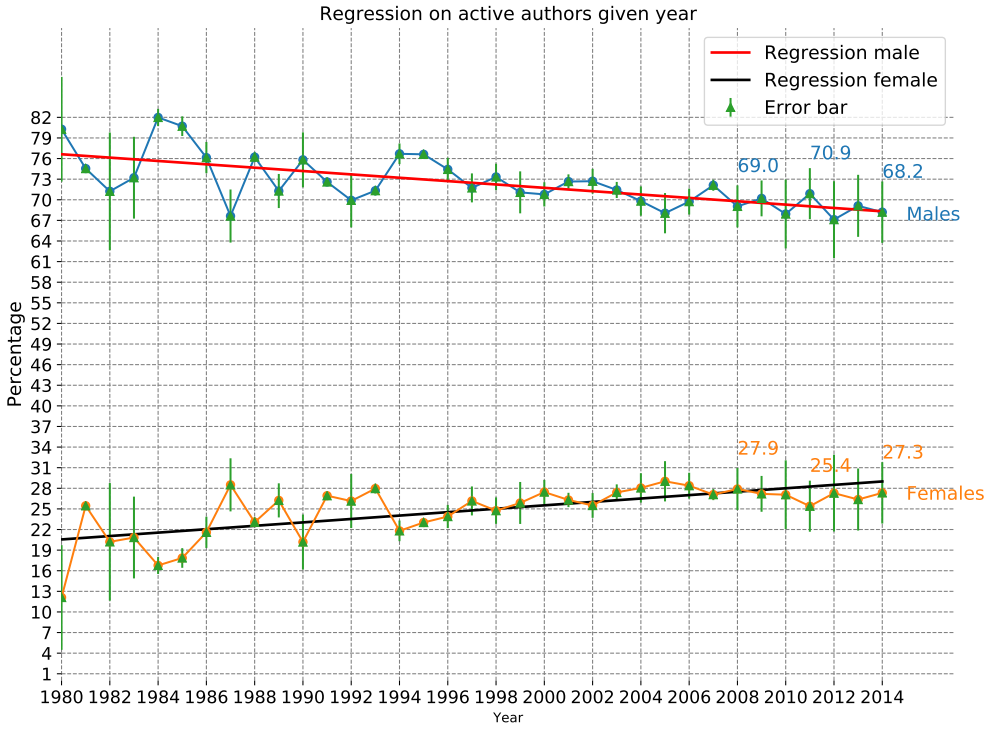

Figure 5.2: Percentage of authors of a given gender publishing at least one paper in a given year. The error bar indicates the percentage of authors classified as "unknown" for that year.

Authorship instances

Authorship: We define an *instance of authorship* as every pair (publication, person) for which the person is listed as a co-author.

At the beginning of this section we determined women represented 26.9% of the authors in the anthology. In terms of authorship instances, they represent 25.5% of all authorship pairs in the dataset. (there are 61611 authorship instances and 18128 unique authors). The small gap in terms of productivity could be explained by more young female scientists joining the field and publishing less papers at the beginning of their careers than more senior scientists. We would expect authors who have been longer in the field to author more papers.

5.3 Productivity and number of active years

The length of publication record and rate of publishing are also important factors defining one's career. It can also uncover properties of our dataset which might correlate with other findings. We split the authors based on the total number of publications in their careers. In Figure 5.3, we present the distribution of publication length record for males and females and find that the distributions are simlar. Out of all females, 55.8% appear with only one published paper, compared to 54.5% for males.

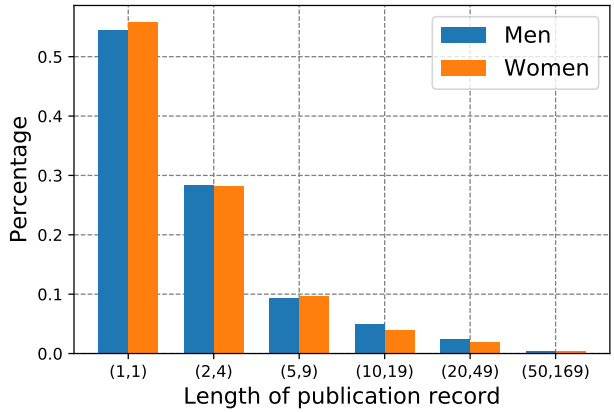

Figure 5.3: Length of publication record distribution for males and females

We take a closer look at the head of the list. We compare the most prolific authors of each gender. Figure 5.4 shows the number of papers published by the top 400 authors of each in gender. They are sorted in decreasing order, so we are looking at the most prolific authors. We note that most prolific male authors have more publications than their female counterparts. The next question we address is whether this is because men have been in the field longer.

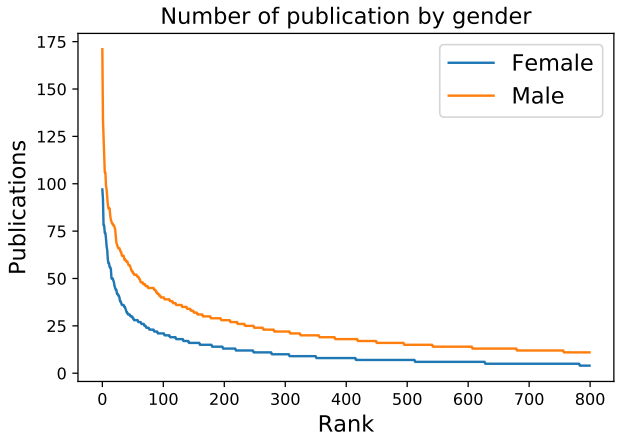

Figure 5.4: Number of publication per author in decreasing order

We consider the number of years between an author's first and last publication as the number of active years. We find that on average, women have been in the field for 2.61 years, compared to 2.94 years for men. To compare this with productivity, we find that on average, women write 3.22 papers, while men write on average 3.61. To test whether men do indeed publish more and have been in the field for longer, we perform a two-sample t-test. Because we are testing that the number male publications is greater than the number of female publications, we use a one-tailed test, based on Welch's t-test.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

where $\bar{X}_1, s_1^2$ and N_1 are the first sample mean, sample variance and sample size, respectively.

Welch's test was used because it is more reliable than Student's t-test when the two samples have unequal variances and unequal sample sizes [Ruxton, 2006], as it is in our case. We find that on average, men indeed publish more.(p=5×10^{-4}, we reject the null hypothesis) Performing the same test on the number of active years, we obtain p=2×10^{-4}.

We calculate the number of papers per active year for every author in our field and take the average. We find that women write on average 1.21 papers per active year, while men write 1.24. We compute this by This difference is again significant ($p = 4 \times 10^{-4}$), suggesting men are more prolific than women.

In terms of productivity, we found that men publish more papers per active year and they have been longer in the field, on average. The most prolific male authors also write more than the most prolific female authors.

5.4 Position in authors list

We consider position in authors list as indicative of one's career progression. Authorship order patterns vary among fields, with some fields listing authors alphabetically (Mathematics, Economics). In Computer Science and Computational Linguistics, the first and last author are considered to be the most prestigious. The first author is usually the person most involved with the paper. In multi-author papers, the last author is usually the group coordinator, a person in a more senior position. [Solomon, 2009]

We look at the first and last position in the list of authors as meaningful with regards to one's career. We expect individuals to progress from being first authors to last authors, as an indicative of academic maturity through supervising the work of others. We also consider being a single credited author as an indicative of academic independence. Solomon [2009] discusses the difficulty of judging one individual's specific contribution. He recommends pairing the list of authors with a list of short description of their work

and dividing authors by level of contribution. Having a solid publishing record is vital for career advancement and the degree of credit that an academic receives can impact their future progress or even receiving tenure.

There is evidence that women tend to get less credit for their work. In Economics, where alphabetical order is usually used, many journalists reporting on a publication switched names to list the man's name first. [Washington Post, Accessed 27 February 2018]. This further enforces the idea that ambiguity in crediting one's work is dangerous and can enforce (unconscious or conscious) bias.

Having motivated the importance of position in authors list, we report the gender composition by authorship position and overall.

Women are under represented as last authors

First, we compute the percentage of women in all authorship positions. For computing the percentage of female first authors, we count all papers with at least 2 authors and observe the gender of the first author. For computing last authors, we take all papers with at least 3 authors and observe the gender of the last author. We present authorship order results in Figure 5.5. We find that women are under represented when it comes to the last author position. This result could be again motivated by the fact that the percentage of females has only been growing recently, with a lot of new researchers who are at the beginning of their careers. We also note that in recent years the same trend continues, with last authorship frequency below the overall female authorship frequency. (Figure 5.6)

We now ask how long it takes for an author to become last author of papers. For this, we only consider authors who have published at least 5 papers before being listen as last author. This is to ensure they are not being listed on that position by chance. We note that on average, women take 7.22 years to before they publish their first paper as last author, while men take 7.58 years to become last author. This is only significant at $p < 0.05$ (p=0.02).

Authorship position of most prolific authors

Having presented the view of authorship position as an indicative of one's career progression, we verify our assumptions by visualising the most prolific females and males, as established in the previous section. Figures 5.7 and 5.8 present the top 15 most prolific females, respectively males. This allows us to embody the average statistics we have reported so far. The names of the authors are not displayed. We distinguish an author as either the solo author of a paper, the first author of a paper with at least 2 authors, the last author of a paper with at least 3 authors, the second author of a paper with 2 authors, or middle position. We consider being the last author of a publication as a sign of seniority. Thus, this visualisation presents the publication timeline of the most prolific authors in the dataset. We notice that both genders tend to become last authors at later stages of their career, which enforces our finding related to the average time required to become last author. All most prolific authors started publishing before the year 2000 (usually around 1980) and they were still publishing in 2014, which is the last year available.

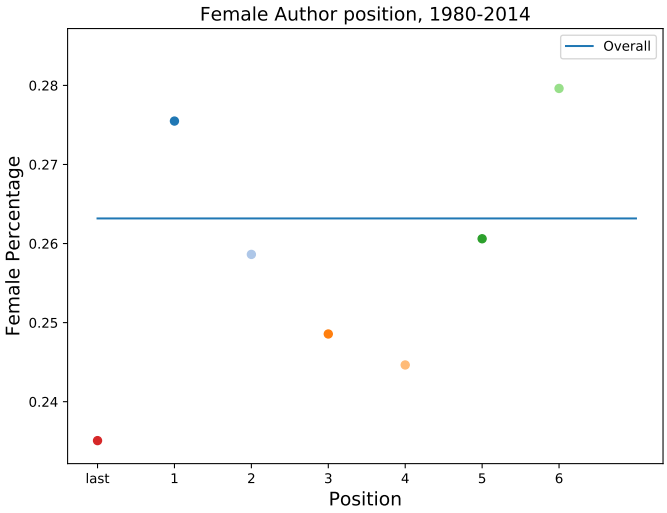

Figure 5.5: Percentage of women at each authorship position. The horizontal line represent the overall frequency of female authorships, over all author-publication pairs. For first author, we look at papers with at least 2 authors. For last author, we look at papers with at least 3 authors.)

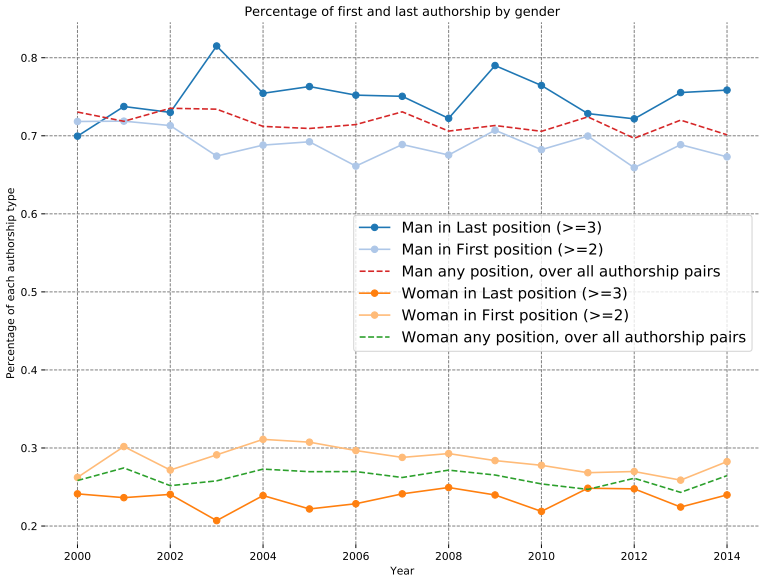

Figure 5.6: Frequency of first and last author position by gender and year. The overall frequency refers to female authorship over all author-publication pairs.

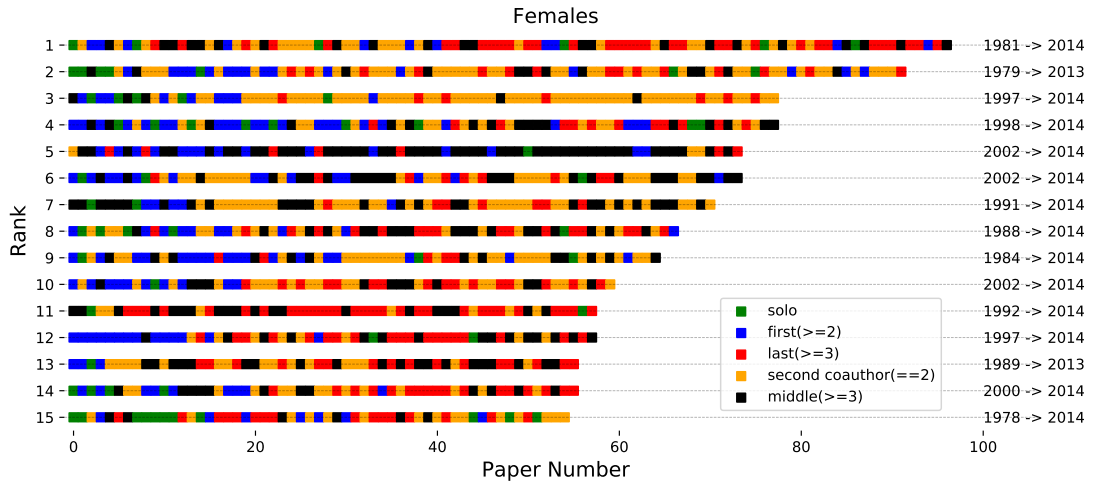

Figure 5.7: Most prolific female authors. Each line is a different author, order by their number of publications. Paper number represents the n-th paper and it's colour represents authorship type.

