He Said, She Said: Gender in the ACL Anthology

Adam Vogel
Stanford University
av@cs.stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Abstract

Studies of gender balance in academic computer science are typically based on statistics on enrollment and graduation. Going beyond these coarse measures of gender participation, we conduct a fine-grained study of gender in the field of Natural Language Processing. We use topic models (Latent Dirichlet Allocation) to explore the research topics of men and women in the ACL Anthology Network. We find that women publish more on dialog, discourse, and sentiment, while men publish more than women in parsing, formal semantics, and finite state models. To conduct our study we labeled the gender of authors in the ACL Anthology mostly manually, creating a useful resource for other gender studies. Finally, our study of historical patterns in female participation shows that the proportion of women authors in computational linguistics has been continuously increasing, with approximately a 50% increase in the three decades since 1980.

1 Introduction

The gender imbalance in science and engineering is particularly striking in computer science, where the percentage of graduate students in computer science that are women seems to have been declining rather than increasing recently (Palma, 2001; Beaubouef and Zhang, 2011; Spertus, 1991; Hill et al., 2010; Singh et al., 2007).

While many studies have examined enrollment and career advancement, less attention has been paid to gender differences in scientific publications. This paper studies author gender in the Association for Computational Linguistics Anthology Network (AAN) corpus (Radev et al., 2009), (based on the ACL Anthology Reference Corpus (Bird et al., 2008)) from which we used 13,000 papers by approximately 12,000 distinct authors from 1965 to 2008.

The AAN corpus disambiguates author names, but does not annotate these names for gender. We first performed a mostly-manual annotation of the gender of each author (details in Section 2). We make these annotation available as a useful resource for other researchers.[1]

We then study a number of properties of the ACL authors. We first address surface level questions regarding the balance of genders in publications. In 2008, women were granted 20.5% of computer science PhDs (CRA, 2008). Does this ratio hold also for the percentages of papers written by women in computational linguistics as well? We explore differences in publication count between genders, looking at total publications and normalized values like publications per year and trends over time.

Going beyond surface level analysis, we then turn to document content. We utilize Latent Dirichlet Allocation (LDA) topic models (Blei et al., 2003) to study the difference in topics that men and women write about.

2 Determining Gender

The gender of an author is in general difficult to determine automatically with extremely high precision. In many languages, there are gender-differentiated names for men and women that can make gender-assignment possible based on gendered name dictionaries. But the fact that ACL authors come from many different language background makes this method prone to error. For example, while U.S. Census lists of frequently occurring names by gender (Census, 2012) can

[1] http://nlp.stanford.edu/projects/gender.shtml

resolve a large proportion of commonly occurring names from authors in the United States and Canada, they incorrectly list the name "Jan" as female. It turns out that authors in the ACL Anthology who are named "Jan" are in fact male, since the name is a very common male name in many parts of Europe, and since US female researchers named "Jan" often use the full form of their name rather than the shortening "Jan" when publishing. Furthermore, a significant percentage of ACL authors have Chinese language names, which are much less clearly linked with personal names (e.g., Weiwei Sun is female whereas Weiwei Ding is male).

We found that Chinese names as well as ambiguous names like "Jan" were poorly predicted by online name gender website algorithms we looked at, leading to a high error rate. To insure high precision, we therefore instead chose to annotate the authors in the corpus with a high-precision method; mainly hand labeling the names but also using some automatic help.

We used unambiguous name lists for various languages to label a large proportion of the name; for example we used the subset of given names (out of the 4221 first names reported in the 1990 U.S. Census) that were unambiguous (occurring consistently with only one gender in all of our name lists) used morphological gender for languages like Czech or Bulgarian which mark morphological gender on names, and relied on lists of Indian and Basque names (from which we had removed any ambiguous names). For all ambiguous names, we next used our personal cognizance of many of the ACL authors, also asking for help from ACL researchers in China, Taiwan, and Singapore (to help label Chinese names of researchers they were familiar with) and other researchers for help on the Japanese and Korean names. Around 1100 names were hand-labeled from personal cognizance or photos of the ACL researchers on their web pages. The combination of name lists and personal cognizance left only 2048 names (15% of the original 12,692) still unlabeled. We then used a baby name website, `www.gpeters.com/names/`, originally designed for reporting the popularity and gender balance of first names, to find the gender of 1287 of these 2048 names.[2] The remaining 761 names remained unlabeled.

[2]The gender balance of these 1287 automatically-determined names was 34% female, 66% male, slightly higher than the average for the whole corpus.

	Total		First Author	
Gender	Papers	%	Papers	%
Female	6772	33%	4034	27%
Male	13454	64%	10813	71%
Unknown	702	3%	313	2%

Table 1: Number of publications by gender. The total publications column shows the number of papers for which at least one author was a given gender, in any authorship position. The first authored publications column shows the number of papers for which a given gender is the first author.

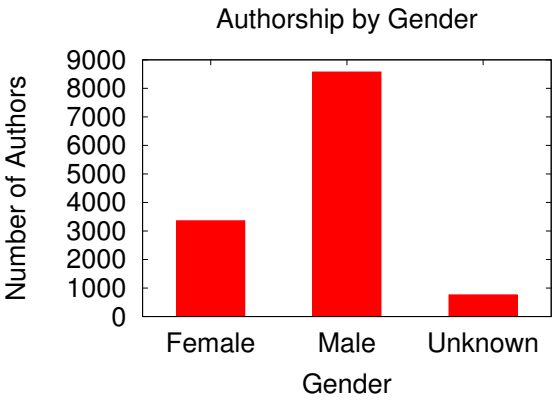

Figure 1: The total number of authors of a given gender.

3 Overall Statistics

We first discuss some overall gender statistics for the ACL Anthology. Figure 1 shows the number of authors of each gender. Men comprised 8573 of the 12692 authors (67.5%) and there were 3359 female authors (26.5%). We could not confidently determine the gender of 761 out of 12692 (6.0%) of the authors. Some of these are due to single letter first names or problems with ill-formatted data.

Table 1 lists the number of papers for each gender. About twice as many papers had at least one male author (64%) as had at least one female author (33%). The statistics for first authorship were slightly more skewed; women were the first author of 27% of papers, whereas men first authored 71%. In papers with at least one female author, the first author was a woman 60% of the time, whereas papers with at least one male author had a male

first author 80% of the time. Thus men not only write more papers, but are also more frequently first authors.

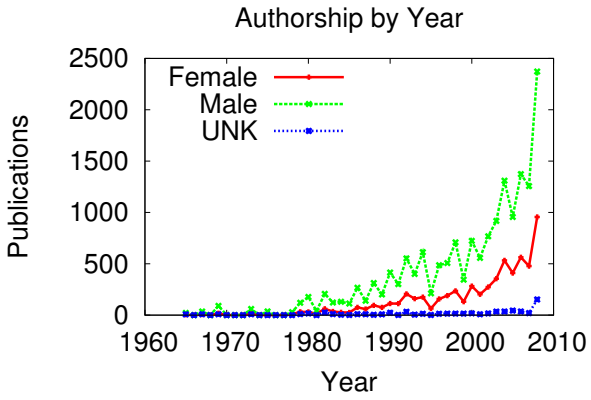

Figure 2: The number of authors of a given gender for a given year.

Figure 2 shows gender statistics over time, giving the number of authors of a given gender for a given year. An author is considered active for a year if he or she was an author of at least one paper. The number of both men and women authors increases over the years, reflecting the growth of computational linguistics.