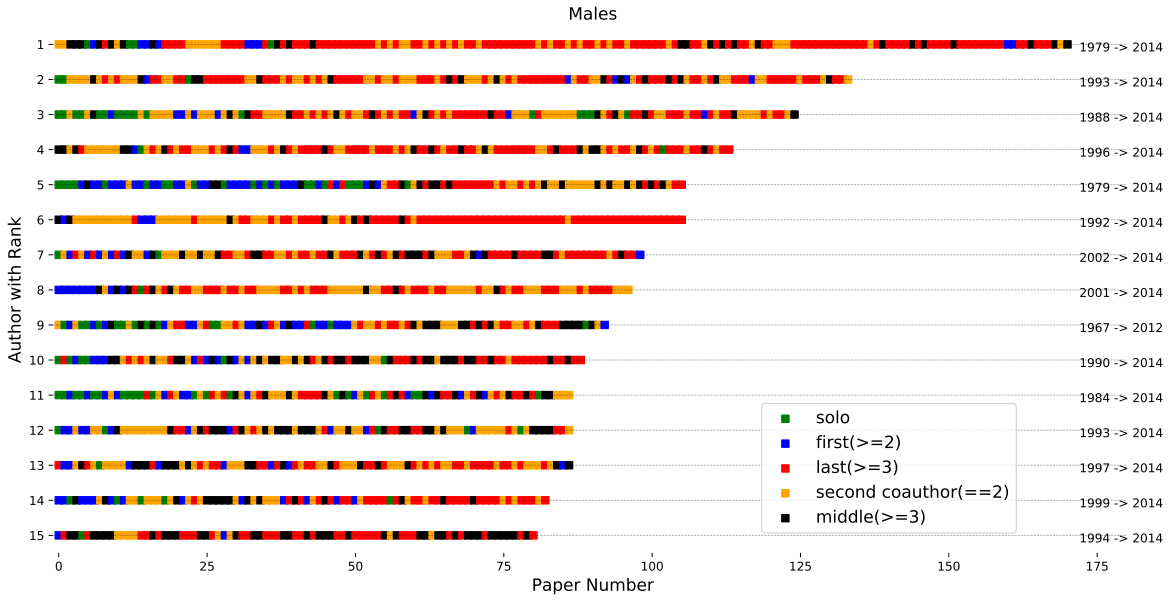

Figure 5.8: Most prolific male authors. Each line is a different author, order by their number of publications. Paper number represents the n-th paper and it's colour represents authorship type.

5.5 Cohort analysis

In order to understand the gender gap among scientists, we analyse the distribution of researchers that successfully sustain an academic career over years. We refer as "cohort" to all authors who begin publishing around the same time, that is, they publish their first article in the same year. Comparing the development of cohorts allows us to unveil patterns over longer periods of times. Since, as our analysis so far shows, there was a smaller fraction of women in the early year of the anthology, it is important to control for length of time in the field. Figure 5.9 shows the number of authors per year and gender with continued careers 5 and 10 years after their first recorded publication. The percentage of authors we identified as female shows an increasing trend. (Figure 5.10) Due to sparse data, we only present the analysis after 1990.

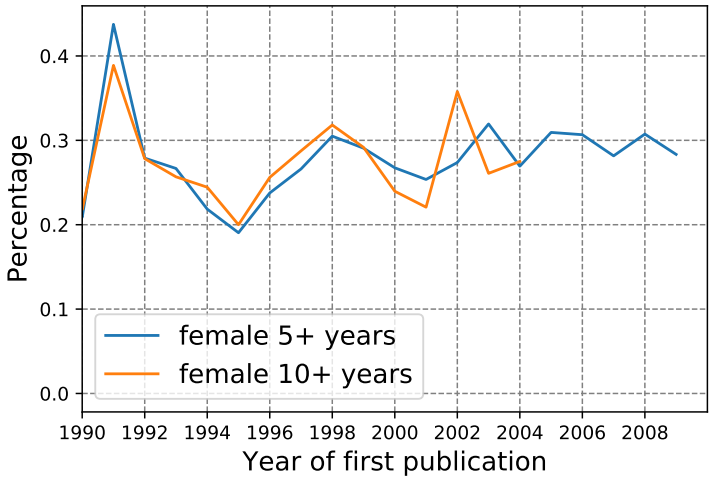

Figure 5.9: Percentage of females per cohort year with a career lasting more than 5 and 10 years

Visualising the cohort of 1990 and their active years

We visualise the top 10 and last 10 authors, in terms of number of papers, who started publishing in 1990, so they belong to the same cohort. Figure 5.11 illustrates our findings. The most striking aspect observed here is that men in this cohort write more papers per year, visible in the darker gradients of the figures. This cohort-specific observation also supports our general finding, with men publishing more papers per year, on average.

Publishing at the beginning of the career

Having a good career start is also important in academia. We have analysed the average number of publications of men and women during the 5 and 10 years after their first article. We find that over a 5 year span, women publish 7.23 papers, while men publish 7.36. Over a 10 year span, women publish on average 12.46 papers, compared to 12.77. The differences are not significant (t-test), which is important, as it might indicate that

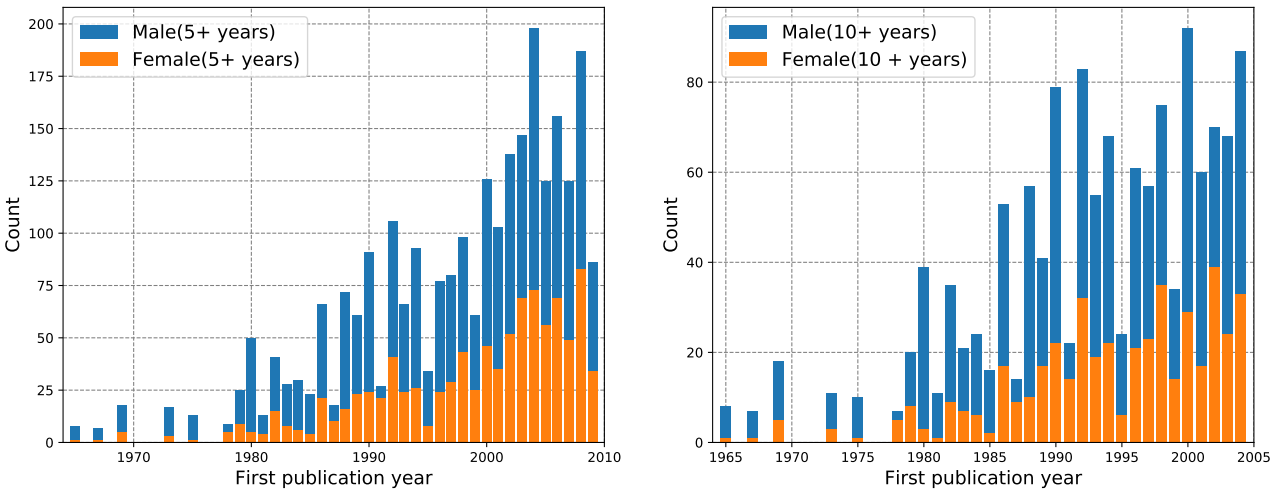

Figure 5.10: Number of authors per cohort year with a career lasting more than 5 and 10 years

the publication records of both genders are similar when they are at the beginning of their careers.

Dropout rate per cohort

We are also interested in the dropout rate of authors. Specifically, those authors who publish their last paper between 5 and 10 years after their first one. We count this as a discounted career, as it might indicate the author decided to leave research, or could not secure a permanent position. We analyse all authors who have published for at least 5 years and their dropout rate. We do this analysis over the cohort of 1974-2004, as we can not determine dropout after this time frame. We find that females have an overall dropout rate of 32.8%, while for men the dropout rate is 31.7%. Figure 5.12 presents the dropout rate for a given year.

5.6 Coauthorship

Coauthorship is another important aspect in one's publication record. Being involved in joint research indicates integration within the community, as well as productivity. Social Science studies tend to suggest that men and women have different ways of networking, which might lead to different patterns when it comes to collaboration with other researchers. Zdenka [2009] claims that women are not good at networking in academia, while Barthauer et al. [2016] discover that women in academia have less dense networks.

Women publish slightly less single authored papers:
Collaboration has been growing in recent years and out of all papers, only 24% are single authored. Out of this, women publish only 23.7% of the solo-authored papers.

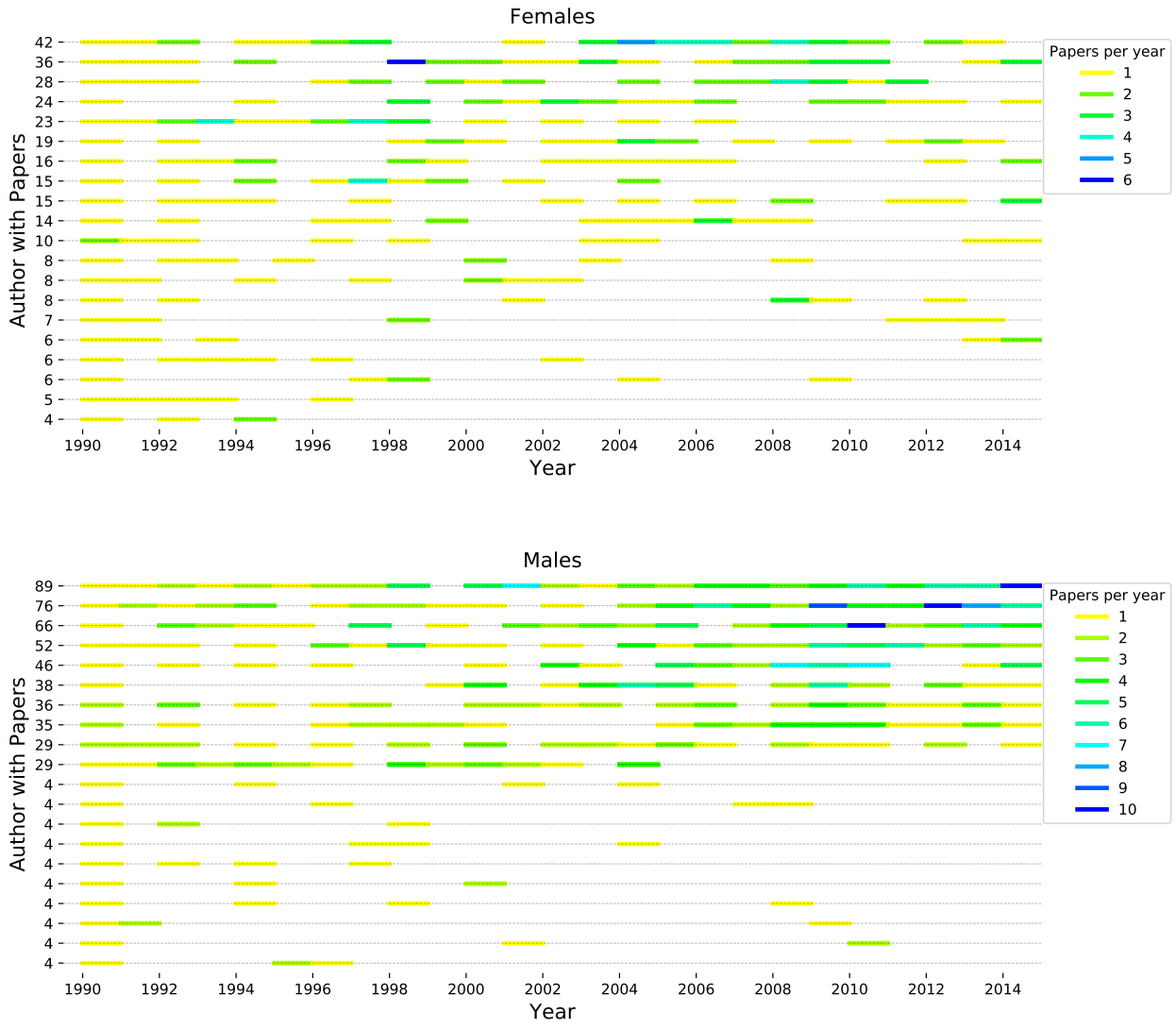

Figure 5.11: Top and bottom 10 authors in terms of number of publications. The colour intensity represents the number of publications in a year.

In academia, publishing a solo paper in regarded as a sign of career development and independence. On average women publish 7.7% of their papers as single-authored papers, compared to 9.2% for men, when taking into consideration authors which appear on at least 2 publications. Out of all women, 15.7% have published at least one single authored paper, compared to 17% for men.

Women have on average less coauthors:
We found that on average, women publish with 6.32 distinct authors, while men publish with 6.65 coauthors. This is a very coarse estimate, as it does not account for the length of the publication record. In Figure 5.13 we measure the mean of the number of distinct coauthor per author, in each range of length of publications. We see that the number of coauthors are very similar in this category, the difference in the overall mean coming from the difference between the most prolific authors. No woman has more than 100

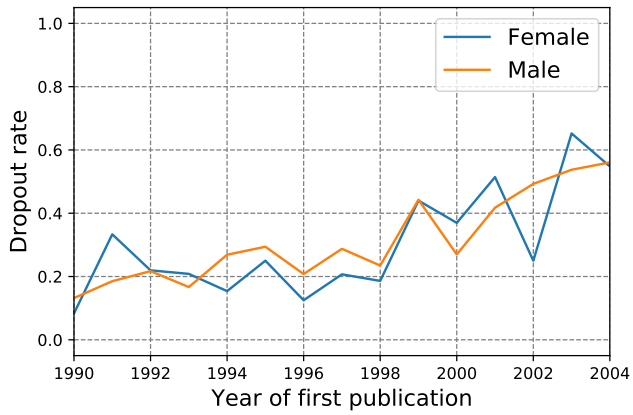

Figure 5.12: Dropout rate: percentage of males and females who have published for at least 5 years and stopped publishing after at most 10 years

papers, while 5 men do.