Figure 3 shows the percentage of authors of a given gender over time. We overlay a linear regression of authorship percentage for each gender showing that the proportion of women is growing over time. The male best fit line has equation $y = -0.3025x + 675.49 (R^2 = 0.41, p = 1.95 \cdot 10^{-5})$ and the female best fit line is $y = 0.3429x - 659.48 (R^2 = 0.51, p = 1.48 \cdot 10^{-5})$. Female authorship percentage grew from 13% in 1980 to 27% in 2007, while male authorship percentage decreased from 79% in 1980 to 71% in 2007. Using the best fit lines as a more robust estimate, female authorship grew from 19.4% to 29.1%, a 50% relative increase.

This increase of the percentage of women authorship is substantial. Comparable numbers do not seem to exist for computer science in general, but according to the CRA Taulbee Surveys of computer science (CRA, 2008), women were awarded 18% of the PhDs in 2002 and 20.5% in 2007. In computational linguistics in the AAN, women first-authored 26% of papers in 2002 and 27% of papers in 2007. Although of course these numbers are not directly comparable, they at least suggest that women participate in computational linguistics research at least as much as in the general computer science population and quite possibly significantly more.

We next turn attention to how the most prolific authors of each gender compare. Figure 4 shows the number of papers published by the top 400 authors of each gender, sorted in decreasing order. We see that the most prolific authors are men.

There is an important confound in interpreting the number of total papers by men and the statistics on prolific authors. Since, as Figure 3 shows, there was a smaller proportion of women in the field in the early days of computational linguistics, and since authors publish more papers the longer they are in the field, it's important to control for length of service.

Figure 5 shows the average number of active years for each gender. An author is considered active in the years between his or her first and last publication in the anthology. Comparing the number of years of service for each gender, we find that on average men indeed have been in the field longer (t-test, $p = 10^{-6}$).

Accounting for this fact, Figure 6 shows the average number of publications per active year. Women published an average of 1.07 papers per year active, while men published 1.03 papers per active year. This difference is significant (t-test, $p = 10^{-3}$), suggesting that women are in fact slightly more prolific than men per active year.

In the field of Ecology, Sih and Nishikawa (1988) found that men and women published roughly the same number of papers per year of service. They used a random sample of 100 researchers in the field. In contrast, Symonds et al. (2006) found that men published more papers per year than women in ecology and evolutionary biology. This study also used random sampling, so it is unclear if the differing results are caused by a sampling error or by some other source.

4 Topic Models

In this section we discuss the relationship between gender and document content. Our main tool is Latent Dirichlet Allocation (LDA), a model of the topics in a document. We briefly describe LDA; see (Blei et al., 2003) for more details. LDA is a generative model of documents, which models documents as a multinomial mixture of *topics*, which in turn are multinomial distributions over words. The generative story proceeds as follows: a document first picks the number of words N it

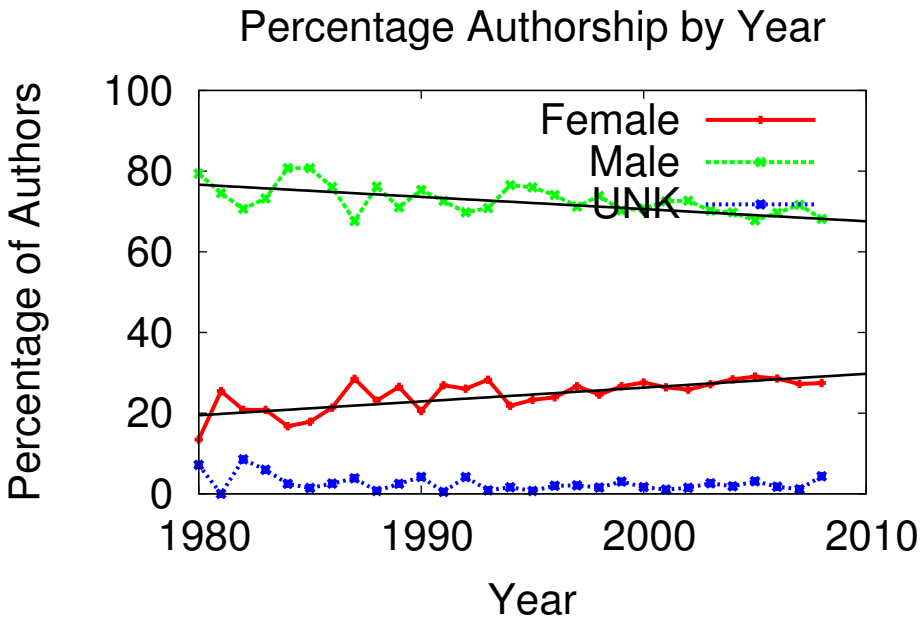

Figure 3: The percentage of authors of a given gender per year. Author statistics before 1980 are sparse and noisy, so we only display percentages from 1980 to 2008.

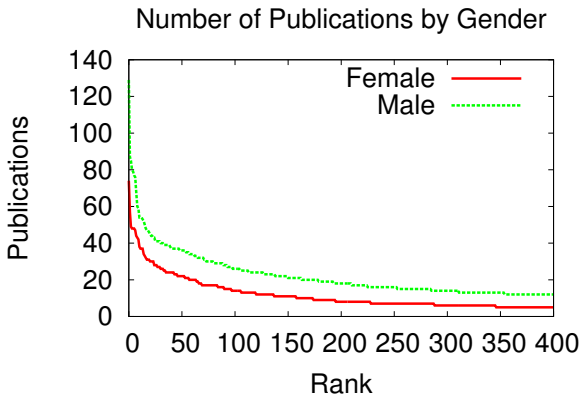

Figure 4: The number of publications per author sorted in decreasing order.

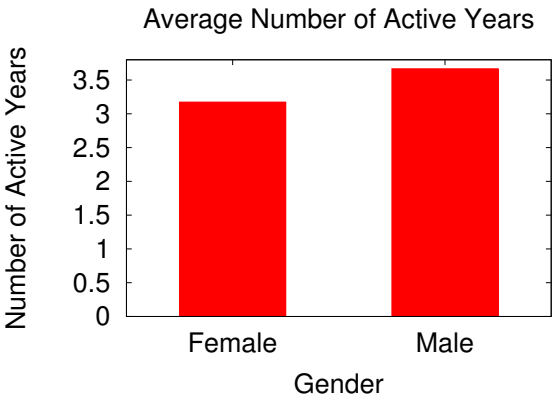

Figure 5: The average number of active years by gender

will contain and samples a multinomial topic distribution $p(z|d)$ from a Dirichlet prior. Then for each word to be generated, it picks a topic z for that word, and then a word from the multinomial distribution $p(w|z)$.

Following earlier work like Hall et al. (2008), we ran LDA (Blei et al., 2003) on the ACL Anthology, producing 100 generative topics. The second author and another senior expert in the field (Christopher D. Manning) collaboratively assigned labels to each of the 100 topics including marking those topics which were non-substantive (lists of function words or affixes) to be eliminated. Their consensus labeling eliminated 27 topics, leaving 73 substantive topics.

In this study we are interested in how documents written by men and women differ. We are mainly interested in $\Pr(Z|G)$, the probability of a topic being written about by a given gender, and $\Pr(Z|Y,G)$, the probability of a topic being written about by a particular gender in a given year. Random variable Z ranges over topics, Y over years, and G over gender. Our topic model gives us $\Pr(z|d)$, where d is a particular document. For a document $d \in D$, let d_G be the gender of the first author, and d_Y the year it was written.

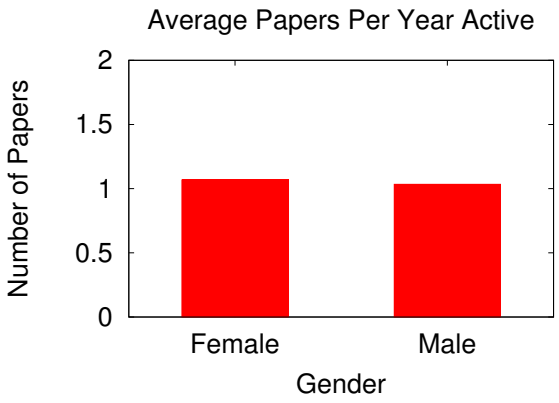

Figure 6: The average number of papers per active year, where an author is considered active in years between his or her first and last publication.