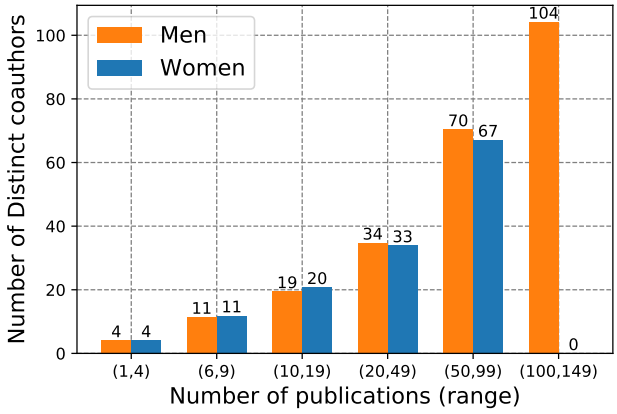

Figure 5.13: Mean of the number of distinct coauthors per author, by length of publication record

Women publish more often with female coauthors

Having found that women have less coauthors on average, we look at the gender distribution of these authors. We find that, on average, 24.19% of a female's distinct coauthors are other females. We then find that 24.03% of a male's distinct coauthors are females. This is not statistically significant. On the other hand, if we consider all coauthors, and not just distinct ones, we find that out of all their coauthorship instances, women have 32.05% females, compared to 24.03% for males, with $p = 1.68 \times 10^{-8}$. This suggests that women publish more often with other women.

5.7 Publication at highly ranked conferences

A very important component in one's research is having their work accepted at major conferences. While there is no official ranking of the conferences, Google Scholar [Accessed 21 March 2018] provides a list of top venues/conferences in Computational Linguistics, based on the h-5 index, the metric used by Google Scholar to quantify impact. It is based on the h index, a widely used metric introduced by Hirsch [2005], and it computes the h index for articles in the past 5 years. According to this ranking and restricting ourselves to the conferences for which we have sufficient publications in the dataset, the ordering as of March 2018 is:

- Meeting of the Association for Computational Linguistics (ACL)
- Conference on Empirical Methods in Natural Language Processing (EMNLP)
- International Conference on Computational Linguistics (COLING)
- Conference of the European Chapter of the Association for Computational Linguistics (EACL)
- Conference on Computational Natural Language Learning (CoNLL)

The highest impact venue is actually considered to be the Computation and Language subsection on arXiv, an online repository. We will turn our attention to arXiv as a separate publishing medium in Chapter 7.

Figure 5.14 presents the percentage of female authors in papers present at well known Computational Linguistics conferences: ACL, EMNLP, EACL, COLING and CoNLL. We include an author in this analysis if they have coauthored a paper present at the respective conference. We compare female representation at these conferences, with overall representation in all conferences. By overall representation, we refer to an instance of authorship i.e an author coauthoring any paper is a data point, as usual. Table 5.1 presents the results, which show that these conferences have a female authorship representation below the overall value of 26.3%. The highest rank conference, ACL, has a female representation of 23.3%.

Conference	Female authorship percentage
All venues	26.3
CoNLL	21.6
EACL	24.7
COLING	23.2
EMNLP	22.4
ACL	23.3

Table 5.1: Female authorship at top Computational Linguistics Conferences

We perform a breakdown of female representation at top ranked conferences over the years. We see little improvement over the years in terms of female representation at these conferences. In the next section, we will correlate the presence of females at top conferences with the top topics that are observed in the accepted publications.

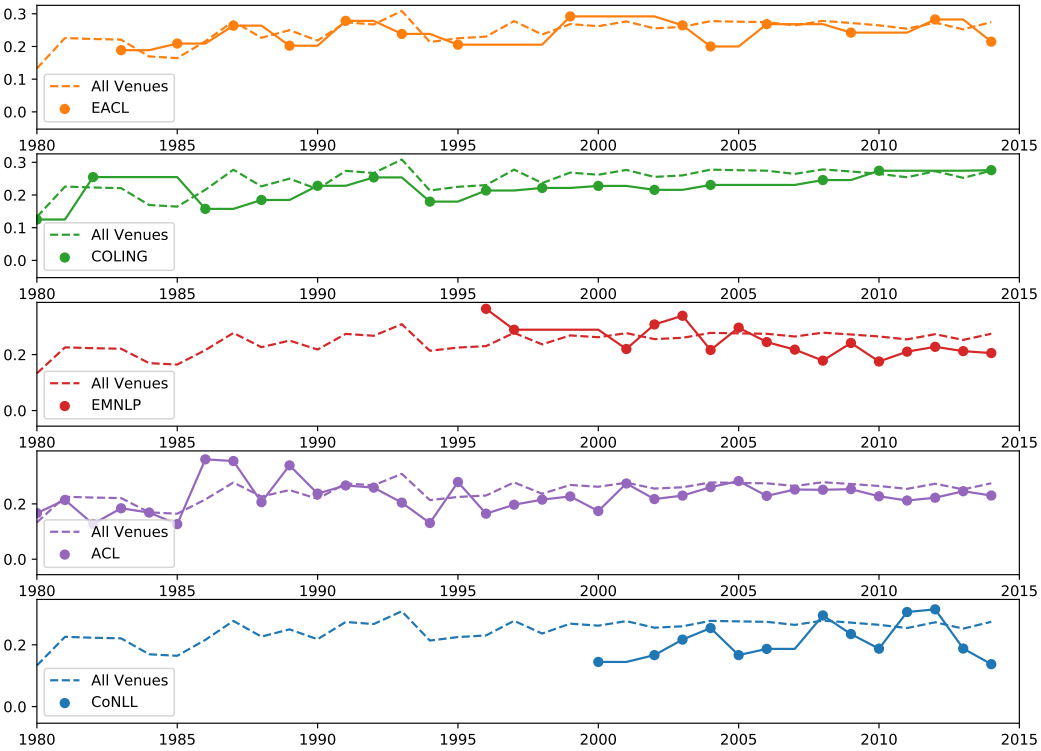

Figure 5.14: Female authorship by conference. Note some conferences started later.

5.8 Summary of publishing patterns

Studying factors in scholarly output proves out to be fruitful, with interesting insights into the gender distribution of the field. Some of out most significant findings were:

- While the total count of females in the field is increasing, the percentile increase have been stagnating in the last 10 years, with 27.9% females in 2014.
- We find that women are underrepresented as last authors, but there has been an increase in their presence as first author. It also takes longer on average for a woman to reach the point of her career when she becomes last author.
- Men have a higher average number of active years in the field, 2.94 compared to 2.61 for women.
- Even consider the above point, women are less prolific and publish less papers per active year (1.21 compared to 1.24). Moreover, the most prolific male authors have more publications than the most prolific female authors
- We find that females have a higher dropout rate (32.8%, compared to 31.7%).

- In terms of coauthorship, we find that women have on average less coauthors. What is more, 32.05% of their coauthorship pairs are with other females, compared to 24.03% for males coauthoring with females.
- Women's presence at the highest ranked Computational Linguistics venues is below their overall presence.

Chapter 6

Topic modeling and the ACL corpus

In this chapter we analyse the connection between gender and document content. Our main tool is topic modeling, which is used for discovering hidden "topics" in a (usually large) collection of documents. Topic modeling has been receiving an increasing focus due to the large collections of data that are readily available. Its applications are varied (it can be used for extracting features, document retrieval, analysing large collections and many more) and for our purposes we will use it as an unsupervised method of finding topics in academic publications.

Our study employs Latent Dirichlet Allocation (LDA Blei et al. [2003]), a generative statistical model that we will use to model topics in our documents. After observing each document as a mixture of topics, we can compute topics that are more linked with each gender. We will also look at how topics evolve over time, to better understand our field and its authors. Our hypothesis is that gender have preferences for certain topics and this can be reflected in the data.

We will start by introducing key concepts about the algorithm, as understanding the model's background will be useful in fine tuning our experiments, especially understanding the importance of priors and the difficulty of evaluating results.

We will then explain the methodology of our experiments. Having laid out all the necessary background, we will extract topics from the ACL corpus and interpret their connection to gender by computing the odds ratio of each topic. Finally, we will analyse the topic composition of the highest ranked Computational Linguistics venues.

6.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation is one of the most common topic models that is used in practice. A topic refers to a multinomial distribution over a fixed vocabulary. Each document is assumed to be a multinomial mixture of topics.

6.1.1 Notation and key distributions

We introduce the following preliminary notation, consistent with Blei et al. [2003]:

- A collection of D documents $\mathcal{D} = \{d_1, d_2, .., d_D\}$
- A vocabulary V with $|V|$ words
- A document d of N_d words is represented as a bag of words $\mathbf{d} = (w_1, w_2, ..w_{N_d})$.
- A fixed number of topics K and we refer to topic n as z_n

The model relies on sampling from a Dirichlet distribution and a Multinomial distribution.

Multinomial distribution

The multinomial distribution is a generalisation of the binomial distribution and has the following probability mass function:

$$P(x_1, x_2, ..., x_K; \theta_1, \theta_2, ..., \theta_K) = \frac{n!}{\prod_{i=1}^{|V|} x_i^{x_i}} \prod_{i=1}^{|V|} \theta_i^{x_i}$$

where $\theta_i = P(x_i)$, probability that the word i is seen and x_i indicates the number of times word w_i of the vocabulary is observed.

$$n = \sum_{i=1}^{|V|} x_i, \quad \sum_{i=1}^{|V|} \theta_i = 1, \theta_i > 0$$

Dirichlet distribution

The Dirichlet distribution has parameters $\alpha_1, \alpha_2, ..., \alpha_K > 0$ and has the following probability density function:

$$f(x_1, x_2, ..., x_K; \alpha_1, \alpha_2, ..., \alpha_K) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma_i(\alpha_i)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

where $\Gamma()$ is the gamma function and $\{x_k\}_{k=1}^{k=K}$ belong to the standard K-1 simplex

$$\sum_{i=1}^{K} x_i = 1 \text{ and } x_i \geq 0 \text{ for all } i \in [1, K]$$

The multinomial distribution can also be expressed in terms of the gamma function. The Dirichlet distribution is an exponential family distribution and it is conjugate to the multinomial. In Bayesian probability theory, if the posterior and the prior are in the same family, then the prior and posterior are called **conjugate distributions**. Conjugacy implies the predictive posterior distribution is analytical, which simplifies inference.

6.1.2 Generative process

The following presentation follows that of Blei [2012] on the topic of Probabilistic Topic Modeling.

The **generative process** for a document with N words $\mathbf{w} = (w_1, ..., w_N)$ goes as follows

1. Choose a number of words N_d for a document d, usually $N_d \sim Poisson(\xi)$
2. Choose a topic mixture for a document according to a Dirichlet distribution over a set of topics. θ is sampled from a Dirichlet$(\alpha_1, \alpha_2, ...\alpha_k)$ distribution
3. For each of the N_d words:
   - Sample a topic $z_n \in 1, ..., k$ from Multinomial(θ)
   - Sample word w_n conditioned on topic z_n, from the multinomial distribution $p(w|z_n)$

Note that we use a unigram bag-of-words model to represent the words in a document.

We summarise the algorithm below, and explain any further notation:

```
For j = 1...T topics:
    Choose φ^(j) ~ Dirichlet(β)

For d = 1...D documents:
    Choose θ^d ~ Dirichlet(α)
    For i = 1...N_d words in document d:
        Choose z_i ~Multinomial(θ^(d))
        Choose w_i ~Multinomial(φ^(z_i))
```

α and β are hyperaparemeters:

- α is a Dirichlet prior of the topic distribution and it controls the mean shape and sparsity of θ
- β is a Dirichlet prior of the per-topic word distributions

The graphical model for LDA is displayed below.

6.1.3 Inference

The inference problem that needs to be solved computes the posterior distribution of the hidden variables given a document.

Given the parameters α and β. the total probability of the model is:

$$P(\mathbf{W}, \mathbf{Z}, \theta, \phi|\alpha, \beta) = \prod_{i=1}^{K} P(\phi_k|\beta) \times \prod_{j=1}^{M} P(\theta_j|\alpha) \times \prod_{t=1}^{N} P(Z_{j,t}|\theta_j) P(W_{j,t}|\phi_{Z_{j,t}})$$

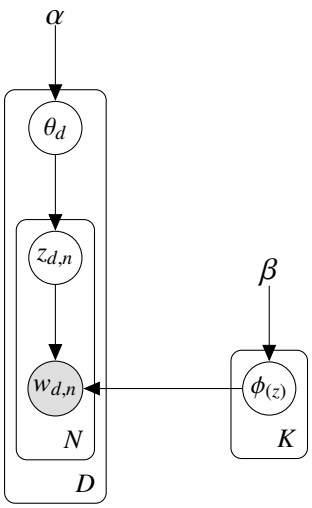

Figure 6.1: LDA Graphical model. Nodes are random variables. Shaded nodes are observed. Edges represent dependence. Plates stand for replicated variables.

As before, N is the number of words in all documents, K is the number of topics, D is the number of documents. We use vector notation: α is a K-dimensional vector, β is a V-dimensional vector. $\mathbf{Z}$ is an N-dimensional vector of integers between 1 and K and represents the topics of all words in all documents, $\mathbf{W}$ is a N-dimensional vector of integers between 1 and V and holds the identity of all words in all documents.

After marginalising the hidden variables θ, ϕ in the joint distribution, computing the evidence $P(\mathbf{W})$ for the posterior $P(\mathbf{Z}|\mathbf{W}, \alpha, \beta)$, becomes intractable. [Blei et al., 2003]. A wide variety of algorithms applicable to approximate inference can be considered, including Laplace approximation, variational inference(which is what Blei et al. [2003] describes) or Gibbs sampling [Griffiths & Steyvers, 2004]. Variational inference is usually implemented in practice due to speed considerations.

6.2 Model evaluation

Topic models are usually evaluated by measuring performance on another task [Griffiths et al., 2007], such as document classification, synonym tests for TOEFL, asking respondents to find the intruder word. Griffiths et al. [2007] detail on other evaluation methods, focusing on the free-association task, in which participants are given a cue word and asked to write down the words that first come to their minds. Topic models can then be evaluated based on how well they predict human word association.

As other statistical tasks, a common metric is the probability of held-out documents given a trained model. Again, the computation of this is intractable. Wallach et al. [2009b] propose methods that are more accurate and efficient estimators. (a Chib-style estimator and a "left-to-right" algorithm for decomposing the posterior). Once we

decide on a way of estimating this probability, log perplexity can be computed:

$$perplexity(D_{holdout}) = \exp\left(-\frac{\sum_{d=1}^{D}\log(p(\mathbf{w}_d))}{\sum_{d=1}^{D}N_d}\right)$$

In many cases, perplexity measures do not always agree with human judgement. Chang et al. [2009] performed a large user study and found that quantitative methods such as those explained by the study of Wallach et al. [2009b] do not agree with how humans measure the interpretability of topics. This motivates additional work for trying to better model human judgement, when perplexity fails.

Visualisations are often a useful alternative to decide on the quality of topics. Sievert & Shirley [2014] introduce LDAVis, a method for visualising and interpreting topics, their meaning and how to interact. Their tool is useful for getting an overall feeling of the data.

Furthermore, topic coherence was introduced by Newman et al. [2010] as a better quantitative way to measure interpretability, claiming high agreement with human opinions. Their proposed measure is term co-occurrence based on a way of scoring each pair, applied to the top N words from the topic and averaging over all pair scores. Their best performing scoring function is Pointwise Mutual Information and it is known as the UCI metric:

$$score_{UCI}(w_i, w_j) = \log\left(\frac{p(w_i, w_j) + \epsilon}{p(w_i)p(w_j)}\right)$$

Further contribution to topic coherence metrics is brought by Mimno et al. [2011], who define the UMass metric, where the score is based on document co-occurrence.

$$score_{UMass}(w_i, w_j) = \log\left(\frac{D(w_i, w_j) + 1}{D(w_i)}\right)$$

where $D(v)$ represent the number of documents with at least one token of type v. This score function is asymmetric and the main advantage it claims is better detection of models with low-quality topics.

6.3 Methodology

Having laid out the basics of topic modeling, we will present the way LDA can be used to discover topics in our collection of documents.

Framework

There are many LDA implementations and topic modeling toolboxes designed for running quick experiments. Because we need more control on our model and want to be able to make changes programatically, we chose Gensim [Řehůřek & Sojka, 2010], a Python framework for topic modeling on large corpora, amongst other uses. The Gensim LDA implements online LDA [Hoffman et al., 2010], which processes the

corpus in chunks of documents and updates the model after each chunk, rather than updating after processing all the documents. If the topic drift is reasonable, model estimation converges faster. Gensim also offers a multicore implementation of LDA, for faster training. For inference, it uses the variational inference approach.

Input

To train our own LDA model, we require the following:

- A corpus of documents, where documents are represented as bag of words
- A fixed number of topics
- α and β priors
- Other training parameters (number of iterations of variational inference, number of passes through each mini-batch, frequency of updates)

6.3.1 Data processing pipeline

Since documents are the output of PDF to text conversion, they have been cleaned as much as possible, as explained in Section 3.1. We use the main body of each published paper, excluding the references. We remove any non alphabetical characters and lowercase all words.

To prepare our corpus, we go through the following pipeline:

1. **Tokenization and preprocessing:** We use the English tokenizer provided by the NLP framework SpaCy [1] to segment each document in words, removing punctuation and removing stopwords.
2. **Lemmatization:** The purpose of lemmatization is to remove inflectional forms and words that relate to some common base forms. In this way, *eats* and *eat* will both be reduced to the same base form, *eat*. In the context of topic modeling, this task is important as it removes inflections of the same word in our topics. The two common way of achieving this are stemming and lemmatization. Stemming is a more crude approach based mostly on heuristics, simply removing the end of the words. For example, *studies* becomes *stud*. Lemmatization, on the other hand, relies on the use of a vocabulary and morphological analysis of words,aiming to return the lemma of the word. For the token *saw*, stemming might return *s*, whereas lemmatization would attempt to return either *see* or *saw*, depending on whether it was used as a verb or noun. (Example based on Manning et al. [2008]). We prefer to use lemmatization as provided by the SpaCy lemmatizer, as a more advanced approach.
3. **Building a dictionary:** At the beginning of this step, all documents are in the form of a list of words in their base form. We collect all the words and their frequencies and construct a dictionary to store word to id correspondence. ($dog- > 0$, $cat- > 1$). We then filter the tokens in the dictionary based on their

[1] https://spacy.io/

frequency: we filter tokens that appear in less than 5 documents, as well as tokens that appear in more that half of the documents. This eliminates any typos or words that are related to citing authors, conferences, while greatly decreasing our vocabulary size from 436,777 tokens to 63,996.

4. **Building a corpus:** The final step of the data processing pipeline converts each document to its bag of words representation, using the dictionary built in the previous step. We make sure to save both our corpus and the dictionary correspondences, so we can recover the words from their ids.

6.4 Topics in the ACL corpus

6.4.1 Experiments and evaluation

The next step after building a corpus is to extract topics using LDA. We will present the methodology of our experiments and use the ACL corpus to illustrate it.

The standard LDA implementation requires a predefined number of topics. Since we are aiming to extract topics in Computational Linguistics, we expect to require a relatively large number of topics, with some popular topics being much more prominent than others. Models trained by Hall et al. [2008] and Vogel & Jurafsky [2012] on a previous version of the ACL corpus both use 100 topics. However, some of the topics contain just random words. It is also necessary to hand select seed words which we use as Dirichlet prior for the topic-word document matrix, in order to improve coverage of the field and to include topics that might have not been found.

Topics	UCI	UMass
50	1.08	-1.40
70	1.20	-1.49
100	1.14	-1.49

Table 6.1: Topics coherence (lower UMass and higher UCI is better)

Our evaluation relies on both topic coherence measures and manual analysis of the top 10 words of the resulting topics.

We report topic coherence for 50, 70 and 100 topics, using the two scoring function (UMass and UCI) explained in Section 6.2. Results are displayed in Table 6.1, with the best model having the highest UCI value and smallest (negative) UMass value. By inspection, the models with a smaller number of topics has less random words topics, while more topics introduce more noise. We keep the model with 70 topics, with the best trade-off between topic coverage and quality of the results.

Dirichlet priors

We mentioned the LDA models required Dirichlet prior of the topic distribution and per-topic word distribution. The default scenario is to use symmetric priors. However,

Wallach et al. [2009a] advocate on the importance of asymmetric priors for achieving performance gains. We use the manually seeded topics provided by Hall et al. [2008] in their study. In order to set the topic-word priors (defined by β in our exposition, or the η parameter in gensim), we initialize matrix *eta* of dimension (n_topics, n_words) to 0. A symmetric prior would be to set all elements to

$$\frac{1}{n\_topics}$$

For every topic for which we have hand picked words, we set

$$eta[topic, word] = 5 \times \frac{1}{n\_topics}$$

For the rest of the words, we set

$$eta[i, j] = \frac{1 - already\_assigned[j]}{n\_topics}$$

where $already\_assigned[j]$ is the sum of the $j-th$ column.

Other parameters

The number of iterations and passes has to be high enough to allow convergence, with the limit of time and computational resources. The number of iterations controls the maximum number of times the E-step (inference) is performed without convergence. If the value is too small, documents will not converge. We monitor the number of documents that converge and set the number of iterations to a maximum of 300. The number of passes controls the number of passes through the corpus. We choose this by monitoring perplexity on the current mini-batch. After 67 passes, there is no change in perplexity (Figure 6.2)

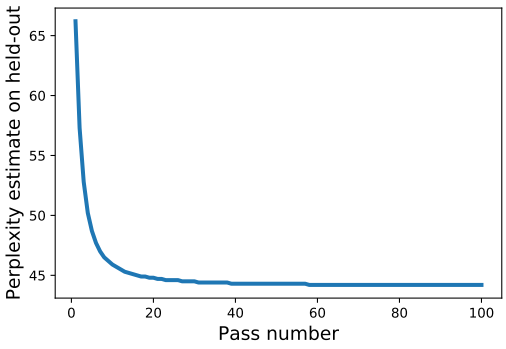

Figure 6.2: LDA Perplexity with 100 topics, 100 passes and 300 maximum E-step iterations

We successfully run LDA on our ACL corpus and now aim to analyse patterns about the topics present in the data. The output we work with consists of the document-topic matrix and the topic-word matrix. The former refers to the topic mixtures of each document, while the latter refers to the mixture of words in each topic. An overview of the process is illustrated in Figure 6.3.

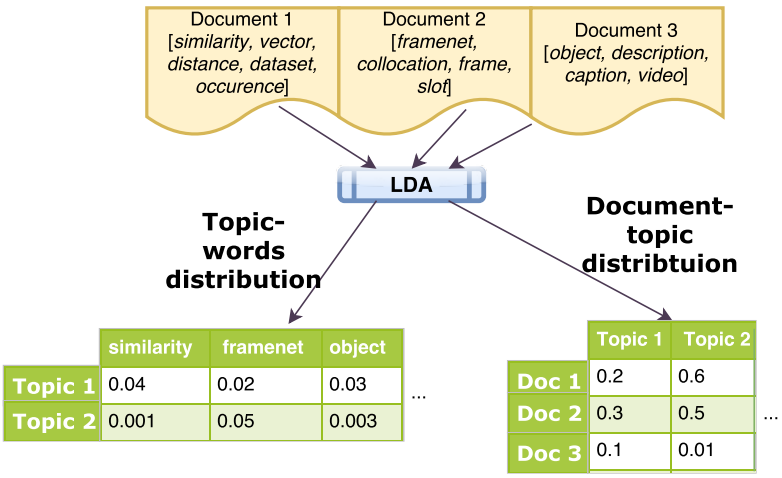

Figure 6.3: Latent Dirichlet Allocation: overview

6.4.2 Labelling topics

Labelling the topics found in the data requires domain knowledge and some topics are easier to label than others. For example, topics like *POS Tagging: [tag, pos, tagger, accuracy, tagging* are easy to recognise from the first few words, but topics like *Multimodal generation :[object, attribute, image, description, visual, scene, game, spatial]* might require more thought. We use the list released by Vogel & Jurafsky [2012] together with their paper, which contains annotated topics and a list of their highest probability words. The most straightforward method would be to find the closest topic for each of our topics, using KL-divergence for discrete probability distributions:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

However, the LDA output we are provided with is incomplete as most words have a listed probability of [0.00], probably due to printing issues. We get around this by using the knowledge that the words are ordered, with the first word having the highest probability etc. We assign the probability of word at position i to be $n\_words - i$ and divide by $\frac{n\_words \times (n\_words+1)}{2}$ for normalization. We can therefore construct a topic-word matrix for the provided labelled topics. Following this methodology, we obtain the correct input for matching our topics with the provided ones. For each topic vector in our topic-word matrix, which represents a probability distribution, we compute KL divergence with each vector from the provided topic-word matrix. We label the topic with the lowest KL divergence.

We inspect the results of this process manually and obtain sensible results, displayed in Table 9.1 (Appendix, Section 9.1). We fix the few misaligned results and find some new topics that were not found in Jurafsky's 2008 analysis, but are detected in our updated corpus: Deep Learning, Social Media Content, Topic Modeling. For most topics, the top 10 words provide a good description of the topics.

6.4.3 Highest probability topics

We compute the probability of each topic over all documents,as given by the document-topic matrix:

$$\begin{aligned} P(z) &= \sum_{\{d \in D\}} P(z|d)P(d) \\ &= \sum_{\{d \in D\}} \frac{P(z|d)}{|D|} \end{aligned}$$

We present the results of topic modeling and the corresponding assigned labels in Table 9.1 in the Appendix. Low quality topics are labelled as *random*, as they contain generic NLP words. The highest probability non-random topics are Probability Theory, Syntax, Discriminative Sequence Models, Statistical Machine Translation, Machine Learning Classification.

6.4.4 Topics by gender

We are interested by the relationship between genders and the topics they prefer. We introduce discrete random variables Z, Y, G which range over topics, year and gender, respectively. We will look at $P(Z|G)$, the probability of a topic given gender and $P(Z|Y,G)$, the probability of a topic given the year and gender. From the document-topic matrix we obtain from running LDA, we are directly given $P(z|d)$.

For each document, we assign to it the gender of its first author, as they are considered to have had the largest contribution. This simply constitutes the rule we use for handling multi-authored papers and there might be other choises available. For a given document d, we use d_G to refer to the gender assigned to it and d_Y for the year of publication.

We compute $P(z|g)$ and $P(z|y,g)$ for each topic:

$$\begin{aligned} P(z|g) &= \sum_{\{d \in D, d_G = g\}} P(z|d,g)P(d|g) \\ &= \sum_{\{d \in D, d_G = g\}} P(z|d)P(d|g) \\ &= \sum_{\{d \in D, d_G = g\}} \frac{P(z|d)}{|\{d \in D, d_G = g|\}} \end{aligned}$$

$$
\begin{aligned}
P(z|g,y) &= \sum_{\{d \in D, d_G = g, d_Y = y\}} P(z|d,g,y)P(d|y,g) \\
&= \sum_{\{d \in D, d_G = g, d_Y = y\}} P(z|d)P(d|y,g) \\
&= \sum_{\{d \in D, d_G = g, d_Y = y\}} \frac{P(z|d)}{|\{d \in D, d_G = g, d_Y = y|\}}
\end{aligned}
$$

We compute our probabilities empirically, from the topic modeling results. For each of the topics, we compute the odds ratio(OR):

$$
OR = \frac{P(z|g = female)(1 - P(z|g = female))}{P(z|g = male)(1 - P(z|g = male))}
$$

The odds ratio is a statistical measure to quantify association between an exposure and an outcome. This metric does not tell us that if one topic is more linked with a gender, the other gender does not publish in that topic. Instead, it simply says that the "odds" are higher for a gender, when adjusting by the total number of publications of that gender.

We display the top male and female topics, as computed with the help of the odds ratio. (Tables 6.2 and 6.3). The overall top male topics are Classic/Dependency Parsing, Probability Theory, NGram Language Models, Finite State Models, Syntatic Trees. The top female published topics are Prosody, Graph theory + Bio NLP, Lexical Acquisition of Verb Subcategarization, Tutoring Systems, Planning/BDI, Dialog.

Figures 6.4 and 6.5 present these topics over time. We note that topics have different times of prominence. The trends of a topic are similar between genders and tend to peak at the same time, reflecting what was happening in the field. For example, for both genders, dialog reached its peak in 1995-2005, but has decreased since, while the topic of dependency parsing has been increasing since 2000.

6.4.5 Topics by conference

In Section 5.7, we find that women are underrepresented at major conferences. We now analyse the most popular topics at these conferences.