To compute $\Pr(z|g)$, we sum over documents whose first author is gender g:

$$\Pr(z|g) = \sum_{\{d \in D | d_G = g\}} \Pr(z|d) \Pr(d|g)$$
$$= \sum_{\{d \in D | d_G = g\}} \frac{\Pr(z|d)}{|\{d \in D | d_G = g\}|}$$

To compute $\Pr(z|y, g)$, we additionally condition on the year a document was written:

$$\Pr(z|y, g) = \sum_{\{d \in D | d_Y = y\}} \Pr(z|d) \Pr(d|y, g)$$
$$= \sum_{\{d \in D | d_Y = y, d_G = g\}} \frac{\Pr(z|d)}{|\{d \in D | d_Y = y, d_G = g\}|}$$

To determine fields in which one gender publishes more than another, we compute the odds-ratio

$$\frac{\Pr(z|g = \text{female})(1 - \Pr(z|g = \text{female}))}{\Pr(z|g = \text{male})(1 - \Pr(z|g = \text{male}))}$$

for each of the 73 topics in our corpus.

5 Topic Modeling Results

Using the odds-ratio defined above, we computed the top eight male and female topics. The top female-published topics are speech acts + BDI, prosody, sentiment, dialog, verb subcategorization, summarization, anaphora resolution, and tutoring systems. Figure 9 shows the top words for each of those topics. Figure 7 shows how they have evolved over time.

The top male-published topics are categorial grammar + logic, dependency parsing, algorithmic efficiency, parsing, discriminative sequence models, unification based grammars, probability theory, and formal semantics. Figure 8 and 10 display these topics over time and their associated words.

There are interesting possible generalizations in these topic differences. At least in the ACL corpus, women tend to publish more in speech, in social and conversational topics, and in lexical semantics. Men tend to publish more in formal mathematical approaches and in formal syntax and semantics.

Of course the fact that a certain topic is more linked with one gender doesn't mean the other gender does not publish in this topic. In particular, due to the larger number of men in the field, there can be numerically more male-authored papers in a female-published topic. Instead, what our analysis yields are topics that each gender writes more about, when adjusted by the number of papers published by that gender in total.

Nonetheless, these differences do suggest that women and men in the ACL corpus may, at least to some extent, exhibit some gender-specific tendencies to favor different areas of research.

6 Conclusion

Our study of gender in the ACL Anthology shows important gains in the percentage of women in the field over the history of the ACL (or at least the last 30 years of it). More concretely, we find approximately a 50% increase in the proportion of female authors since 1980. While women's smaller numbers means that they have produced less total papers in the anthology, they have equal (or even very slightly higher) productivity of papers per year.

In topics, we do notice some differing tendencies toward particular research topics. In current work, we are examining whether these differences are shrinking over time, as a visual overview of Figure 7 seems to suggest, which might indicate that gender balance in topics is a possible outcome, or possibly that topics first addressed by women are likely to to be taken up by male researchers. Additionally, other applications of topic models to the ACL Anthology allow us to study the topics a single author publishes in over time (Anderson et al., 2012). These techniques would allow us to study how gender relates to an author's topics throughout his or her career.