To investigate this aspect, we introduce the random variable C, over conferences. We write:

$$
\begin{aligned}
P(z|c) &= \sum_{\{d \in D, d_C = c|\}} P(z|c,g)P(d|c) \\
&= \sum_{\{d \in D, d_C = c\}} P(z|d)P(d|c) \\
&= \sum_{\{d \in D, d_C = c|\}} \frac{P(z|d)}{|\{d \in D, d_C = c|\}}
\end{aligned}
$$

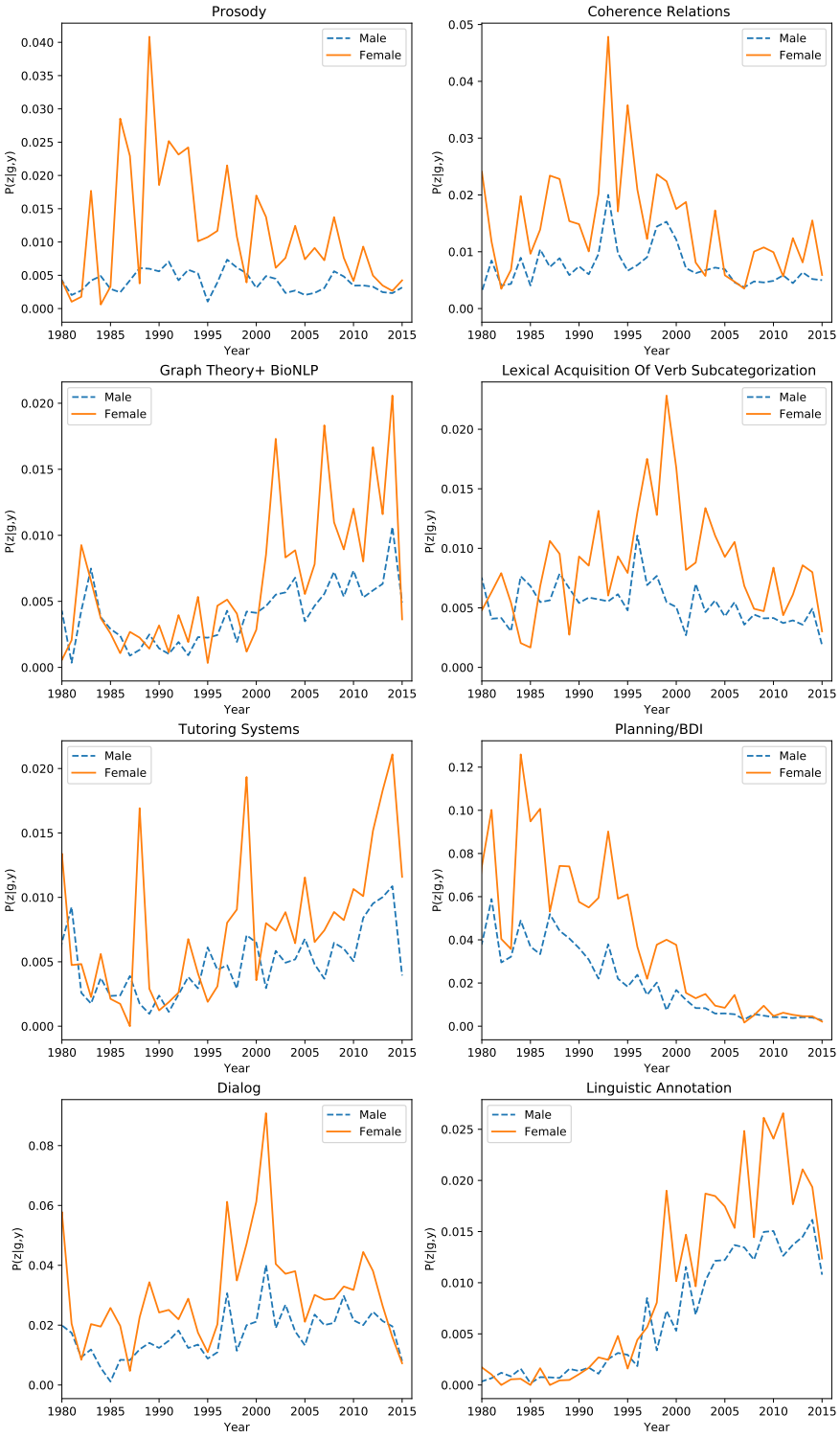

Figure 6.4: Topics with P(topic|female)>P(topic|male). The scale of the y axis differs.

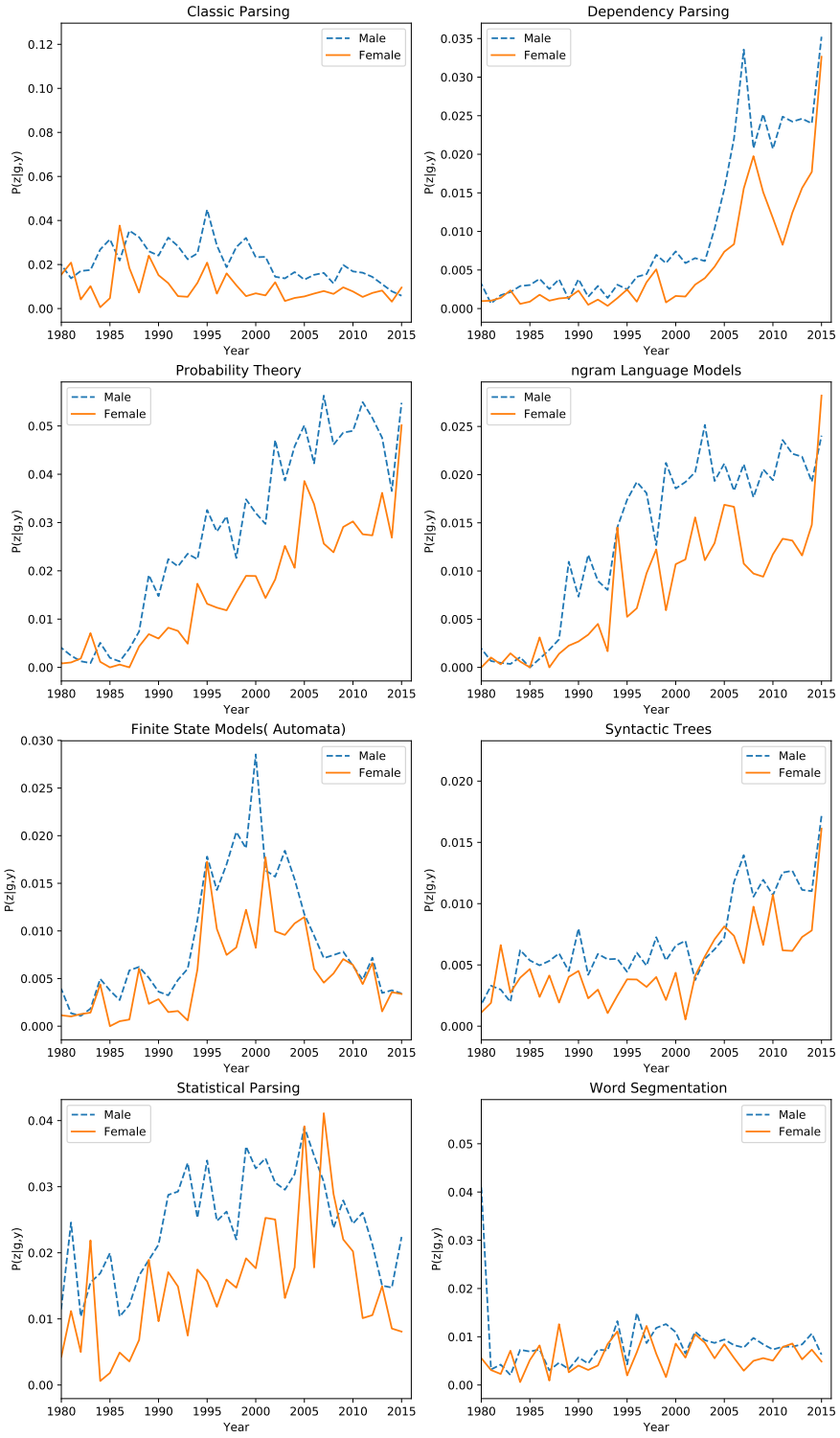

Figure 6.5: Topics with P(topic|male)>P(topic|female). The scale of the y axis differs.

Topic	OR	Top words
1. Prosody	2.40	prosodic game pitch speaker pause accent player tone boundary prosody
2. Coherence Relations	1.86	discourse connective marker coherence clause rhetorical rst causal implicit course
3. Graph Theory+ BioNLP	1.84	medical patient disease record clinical drug health uml symptom field
4. Lexical Acquisition Of Verb Subcategorization	1.72	verb particle object light alternation meaning transitive verbal tense change
5. Tutoring Systems	1.70	student genre write essay style read dialect msa grade readability
6. Planning/BDI	1.60	plan action goal agent act belief speaker proposition utterance planning
7. Dialog	1.56	dialogue user utterance dialog turn act interaction conversation human action
8. Linguistic Annotation	1.55	annotation annotate annotator scheme annotated corpora mark guideline manual automatic
9. Agreement	1.50	agreement annotator human expert judge quality worker agree judgment rating
10. Negation Detection	1.48	cue scope negation simplification token modality uncertainty detection modal speculation
11. Sentiment Analysis	1.48	sentiment negative positive opinion polarity lexicon target subjective neutral expression
12. MUC Era Information Extraction	1.46	template entailment metaphor fill slot object literal rte target analyst
13. TemporalIE/ Aspect	1.40	event temporal trigger expression tense interval date past aspect reference
14. Summarization	1.31	document summary summarization news content article keyword length topic human
15. Question Answering	1.28	question answer response answering passage ask swer candidate yes trec

Table 6.2: Topics women write more about

Similarly to the gender variable, we also condition on the year of the conference, obtaining $P(z|y,c)$. For each conference, we present the topics which give the highest five $P(z|c)$. (Figure 6.4). We also present the odds ratio score we obtained earlier. Machine Learning Classification and Discriminative Sequence Models have values closest to 1, however for the rest of the topics we note that they had lower odds ratios for females. However, we can not say whether the conferences encourage topics that attract a particular gender, or whether they attract a particular gender and this influences their topic composition, as the cause effect relationship is not clear.

Hall et al. [2008] provide some insights regarding the traditional view on the computational linguistics community. COLING is considered to have a wider variety of topics, while ACL is more narrow in scope. Another conference, EMNLP, which started as a smaller and narrower conference, has begun to broaden in recent years. We have also included CoNLL in our exploration, which is a fairly new conference, with the first

Topic	OR	Top words
1. Classic Parsing	0.45	grammar string symbol derivation free terminal let production finite nonterminal
2. Dependency Parsing	0.63	dependency parser parse head tree arc parsing treebank dependent projective
3. Probability Theory	0.64	probability parameter distribution estimate weight sample variable log prior likelihood
4. Ngram Language Models	0.65	gram bigram probability size count trigram unigram entropy frequency estimate
5. Finite State Models(Automata)	0.69	sequence chunk transducer finite transition regular automaton expression fst path
6. Syntactic Trees	0.71	graph edge node path vertex weight connect walk direct label
7. Statistical Parsing	0.72	parse parser grammar constituent treebank parsing tree pars accuracy head
8. Word Segmentation	0.73	character token code letter string length repair sequence abbreviation line
9. Unification Based Grammars	0.73	grammar unification formalism description representation category functional specify constraint specification
10. Statistical Machine Translation	0.76	translation bleu smt target baseline reordering decoder hypothesis reorder decode
11. PPAttachment	0.77	constraint local ambiguity global preference solution ambiguous heuristic attachment straint
12. Tree Adjoining Grammars	0.79	tree node root subtree child label fragment forest leaf parent
13. Categorial Grammar/-Logic	0.80	interpretation meaning representation logical inference john logic expression theory predicate
14. Automata Theory	0.86	match paraphrase edit matching distance substitution string transformation original deletion
15. Deep Learning	0.86	vector representation matrix space layer neural network dimension embedding learn

Table 6.3: Topics men write more about

publications appearing in 2000.

In their study, Hall et al. [2008] show that COLING, EMNLP and ACL started with a narrow focus, but became broader and are catching up with each other. They measure *topic entropy* and find ACL and COLING to be converging in terms of breadth, especially starting with the year 2000. Topic entropy can be computed from the empirical conditional probabilities:

$$H(z|c,y) = -\sum_{i=1}^{K} P(z_i|c,y) \log P(z_i|c,y)$$

Conference	Top Topics	Odds ratio (female)
ACL	1 Probability Theory	0.64
	2 Discriminative Sequence Models	0.91
	3 Statistical Parsing	0.72
	4 Statistical Machine Translation(More Phrase Based)	0.76
	5 Unification Based Grammars	0.73
EMNLP	1 Probability Theory	0.64
	2 Discriminative Sequence Models	0.91
	3 Statistical Machine Translation(More Phrase Based)	0.76
	4 Machine Learning Classification	0.99
	5 Statistical Parsing	0.72
EACL	1 Unification Based Grammars	0.73
	2 Categorial Grammar/ Logic	0.80
	3 Probability Theory	0.64
	4 Classic Parsing	0.45
	5 Discriminative Sequence Models	0.91
COLING	1 Unification Based Grammars	0.73
	2 Categorial Grammar/ Logic	0.80
	3 Classic Parsing	0.45
	4 Probability Theory	0.64
CoNLL	1 Machine Learning Classification	0.99
	2 Discriminative Sequence Models	0.91
	3 Probability Theory	0.64
	4 Dependency Parsing	0.63
	5 Statistical Parsing	0.72

Table 6.4: Topics at conferences, ordered by P(topic|conference)

We update their study regarding topic convergence, to include more recent years of the anthology.In order to do this, we compute the pairwise Jensen-Shannon divergence (JS), a symmetric measure of similarity of two probability distributions. It is defined formally as:

$$D_{JS}(P||Q) = \frac{1}{2}D_{KL}(P||R) + \frac{1}{2}D_{KL}(Q||R)$$
$$R = \frac{1}{2}(P+Q)$$

for two distributions P and Q.

Results are presented in Figure 6.6. It can be observed that the pairs ACL, COLING and ACL, EMNLP are increasing in similarity and their differences are almost absent in recent years. The differences in topics we observed with regards to CoNLL, which is a newer conference are also reflected in its JS Divergence with COLING, higher than for

the other pairs, but still exhibiting a decreasing trend.

The fact that come of these conferences are changing can be regarded as a good sign for diversity.

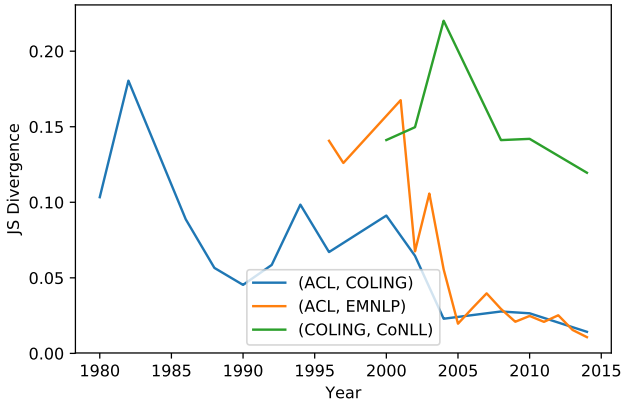

Figure 6.6: JS Divergence between major conferences

6.5 Summary of topics in ACL

We find that there are some different tendencies towards particular genders.

- Males are more likely to write about Classic Parsing, Dependency Parsing, Probability Theory, Ngram Language Models, Finite state models, Syntactic trees, Statistical parsing, Word segmentation
- Females are more likely to write about Prosody, Coherence Relations, Graph Theory and BioNLP, Lexical Acquisition of Verb Subcategorization, Tutoring Systems, Planning, Dialog, Linguistic Annotation
- Topics present at highest ranked conferences are usually topics that male write more about. (Probability Theory, Statistical Parsing, Categorial Grammar). However, these conferences are also becoming more broad, which could indicate a change in the future.

Chapter 7

Publishing in online repositories: arXiv

The ACL corpus is a great way of painting an image of the evolution of the field. However, the dataset lacks the past 4 years (2014 to 2018) in Computation Linguistics, where there has been a big shift towards deep learning methods. We propose analysing a new dataset, to understand this shift. Another reason for this analysis is to better capture the NLP community, without relying too much on just one corpus. The arXiv dataset is different in composition in the sense that publishers do not have to be accepted into a conference to be able to upload their paper. It also means that is is more difficult to draw any causal conclusions, as one gender could be more inclined to have an online presence. This dataset was described in Section 3.2.

The fact that authors publish their papers online has raised multiple concerns. There have been studies that argue in favour of double blind reviews (Budden et al. [2008], Tomkins et al. [2017]),where the reviewers and the authors do not reveal their names. One of their findings is that double blind reviews can even improve the representation of minorities. However, with the rising popularity of arXiv as a medium for authors to distribute their work in progress, there is a clear impact on anonymity. In response of this problem, the ACL policy as of January 1, 2018 [ACL, 1 January 2018] enforces that entries should not have been posted on arXiv while being under review, or within a month prior to the review. The goal is that reviewers should not be able to infer the author of a paper they are reviewing, as this could bias their decision.

There are of course many voices in favour of arXiv, as it promotes open access to publications, while allowing authors to be visible without the competitiveness of highly regarded conferences and journals. In his "Proposal for a new publishing model in Computer Science", LeCun addressed the concerns of people in academia who are worried that the very double blind review process and the highly selective conferences can hinder innovation. In his proposal, he encourages submitting publications to online repositories such as arXiv, as a way to save papers from a cycle of getting rejected and resubmitting to a different conference.

It is therefore very relevant to our study to turn our attention to the online medium and

the differences to the traditional peer-reviewed conferences and journals.

We will start by analysing the gender composition of the Computational Linguistics subsection on arXiv. As the time frame of arXiv publications is much shorter, we will not be able to perform a direct comparison to the authors in the ACL or track the careers of individual academics. We will instead perform an analysis of the topics composition of the dataset.

7.1 Overall statistics

The arXiv dataset covers the 1994-2018 period with a total of 7934 publications and 11918 unique authors. The overall gender distribution is 22% females, 75% male and 3% authors we could not confidently classify. We note that the overall female representation in ACL was much higher, with 26.9% female authors. Similar to the ACL corpus analysis, we fit a regression line of the percentage of authors publishing in a given year (Figure 7.1), starting with 2009, due to the spartsit of the data before this. We find that the percentage has been growing between 2009 and present.

Given the large number of authors compared to their publishing output, we look at the length of their publications record. We find that 67.9% of the authors in our dataset appear in only one publication (with 70.3% of the female authors and 66.5% of the male authors). In the ACL corpus, just over 55% of the authors have just one publication. This result could be due to the fact that authors in the arXiv do not have to be established researchers. It is also the case that not all researchers upload their work to arXiv. Because of these observations, it is not relevant to study the publishing patterns of authors in a similar way as we did with ACL, as they are not necessarily represenative of the community.

7.2 Topics

We employ LDA for finding topics in the arXiv corpus. We note that this analysis is done on abstracts only, which means the data is less noisy, with no tables, figures or citations. Since there are only 20 years of publications, the total number of topics is smaller. We employ the same topic modeling methodology as outlined in Chapter 6.4. Using topic coherence and manual inspection of the topic modeling results, we choose a model with 50 topics.

The results show the rise of deep learning methods in recent years, as well as topics not found in the ACL, such as Reinforcement Learning:

```
human agent learning learn action natural policy
reinforcement system robot
```

An interesting experiment is to match topics between the two datasets. We do this by using KL divergence between each pairs of topics in (arXiv, ACL) and finding the ACL

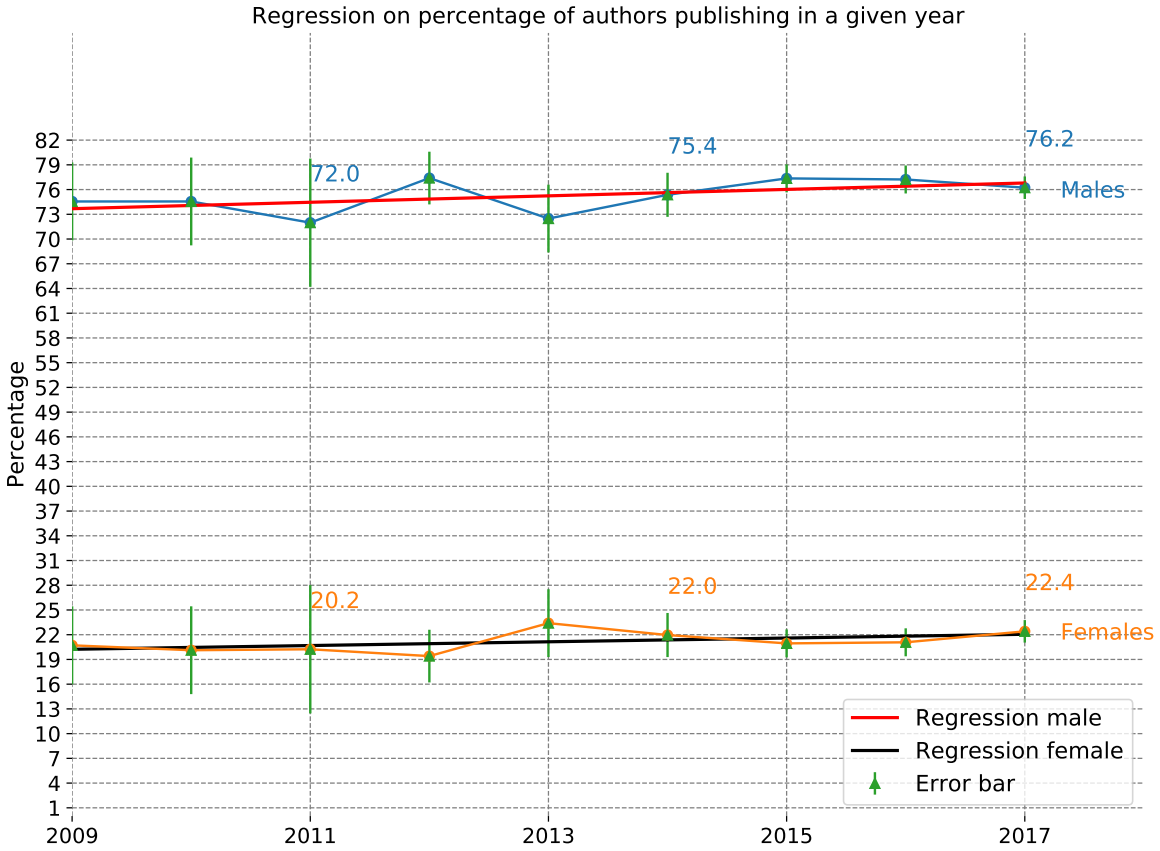

Figure 7.1: Percentage of authors of a given gender publishing at least one paper in a given year in arXiv. The error bar indicates the percentage of authors classified as "unknown" for that year.

topic which is closest. We compare the 50 topics arXiv model with the 100 topics ACL model, for a higher chance of matching all arXiv topics. The size of the vocabulary is different, therefore we have to limit the ACL vocabulary to match that of arXiv, which is smaller. After normalizing the new probability distribution, we obtain a sensible match. The results are presented in Table 9.2, included in the Appendix. As arXiv contains more recent data, it is interesting to note the shift in the most probable words of each topic distribution. For example, the topic of *Neural Networks* changed from:

```
[network layer neural citation hidden cite deep
comparative hide architecture]
```

to

```
[network neural deep recurrent rnn architecture state
lstm task convolutional]
```

outlining the focus on convolutional and recurrent neural networks in deep learning.

We compute the highest probability topics in the arXiv corpus, based on $P(topic)$ as given by the document topic matrix. We note the rise of neural networks and deep learning topics, which had positions 34 and 36 in the ACL, but are now positions 7 and 3 in arXiv. Computing the odds ratio for each gender, we note fewer differences in topic preferences, with 44% of the topics having odds ratio between 0.9 and 1.1, compared to 21% for ACL. Table 7.1 shows the topics with the highest odds ratio for each gender.

Top male	Top Female
System Architectural 0.69	Discourse Coherence 1.64
Algorithm Efficiency 0.71	Multimodal Image Captioning 1.45
Knowledge Representation 0.74	Text Classification 1.45
Speech Recognition 0.78	Human Evaluation 1.41
Semantic structure 0.80	Sentiment Analysis 1.34
Distributional Semantics 0.83	Dialog 1.30

Table 7.1: Top Topics by gender in arXiv

We find that there are differences in the topics preferred by each gender even in the case of an online repository dataset. However, the most interesting result of this chapter is the evolution of topics in the field, with an emphasis on practical and experimental applications (Speech recognition, Image captioning, Sentiment Analysis) which indicate an interesting future for Computational Linguistics.

Chapter 8

Conclusion

8.1 Discussion of the results

We find several differences between the publishing patterns of men and women, most of them with a direct influence on the regarded prestige of authors and the progress of their academic career.

The most worrying result is that in the past 5 years, the percentage of females (27.3%) is below the trend predicted by the regression line and it is actually close to what it was in 1986, when it reached its peak. This is somewhat surprising, as one would expect the gender gap to be slowly closing, with the amount the studies that recognise the existence of the gender gap and the measures taken by several organisations to increase female participation in STEM fields.

We find that men have a higher number of years of activity. This finding could motivate why the most prolific men have a lot more publications that the most prolific female authors. However, we have also found that women author less papers per active year and this could affect their career, as they become less visible. This effect is only present because of the way academic success is regarded: productivity should only be one of the aspects that we consider.

While there is a clear improvement in the percentage of women in first author positions, the proportion of women in last position remains low. We also find that it takes longer, on average, for a female to be credited as last author for the first time. This could be the result of how women interact with their peers and establish authority, and not necessarily an indicative of their merits. In terms of coauthorship, we find that women have on average less coauthors. What is more, 32.05% of their coauthorship pairs are with other females, compared to 24.03% for males coauthoring with females

Women are also underrepresented at the highest ranked Computational Linguistics conferences, which have a female representation below the overage percentage. There are several initiatives that now encourage women to be more active in research, including special workshops at different venues, for example Women in Machine Learning at

NIPS [1] or Women and Underrepresented Minorities in NLP at ACL 2017 [2].

The most likely topics are also different between genders. Men tend to write more about Parsing, Probability Theory, FSMs, while women write more about Prosody, BioNLP, Tutoring Systems, Dialog. There is also a clear shift towards deep learning when it comes to recent publications. In this regard, we find that women and men are as likely to write about this topic, in the context of NLP. We consider these differences to be a positive aspect of gender diversity, as it enriches the field of Computational Linguistics by exploring a broader range of topics.

While there are differences between genders when it comes to their publication output, we argue that with sustained efforts of encouraging equal opportunities and a welcoming environment we can slowly decrease the gender gap and encourage more diverse research.

Why is AI a sea of dudes? Perhaps the answer can be found in the question itself. Given the fact that AI has been "a sea of dudes" from its very beginning, in times where women were not encouraged to pursue careers, it is difficult to break from a tradition with a dominating gender. Fortunately, positive steps have been taken in this direction, with research recognising bias in the data (with the study on word embeddings by Bolukbasi et al. [2016] or the study on facial recognition by Buolamwini & Gebru [2018]) and models trained on unrepresentative datasets giving disastrous results. What this thesis accomplishes is to present a detailed picture of the research output of each gender, revealing several interesting patterns in the data, as way as identifying those experiments that have potential of outputting revealing results.

8.2 Concluding remarks

As there is limited quantitative analysis of women and men's scholarly output in Computational Linguistics (most analysis just reports statistics on the number of graduates), this project fills a gap in our knowledge about gender and publication productivity. We also report on topics in research papers based on gender. Our main contributions are new name classification approaches and new publishing patterns investigations, as well as topic modeling across two different datasets and across conferences.

We presented a pipeline for research applicable to any collection of documents. This included preprocessing a corpus, performing name classification, finding and visualising patterns in the data, identifying high quality topics and their evolution over time. Our preprocessing step included original methods for ensuring the quality of our input. Thus, we checked language coherence by employing language identification tools on parsed PDF files. We corrected failed parsing by employing modern OCR techniques.