Our gender labels for ACL authors (available at

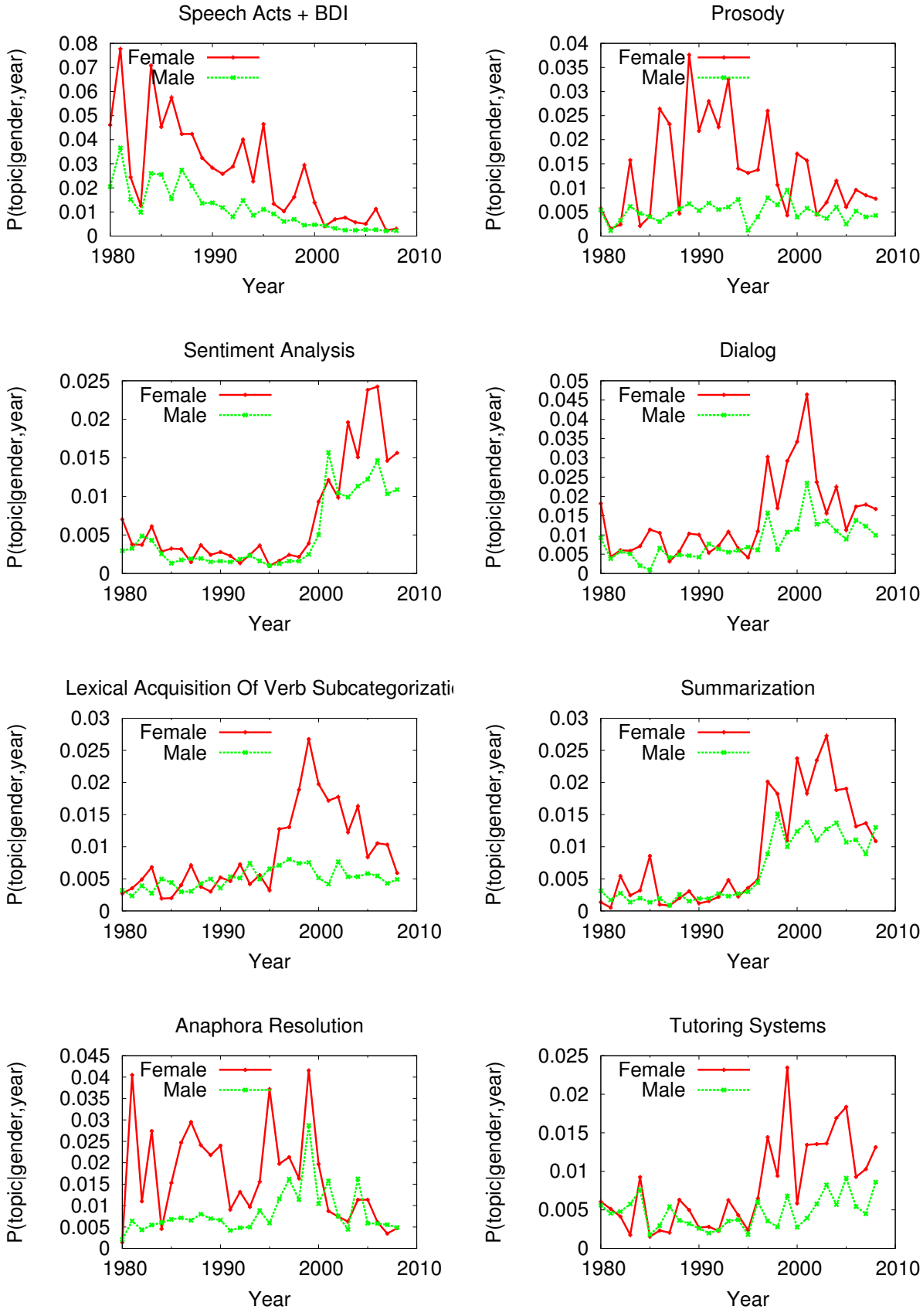

Figure 7: Plots of some topics for which $P(\text{topic}|\text{female}) > P(\text{topic}|\text{male})$. Note that the scale of the y-axis differs between plots.

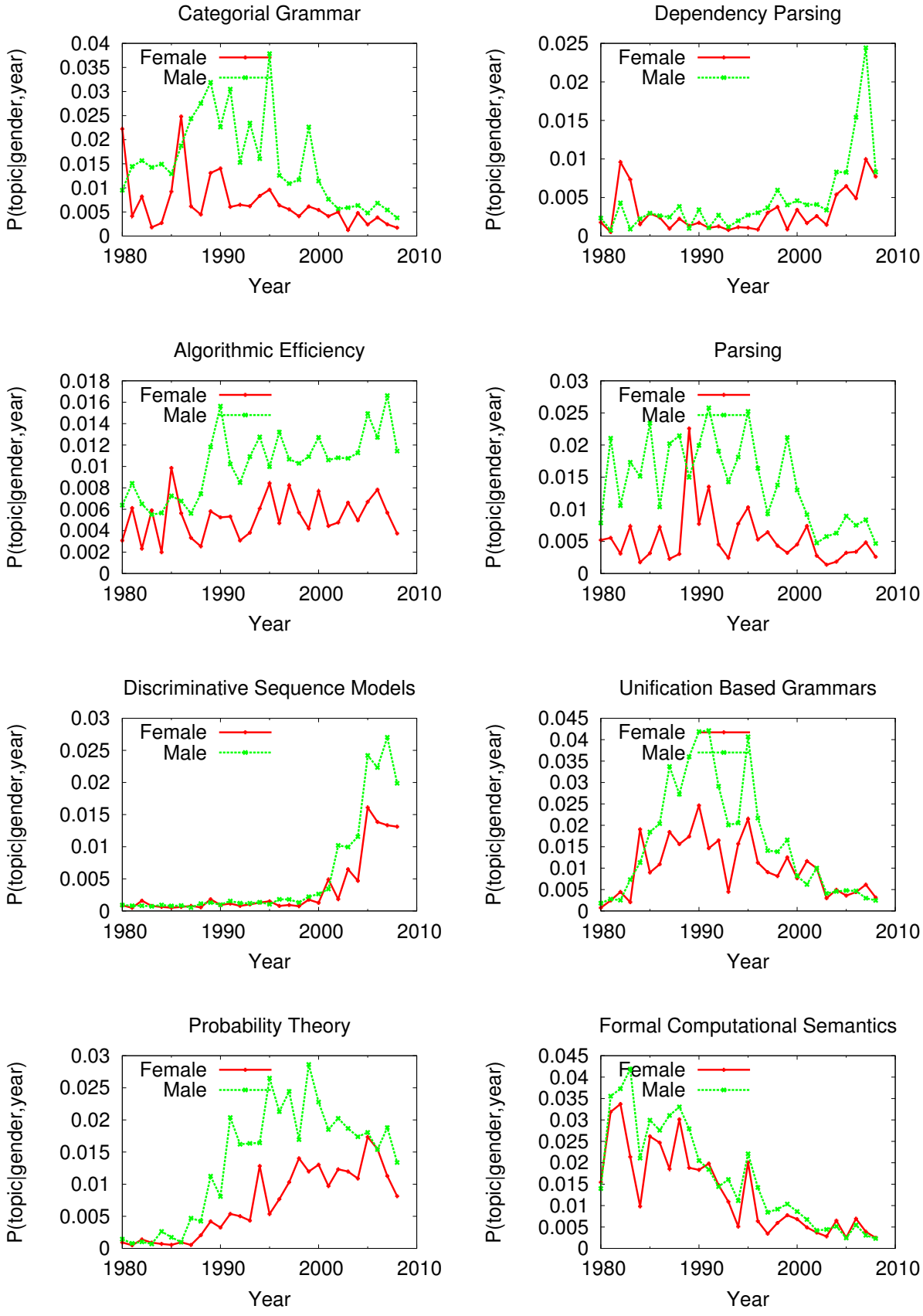

Figure 8: Plots of some topics for which $P(\text{topic}|\text{male}) > P(\text{topic}|\text{female})$. Note that the scale of the y-axis differs between plots.