We explored the efficiency of different name classification methods. We identified the impossibility of correct classification based on first names only and proposed the

[1] http://wimlworkshop.org/

[2] http://www.winlp.org/winlp-workshop/

need for additional information fed into the classifier. Inspired by character level LSTMs for nationality identification, we implemented this idea for our own task and found promising results. The merit of this approach is that it is able to include information contained in the last name, which often included nationality data. Additionally, we proposed face detection as a way to include imagine information, on top of textual information. Our face detection algorithm includes using search engine results on scholarly websites to retrieve images of authors. Further on, we rely on modern deep learning methods for face landmark extraction and classification. For the purpose of our task, we identify a name as being ambiguous if it is associated with both genders in any of the available census databases and we manually classify this cases. We evaluate the robustness of our name classifier, which achieves 99.2% precision and we use this to classify the genders of the authors in our datasets.

Having performed these preliminary steps, we performed a thorough analysis of our data.
The first part of the study was concerned with scholarly productivity of men and women. We scanned a range of determinants that contribute to an academic career, aiming to place the status of women in the research landscape. We formulated research questions and answered them based on data, with many results being statistically significant. Our study uses all publications in the ACL corpus between 1974-2014. We have looked into various publication patterns, including productivity, position in authors list, coauthorsip and collaboration, dropout rates, presence in highly ranked venues.

The second part of the study explored the topics in Computational Linguistics research and their relation to gender. We introduced the theory behind Latent Dirichlet Allocation and identified the importance of Dirichlet priors for a better coverage of the topics. We set fixed seed words for finding topics that were only introduced later in the dataset. We showed how KL divergence can be used to match our results with the results of other studies. We label all 70 topics in the ACL dataset. We used the probabilities returned by the document-topic matrix for sorting the most probable topics. In order to understand the evolution of topics and their relation to gender, we introduced *year* and *gender* random variables. In order to understand topics in which a gender is more likely to publish, we computed the odds ratio and found clear differences.

Finally, we investigated the differences between a traditional dataset with publications from various conferences(ACL corpus) with an online medium(arXiv corpus). We recognised how the composition of arXiv might be different, due to the fact that anyone is able to choose if they want their publication included. We extracted the topics from arXiv and found that similarity based on KL divergence achieved very good results in terms of matching this output with our ACL output. We were therefore able to offer a side by side comparison of topics in the two corpora. We found new and exciting directions in Computational Linguistics, with the shift towards deep learning methods and even the introduction of new topics, such as Reinforcement Learning.

8.3 Further work

There are many other interactions in the data that could be analysed. We could look at citations information to test whether a gender tends to cite the other gender more. However, extraction citations from PDF files is prone to mistakes. This step would be easier to perform with the help of Google Scholar, which records outgoing citations of an author.

Further research could be pursued in the direction of face classificatin methods, which showed to be promising. In a future iteration, we would replace our Microsoft Azure based face recognition system with a cheaper system, that would also return confidence scores for gender classification. As pretrained networks with large datasets of faces are available [3], we could build our own face classifier.

There are also other topic models that could be employed. Author topic models [Rosen-Zvi et al., 2012] could be used to identify the topics of a single author over time.

Potentially, there could be interesting observations with regards to the style of writing of different authors. According to sociolinguistics theories [Lakoff, 1975], there are differences between the style of women's and men's writing, in terms of lexical choices, linguistic cues and discourse behaviour. However, since academic papers tend to use formal language, lexical cues that are easiest to detect are suppressed. Stylometric studies usually look at a combination of bag of words (BOW), syntax (Context Free Grammar rules) and style (punctuation, stop words, Latin abbreviation). Since we already have the documents of each author, this could be a potential straightforward extension to be made.

[3] http://www.cbsr.ia.ac.cn/english/CASIA-WebFace-Database.html

Chapter 9

Appendix

9.1 ACL topics

Table 9.1: ACL 70 topics sorted by P(topic), top 10 words

1.**random(commonwords)**	suggest look change issue come clear view involve expect kind
2.**Probability Theory**	probability parameter distribution estimate weight sample variable log prior likelihood
3.**random(System Architectural)**	user tool module component interface file database design project architecture
4.**Syntax**	clause verb subject head object construction preposition modifier noun complement
5.**Discriminative Sequence Models**	label learn baseline learning supervised dataset crf unlabelled unsupervised target
6.**random(misc)**	computer university computational program proceeding science technology project course conference
7.**Statistical Machine Translation**	translation bleu smt target baseline reordering decoder hypothesis reorder decode
8.**Machine Learning Classification**	classifier classification accuracy svm kernel learn classify learning decision binary
9.**Statistical Parsing**	parse parser grammar constituent treebank parsing tree pars accuracy head
10.**Dialog**	dialogue user utterance dialog turn act interaction conversation human action
11.**Unification Based Grammars**	grammar unification formalism description representation category functional specify constraint specification
12.**Categorial Grammar/ Logic**	interpretation meaning representation logical inference john logic expression theory predicate
13.**Bilingual Word Alignment**	translation alignment target parallel align bilingual translate corpora monolingual transfer

14.**Multi Lingual Resources**	transliteration hindi korean russian script urdu japanese name hybrid thai
15.**Clustering+ Distributional Similarity**	similarity vector distance dataset compute distributional cosine weight occurrence space
16.**Summarization**	document summary summarization news content article keyword length topic human
17.**random**	ion t ica ie in i par descr con iona
18.**Named Entity Recognition**	entity name person extraction ner location organization recall precision mention
19.**Anaphora Resolution**	mention coreference chain resolution entity link ace pronoun nps document
20.**ngram Language Models**	gram bigram probability size count trigram unigram entropy frequency estimate
21.**Speech Recognition**	recognition speaker asr rate error acoustic spoken recognizer phone transcription
22.**Classic Parsing**	grammar string symbol derivation free terminal let production finite nonterminal
23.**Document Retrieval**	query web retrieval document page collection retrieve engine relevant user
24.**Dependency Parsing**	dependency parser parse head tree arc parsing treebank dependent projective
25.**Metrics**	metric human rank reference correlation quality ranking translation automatic average
26.**random(misc)**	item stack right reduce action strategy memory cost left operation
27.**Word Sense Disambiguation**	sense sens disambiguation wordnet noun wsd target disambiguate ambiguous grain
28.**Tree Adjoining Grammars**	tree node root subtree child label fragment forest leaf parent
29.**Collocations Measures**	candidate frequency precision occurrence mwe expression collocation filter association recall
30.**Planning/BDI**	plan action goal agent act belief speaker proposition utterance planning
31.**SRL/ Framenet**	frame slot framenet filler element fill target collocation subcategorization mutual
32.**Multimodal(Mainly Generation)**	object description scene expression spatial gesture action visual video reference
33.**Linguistic Annotation**	annotation annotate annotator scheme annotated corpora mark guideline manual automatic
34.**Neural Networks/ Human Cognition**	prediction predict learn child learning learner effect hypothesis acquisition predictor
35.**Spell Correction**	error correction detection detect rate spelling preposition incorrect learner recall
36.**Deep Learning**	vector representation matrix space layer neural network dimension embedding learn

37.**Chinese KoreanNLP**	chinese character segmentation oov candidate respectively china adopt denote nese
38.**POS Tagging**	tag pos tagger accuracy tagging token tagset unknown sequence hmm
39.**Agreement**	agreement annotator human expert judge quality worker agree judgment rating
40.**Morphology**	morphological arabic stem suffix morpheme morphology root prefix affix analyser
41.**Dictionary Lexicons**	dictionary lexicon entry definition database construct lexeme vocabulary meaning coverage
42.**Question Answering**	question answer response answering passage ask swer candidate yes trec
43.**Relation Extraction**	pattern seed post extraction learn thread precision tern bootstrapping match
44.**Sentiment Analysis**	sentiment negative positive opinion polarity lexicon target subjective neutral expression
45.**PP Attachment**	constraint local ambiguity global preference solution ambiguous heuristic attachment straint
46.**Syntactic Trees**	graph edge node path vertex weight connect walk direct label
47.**Automata Theory**	match paraphrase edit matching distance substitution string transformation original deletion
48.**TemporalIE/ Aspect**	event temporal trigger expression tense interval date past aspect reference
49.**Word Segmentation**	character token code letter string length repair sequence abbreviation line
50.**Coherence Relations**	discourse connective marker coherence clause rhetorical rst causal implicit course
51.**MUC Era Information Extraction**	template entailment metaphor fill slot object literal rte target analyst
52.**Topic Modeling**	topic document lda distribution author latent dirichlet topical draw sample
53.**Finite State Models(Automata)**	sequence chunk transducer finite transition regular automaton expression fst path
54.**Discourse Segmentation**	segment segmentation unit boundary block length sequence recall segmenter comma
55.**Clustering**	cluster clustering group induce merge induction ter unsupervised partition similarity
56.**Natural Language Generation**	generation selection generator surface choice nlg compression realization ilp ordering
57.**Biomedical Named Entity Recognition**	protein gene extraction entity biomedical cell interaction genia name patent
58.**Concept Ontologies/ Knowledge Rep**	network hierarchy taxonomy relationship hypernym node hierarchical definition hyponym activation
59.**Tutoring Systems**	student genre write essay style read dialect msa grade readability

60.**Multilingual Ontologies**	concept ontology image conceptual representation triple property object caption cept
61.**Graph Theory+ BioNLP**	medical patient disease record clinical drug health uml symptom field
62.**Social Media**	tweet emotion social user twitter medium blog post emotional hashtag
63.**Recommender System**	review aspect product movie book rating restaurant user food price
64.**Wordnet**	link wordnet attribute resource synset synonym anchor gloss mapping net
65.**Lexical Acquisition Of Verb Subcategorization**	verb particle object light alternation meaning transitive verbal tense change
66.**Computational Phonology**	vowel syllable phoneme consonant phonetic stress sound phonological pronunciation grapheme
67.**random(Pronouns Commonnouns)**	message story email city location day send region actor zone
68.**Prosody**	prosodic game pitch speaker pause accent player tone boundary prosody
69.**Web Search+ Wikipedia**	category article wikipedia compound page hedge title categorization wiki split
70.**Negation Detection**	cue scope negation simplification token modality uncertainty detection modal speculation

9.2 ArXiv topics

Table 9.2: Top 10 words: arXiv 50 topics sorted by P(topic|documents) and their best match in the ACL topics using KL divergence

arXiv: 1 Semantic Structure	grammar semantic structure theory context natural linguistic present constraint logic
ACL: Categorial Grammar/ Logic	definition variable proof logic formula property let condition description axiom
arXiv: 2 Word Sense Disambiguation	word lexical rule corpus lexicon dictionary present sense morphological resource
ACL: Word Sense Disambiguation	sense sens wordnet synset wsd disambiguation target hypernym definition gloss
arXiv: 3 Deep Learning	network neural deep recurrent rnn architecture state lstm task convolutional
ACL: Neural Networks/ Human Cognition	network layer neural citation hidden cite deep comparative hide architecture

arXiv: 4 Machine Translation	translation machine english nmt system source neural target parallel sentence
ACL: Machine Translation(Non Statistical+ Bitexts)	translation target translate parallel bilingual corpora monolingual spanish french lingual
arXiv: 5 random	datum training task method learning train performance learn propose approach
ACL: Discriminative Sequence Models	label learn learning supervised unlabelled parameter weight baseline dataset iteration
arXiv: 6 Information Retrieval	information query datum system retrieval search web tool processing research
ACL: Web Search+ Wikipedia	resource tool file project format xml element field record database
arXiv: 7 Neural Networks	attention propose sentence sequence neural mechanism task network encoder memory
ACL: Neural Networks/ Human Cognition	network layer neural citation hidden cite deep comparative hide architecture
arXiv: 8 Temporal Information Extraction	time study linguistic change analysis find human different pattern statistical
ACL: Collocations Measures	frequency occurrence effect association corpora distribution size count item average
arXiv: 9 Word embeddings	word embedding learn vector task representation method embed train context
ACL: Clustering+ Distributional Similarity	vector space matrix representation dimension embedding similarity lsa dimensional compute
arXiv: 10 Distributional Similarit	similarity word method measure semantic distance vector approach propose matrix
ACL: Clustering+ Distributional Similarity	vector space matrix representation dimension embedding similarity lsa dimensional compute
arXiv: 11 Algorithmic Efficiency	algorithm parameter search number efficient approach result technique method problem
ACL: Algorithmic Efficiency	weight lattice prune decoding beam decode size compute space span

arXiv: 12 Machine Learning Classification	feature classification classifier text class machine result tweet performance accuracy
ACL: Machine Learning Classification	classifier classification accuracy baseline learn learning decision classify predict prediction
arXiv: 13 Dependency Parsing	sentence tree parser dependency parse structure syntactic parsing treebank transition
ACL: Dependency Parsing	dependency parser parse treebank head tree parsing accuracy label constituent
arXiv: 14 Distributional Similarity	representation semantic vector learn space word distribute structure method task
ACL: Clustering+ Distributional Similarity	vector space matrix representation dimension embedding similarity lsa dimensional compute
arXiv: 15 Knowledge Representation	knowledge learn reasoning ontology natural inference approach structured learning method
ACL: Formal Computational Semantics	representation logical triple ccg learn meaning logic parse composition derivation
arXiv: 16 Reinforcement Learning	human agent learning learn action natural policy reinforcement system robot
ACL: UI/ Natural Language Interface	participant action interaction policy subject human instruction condition learn control
arXiv: 17 ASR	speech recognition speaker acoustic system automatic asr error signal feature
ACL: Speech Recognition	recognition speaker rate error acoustic phone recognizer adaptation hmm vocabulary
arXiv: 18 Collocations Measures	word distribution frequency probability law information length number estimate size
ACL: Collocations Measures	frequency occurrence effect association corpora distribution size count item average
arXiv: 19 Speech Recognition	sequence end train system training recognition speech word loss rate

ACL: Speech Recognition	recognition speaker rate error acoustic phone recognizer adaptation hmm vocabulary
arXiv: 20 Linguistic Annotation	system task arabic describe evaluation set result test challenge share
ACL: Linguistic Annotation	arabic msa abbreviation dialect habash morphological dialectal write bic diacritic
arXiv: 21 random(Social Media)	social medium user news online post community content identify forum
ACL: random(Pronouns Commonnouns)	post tweet social twitter worker thread medium people hit online
arXiv: 22 POS Tagging	tag feature task cluster pos nlp word tagging hand art
ACL: POS Tagging	tag pos tagger accuracy tagging tagset noun unknown treebank hmm
arXiv: 23 Lexical Acquisition Of Verb Subcategorization	annotation verb argument semantic structure pattern role subject pronoun resolution
ACL: Lexical Acquisition Of Verb Subcategorization	verb subject object construction noun verbal light preposition passive transitive
arXiv: 24 Dialog	dialogue system dialog conversation response human generation utterance evaluation task
ACL: Dialog	dialogue utterance act turn conversation speaker gesture human meeting conversational
arXiv: 25 Document Retrieval	document summarization summary sentence article story text approach method scientific
ACL: Document Retrieval	document article news keyword collection paragraph title content extraction relevant
arXiv: 26 Sentiment Analysis	sentiment analysis review opinion aspect emotion polarity negative positive product
ACL: Sentiment Analysis	negative sentiment positive polarity lexicon tweet neutral classification twitter label
arXiv: 27 Text Categorization	text corpus write method movie style generate content book author

ACL: Text Categorization	author genre style book email write quote stylistic zone novel
arXiv: 28 Relation Extraction	relation extraction graph pair method temporal approach extract task knowledge
ACL: Relation Extraction	pattern extraction seed learn bootstrapping tern acquire automatically precision acquisition
arXiv: 29 random	problem vqa research machine dataset provide different natural result processing
ACL: Machine Learning Classification	classifier classification accuracy baseline learn learning decision classify predict prediction
arXiv: 30 Topic Models	topic latent document modeling variable lda distribution method generative propose
ACL: Topic Models	topic document lda distribution topical latent blei dirichlet collection probability
arXiv: 31 Discriminative Sequence Models	label dataset text approach supervised semi datum detection large report
ACL: Discriminative Sequence Models	label learn learning supervised unlabelled parameter weight baseline dataset iteration
arXiv: 32 Chinese word segmentation	word character level chinese segmentation segment unit sequence result japanese
ACL: Chinese KoreanNLP	chinese character segmentation unknown china respectively adopt nese dictionary sequence
arXiv: 33 Question Answering	question answer answering dataset system task art state performance large
ACL: Sentiment Analysis	question answer answering passage ask swer trec factoid return expect
arXiv: 34 Language Generation	paraphrase learn large approach sequence datum result generation source dataset
ACL: Discriminative Sequence Models	label learn learning supervised unlabelled parameter weight baseline dataset iteration
arXiv: 35 random	user error item level approach system datum recommendation learning propose

ACL: UI/ Natural Language Interface	user profile response item request interaction option interactive expert help
arXiv: 36 Discourse Coherehce	discourse text approach stance coherence target identify non present classification
ACL: Machine Learning Classification	classifier classification accuracy baseline learn learning decision classify predict prediction
arXiv: 37 System Architectural	system set technique multilingual graph method present knowledge context datum
ACL: random(System Architectural)	component interface user tool design architecture software display access development
arXiv: 38 Web Search+ Wikipedia	attribute category analysis author method political semantic wikipedia title study
ACL: Web Search+ Wikipedia	category link wikipedia article anchor page categorization wiki title dataset
arXiv: 39 System architectural	code software source toolkit feature implementation describe function open design
ACL: random(System architectural)	component interface user tool design architecture software display access development
arXiv: 40 Metrics+ Human Evaluation	video method semantic precision property recall problem matching good correlation
ACL: Metrics+ Human Evaluation	metric human reference quality correlation rank judgment automatic judge ranking
arXiv: 41 Multimodal Image Captioning	image visual description object caption generate dataset multimodal task generation
ACL: Multimodal(Mainly Generation)	object attribute image description visual scene game spatial video expression
arXiv: 42 Collocations Measures	phrase noun word association method sentence prosodic experiment run result
ACL: Collocations Measures	frequency occurrence effect association corpora distribution size count item average
arXiv: 43 Anaphora Resolution	event system coreference mention detection detect narrative emotion type speaker

ACL: Anaphora Resolution	mention coreference chain entity resolution link ace head muc document
arXiv: 44 Automated Essay Scoring	domain specific adaptation adapt component method general spelling target system
ACL: Spell Correction	error correction edit spelling learner preposition rate detect incorrect detection
arXiv: 45 UI/ Natural Language Interface	user twitter preference health tweet spatial location term prediction public
ACL: UI/ Natural Language Interface	user profile response item request interaction option interactive expert help
arXiv: 46 System Architectural	normalization task different propose author passage additional work text information
ACL: random(System Architectural)	component interface user tool design architecture software display access development
arXiv: 47 random	concept message medical emoji gender service web email hashtag filter
ACL: random(Pronouns Commonnouns)	post tweet social twitter worker thread medium people hit online
arXiv: 48 Named Entity Recognition	entity name recognition ner knowledge link mention type graph information
ACL: Named Entity Recognition	entity name person location ner organization recognition proper ne org
arXiv: 49 Multimodality	modality information modal semantic incremental different audio datum process multimodal
ACL: Dialog	dialogue utterance act turn conversation speaker gesture human meeting conversational
arXiv: 50 Computational Phonology	phoneme synthesis system phonological pronunciation voice phonetic customer syllable compression
ACL: Computational Phonology	vowel syllable phoneme letter consonant stress phonetic pronunciation phonological sound

Bibliography

AAUP. Faculty gender equity indicators. 2006. URL https://www.aaup.org/reports-publications/publications/see-all/aaup-faculty-gender-equity-indicators-2006.

ACL. ACL Policies for Submission, Review and Citation. 1 January 2018. URL https://www.aclweb.org/adminwiki/index.php?title=ACL_Policies_for_Submission,_Review_and_Citation.

Anderson, Ashton, McFarland, Dan, and Jurafsky, Dan. Towards a Computational History of the ACL: 1980-2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, ACL '12, pp. 13–21, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

Athena Swan Award. Diversity at The University of Edinburgh. 2016. URL https://www.ed.ac.uk/informatics/about/work-with-us/equality-diversity/athena-swan-award.

Barthauer, L., SpurkD., and Kauffeld, S. Women's social capital in academia: A personal network analysis. *International Review of Social Research*, 6:195–205, 2016.

BBC. Microsoft chatbot is taught to swear on twitter. 2016. URL http://www.bbc.com/news/technology-35890188.

Blei, David. Probabilistic topic models. *ICML Tutorial*, 2012. URL http://www.cs.columbia.edu/~blei/talks/Blei_ICML_2012.pdf.

Blei, David M. and Lafferty, John D. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 113–120, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

Bloomberg. Artificial intelligence has a sea of dudes problem. 2016. URL https://www.bloomberg.com/news/articles/2016-06-23/artificial-intelligence-has-a-sea-of-dudes-problem.

Bolukbasi, Tolga, Chang, Kai-Wei, Zou, James Y., Saligrama, Venkatesh, and Kalai, Adam. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520, 2016.

Budden, Amber E., Tregenza, Tom, Aarssen, Lonnie W., Koricheva, Julia, Leimu, Roosa, and Lortie, Christopher J. Double-blind review favours increased representation of female authors. *Trends in Ecology Evolution*, 23(1):4 – 6, 2008. ISSN 0169-5347.

Buolamwini, Joy and Gebru, Timnit. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, Sorelle A. and Wilson, Christo (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

Chang, Jonathan, Gerrish, Sean, Wang, Chong, Boyd-graber, Jordan L., and Blei, David M. Reading tea leaves: How humans interpret topic models. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 22*, pp. 288–296. Curran Associates, Inc., 2009.

CRA. CRA Taulbee Survey. 2016. URL https://cra.org/resources/taulbee-survey/.

Das, Rajarshi, Zaheer, Manzil, and Dyer, Chris. Gaussian lda for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2015.

DeFrancisco, V.L., DeFrancisco, V.P., Palczewski, C.H., and McGeough, D.D. *Gender in Communication*. SAGE Publications, 2013. ISBN 9781452220093. URL https://books.google.co.uk/books?id=lb6hAQAAQBAJ.

Google Scholar. Computational Linguistics Conferences. Accessed 21 March 2018. URL https://scholar.google.co.uk/citations?view_op=top_venues&hl=en&vq=eng_computationallinguistics.

Griffiths, T. L. and Steyvers, M. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April 2004.

Griffiths, Thomas L., Steyvers, Mark, and Tenenbaum, Joshua B. Topics in semantic representation. *Psychological review*, 114 2:211–44, 2007.

Guardian, The. Google photos labels black people as 'gorillas'. 2015. URL http://www.telegraph.co.uk/technology/google/11710136/Google-Photos-assigns-gorilla-tag-to-photos-of-black-people.html.

Hall, David, Jurafsky, Daniel, and Manning, Christopher D. Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pp. 363–371, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

Hirsch, J. E. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, 2005.

Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Hoffman, Matthew D., Blei, David M., and Bach, Francis. Online learning for latent dirichlet allocation. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, pp. 856–864, USA, 2010. Curran Associates Inc.

Karimi, Fariba, Wagner, Claudia, Lemmerich, Florian, Jadidi, Mohsen, and Strohmaier, Markus. Inferring gender from names on the web: A comparative evaluation of gender detection methods. *CoRR*, abs/1603.04322, 2016.

Lakoff, R.T. *Language and woman's place*, volume 2 of *Harper colophon books*. Harper & Row, 1975.

LeCun, Yann. Proposal for a new publishing model in Computer Science. Accessed 18 March 2018. URL http://yann.lecun.com/ex/pamphlets/publishing-models.html.

Lee, Jinhyuk, Kim, Hyunjae, Ko, Miyoung, Choi, Donghee, Choi, Jaehoon, and Kang, Jaewoo. Name nationality classification with recurrent neural networks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 2081–2087, 2017.

Liu, W and Ruths, D. What's in a name? using first names as features for gender inference in twitter. pp. 10–16, 01 2013.

Liu, Yang, Liu, Zhiyuan, Chua, Tat-Seng, and Sun, Maosong. Topical word embeddings. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pp. 2418–2424. AAAI Press, 2015. ISBN 0-262-51129-0.

Lui, Marco and Baldwin, Timothy. Cross-domain feature selection for language identification. In *In Proceedings of 5th International Joint Conference on Natural Language Processing*, pp. 553–561, 2011.

Manning, Christopher D. Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707, 2015.

Manning, Christopher D., Raghavan, Prabhakar, and Schütze, Hinrich. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2008. ISBN 978-0-521-86571-5.

Mihaljević-Brandt, Helena, Santamaría, Lucía, and Tullney, Marco. The effect of gender in the publication patterns in mathematics. *PLOS ONE*, 11(10):1–23, 10 2016.

Mimno, David, Wallach, Hanna M., Talley, Edmund, Leenders, Miriam, and McCallum, Andrew. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp. 262–272, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-11-4.

Moss-Racusin, Corinne A., Dovidio, John F., Brescoll, Victoria L., Graham, Mark J., and Handelsman, Jo. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41):16474–16479, 2012. ISSN 0027-8424.

National Science Foundation. Women, minorities, and persons with disabilities in science and engineering: 2017. *Special Report NSF 17-310*, 2017. URL www.nsf.gov/statistics/wmpd/0.

Newman, David, Lau, Jey Han, Grieser, Karl, and Baldwin, Timothy. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pp. 100–108, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5.

Niu, Liqiang and Dai, Xin-Yu. Topic2vec: Learning distributed representations of topics. *CoRR*, abs/1506.08422, 2015.

Radev, Dragomir R., Muthukrishnan, Pradeep, Qazvinian, Vahed, and Abu-Jbara, Amjad. The acl anthology network corpus. *Language Resources and Evaluation*, pp. 1–26, 2013. ISSN 1574-020X.

Řehůřek, Radim and Sojka, Petr. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

Report prepared by Laboratory for Computer Science and the Artificial Intelligence Laboratory at MIT. *Barriers to equality in academia: women in computer science at M.I.T.* M.I.T., 1983. URL https://books.google.co.uk/books?id=wm5ZAAAAYAAJ.

Rosen-Zvi, Michal, Griffiths, Thomas L., Steyvers, Mark, and Smyth, Padhraic. The author-topic model for authors and documents. *CoRR*, abs/1207.4169, 2012.

Ruxton, Graeme D. The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral Ecology*, 17(4):688–690, 2006.

Schäfer, Ulrich, Read, Jonathon, and Oepen, Stephan. Towards an ACL Anthology Corpus with Logical Document Structure: An Overview of the ACL 2012 Contributed Task. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, ACL '12, pp. 88–97, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

Shi, Bei, Lam, Wai, Jameel, Shoaib, Schockaert, Steven, and Lai, Kwun Ping. Jointly learning word embeddings and latent topics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pp. 375–384, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5022-8.

Sievert, C and Shirley, K.E. Ldavis: A method for visualizing and interpreting topics. pp. 63–70, 01 2014.

Smith, R. An overview of the tesseract ocr engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02*, ICDAR '07, pp. 629–633, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 0-7695-2822-8.

Solomon, Justin. Programmers, professors, and parasites: Credit and co-authorship in computer science. 15:467–89, 03 2009.

Tang, Cong, W. Ross, Keith, Saxena, Nitesh, and Chen, Ruichuan. What's in a name: A study of names, gender inference and gender behavior in facebook, 04 2011.

Teh, Yee Whye, Jordan, Michael I, Beal, Matthew J, and Blei, David M. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Tomkins, Andrew, Zhang, Min, and Heavlin, William D. Single versus double blind reviewing at WSDM 2017. *CoRR*, abs/1702.00502, 2017.

UCAS. 2014. URL https://www.ucas.com/.

Vogel, Adam and Jurafsky, Dan. He said, she said: Gender in the acl anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, ACL '12, pp. 33–41, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

Wallach, Hanna M., Mimno, David M., and McCallum, Andrew. Rethinking lda: Why priors matter. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 22*, pp. 1973–1981. Curran Associates, Inc., 2009a.

Wallach, Hanna M., Murray, Iain, Salakhutdinov, Ruslan, and Mimno, David. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 1105–1112, New York, NY, USA, 2009b. ACM. ISBN 978-1-60558-516-1.

Washington Post. Why men get all the credit when they work with women. Accessed 27 February 2018. URL https://www.washingtonpost.com/news/wonk/wp/2015/11/13/why-men-get-all-the-credit-when-they-work-with-women/?utm_term=.7179248dc365.

West, Jevin D., Jacquet, Jennifer, King, Molly M., Correll, Shelley J., and Bergstrom, Carl T. The role of gender in scholarly authorship. *PLOS ONE*, 8(7):1–6, 07 2013.

Zdenka, Šadl. 'We Women Are No Good at It': Networking in Academia. *Sociologický Časopis / Czech Sociological Review*, 45(6):1239–1263, 2009.