Speech Acts + BDI	speaker utterance act hearer belief proposition acts beliefs focus evidence
Prosody	prosodic pitch boundary accent prosody boundaries cues repairs speaker phrases
Sentiment	question answer questions answers answering opinion sentiment negative trec positive
Dialog	dialogue utterance utterances spoken dialog dialogues act turn interaction conversation
Verb Subcategorization	class classes verbs paraphrases classification subcategorization paraphrase frames acquisition
Summarization	topic summarization summary document news summaries documents topics articles content
Anaphora Resolution	resolution pronoun anaphora antecedent pronouns coreference anaphoric definite reference
Tutoring Systems	students student reading course computer tutoring teaching writing essay native

Figure 9: Top words for each topic that women publish in more than men

Categorial Grammar + Logic	proof logic definition let formula theorem every defined categorial axioms
Dependency Parsing	dependency dependencies head czech depen dependent treebank structures
Algorithmic Efficiency	search length size space cost algorithms large complexity pruning efficient
Parsing	grammars parse chart context-free edge edges production symbols symbol cfg
Discriminative Sequence Models	label conditional sequence random labels discriminative inference crf fields
Unification Based Grammars	unification constraints structures value hpsg default head grammars values
Probability Theory	probability probabilities distribution probabilistic estimation estimate entropy
Formal Semantics	semantics logical scope interpretation logic meaning representation predicate

Figure 10: Top words for each topic that men publish in more than women

`http://nlp.stanford.edu/projects/gender.shtml`) provide an important resource for other researchers to expand on the social study of computational linguistics research.

7 Acknowledgments

This research was generously supported by the Office of the President at Stanford University and the National Science Foundation under award 0835614.

Thanks to Steven Bethard and David Hall for creating the topic models, Christopher D. Manning for helping label the topics, and Chu-Ren Huang, Olivia Kwong, Heeyoung Lee, Hwee Tou Ng, and Nigel Ward for helping with labeling names for gender. Additional thanks to Martin Kay for the initial paper idea.

References

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

Theresa Beaubouef and Wendy Zhang. 2011. Where are the women computer science students? *J. Comput. Sci. Coll.*, 26(4):14–20, April.

S. Bird, R. Dale, B.J. Dorr, B. Gibson, M. Joseph, M.Y. Kan, D. Lee, B. Powley, D.R. Radev, and Y.F. Tan. 2008. The ACL Anthology Reference Corpus: A reference dataset for bibliographic research in computational linguistics. In *LREC-08*, pages 1755–1759.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March.

US Census. 2012. First name frequency by gender. http://www.census.gov/genealogy/names/names_files.html.

CRA. 2008. CRA Taulbee Survey (web site). http://www.cra.org/resources/taulbee/.

David L.W. Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of Conference on Empirical Methods on Natural Language Processing*.

Catherine Hill, Christianne Corbett, and Andresse St Rose. 2010. *Why So Few? Women in Science, Technology, Engineering, and Mathematics*. American Association of University Women.

Paul De Palma. 2001. Viewpoint: Why women avoid computer science. *Commun. ACM*, 44:27–30, June.

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The ACL Anthology Network corpus. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, NLPIR4DL '09, pages 54–61, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andrew Sih and Kiisa Nishikawa. 1988. Do men and women really differ in publication rates and contentiousness? an empirical survey. *Bulletin of the Ecological Society of America*, 69(1):pp. 15–18.

Kusum Singh, Katherine R Allen, Rebecca Scheckler, and Lisa Darlington. 2007. Women in computer-related majors: A critical synthesis of research and theory from 1994 to 2005. *Review of Educational Research*, 77(4):500–533.

Ellen Spertus. 1991. Why are there so few female computer scientists? Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA.

Matthew R.E. Symonds, Neil J. Gemmell, Tamsin L. Braisher, Kylie L. Gorringe, and Mark A. Elgar. 2006. Gender differences in publication output: Towards an unbiased metric of research performance. *PLoS ONE*, 1(1):e127, 12.