Gender and Language Use in Changing Demographics: A Case Study in Scientific Writing

Anonymous EMNLP submission

Abstract

We present a study of the relationship between gender, demographics, and linguistic styles, using a corpus of scientific writings. Prior empirical work on gender treats it as an unchanging binary opposition. In contrast, we present an approach that views gender as a social identity. Drawing inspiration from theories in social psychology, we find that aspects of inter-gender linguistic differences as well as commonalities change with shifting gender demographics of a group. This integration of empirical methods with social theories offers new insights into how gender transpires in response to and/or as a reinforcement of social groups.

1 Introduction

A formidable body of work in sociolinguistics has argued that there is a connection between language and social identities such as gender, ethnicity, and age. With the vast amount of data becoming increasingly available, large-scale computational analyses of such connections have flourished. The primary goal is to either build predictive models of these social attributes or to understand stylometric differences. Several predictive models have been impressively accurate. However, they present an overly simplistic picture of the relation between language use and these attributes.

We present a study of the relation between language, demographics, and a salient social identity – gender (Sherif, 1982; Deaux, 1984). In doing so, we address an important constraint of previous computational analyses of language and gender.

Prior work has focused on lexico-syntactic differences between the language use by women and men, but based only on gender. This creates a threefold problem: (i) it amplifies the perceived gender differences without accounting for overlap, thereby leading to stereotypical interpretations (Koolen and van Cranenburgh, 2017), (ii) it does not account for linguistic differences that may be due to different social contexts (e.g., Baker (2014, p. 30)), and (iii) the analyses rest on the *a priori* assumption that 'women' and 'men' are always distinct and stable binary categories (Bamman et al., 2014; Larson, 2017). Further, this disregards social theories and evidence from qualitative studies that gender can be performative (Butler, 2011), and thus emerging as a response to social contexts and objectives while at the same time, being constrained by them (Brewer, 1991; Leonardelli et al., 2010). Our study uses formal scientific literature from a single domain to control for the writers' external social context. We find that linguistic differences change in style as well as magnitude when the demographics of the social group change.

2 Further Related Work

There is a long history of qualitative (e.g., Tannen (1990)) and quantitative (Pennebaker et al., 2003; Argamon et al., 2003) work connecting gender and language use. Both bodies of work have presented general characteristics of gender-specific language use. Computational approaches, too, analyzed such differences and have been successful at developing classifiers (Mukherjee and Liu, 2010; Sarawgi et al., 2011; Bergsma et al., 2012).

The latter has taken what is dubbed the "folk" view of gender (Larson, 2017). But if one takes the performative view of gender, studies must account for the behavioral aspects (which would include language use) change vis-á-vis social context.Outside of qualitative studies, only a very few have adopted this. Notable among them are Ellis (2009), Filippova (2012), and Bamman et al. (2014). This body of work, however, investigates general social media, where linguistic variation exists due to a multitude of uncontrollable factors. Associating language with any social identity under such circumstances can be misleading, as has been argued in detail by research in sociolinguistics (such as Eckert (2008), among others). In terms of analysing gender as a social identity, our work is similar in spirit to Bamman et al. (2014). But, while they present insightful results of cases where gender-based linguistic behavior changes, their model falls back on analyses of word classes instead of exploring deeper linguistic constructs.

Due to the above factors, we focus on a domain that reflects language use by a well-defined community where most writers are likely to have less influence on their writing styles from outside the community – scientific writings. Unlike Sarawgi et al. (2011) or Bergsma et al. (2012), who also looked at such texts, we explore two new frontiers in terms of model building. First, we work with a much larger dataset by including documents with multiple authors. This allows us to study how the stereotypical characterization of language-gender links changes when the *gender composition* of authors change. Second, we explore non-lexical syntactic features to control for topic. Not doing so can falsely attribute stylometric traits to social identities like gender, as was explicitly demonstrated by Herring and Paolillo (2006).

3 Data

Our research is based on analyses of the ACL Anthology Reference Corpus (Bird et al., 2008). The goal was to create a subset such that the articles pertain to research findings, and are comparable in terms of the complexity or magnitude of the scientific work being presented, so we filtered out front/back matter files and student workshop publications. We also excluded articles whose abstract and introduction sections put together consisted of fewer than 500 characters To assign gender to the authors, we only considered first names. For gold-standard labels, we used historical census information from U.S. Social Security Administration data from 1880 to 2016.A large fraction of these names are gender-neutral, so we adopted the frequentist approach, and assigned a score $s \in [0, 1]$, with 0 and 1 indicating exclusively male and exclusively female names, respectively. The score is simply the ratio of women with a specific name to the total number of people with that name in the corpus. For example, “Alex” and “Laurel” were scored at 0.0311 and 0.9589, respectively. Since this data does not cover all international names, we also included unambiguous name-gender associations from `http://www.behindthename.com`. If

—Dataset—	—Characteristics—
D1	4,578 articles 7,463 unique authors
D2	1,739 articles with all-male authorship 362 articles with all-female authorship

Table 1: Dataset Overview

a name was not found in either resource, we assigned it a gender score of $\bar{s}_{\text{corpus}} = 0.2477$, which is the gender ratio in the entire collection of author names from the corpus.

Finally, articles were discarded if all authors were assigned $\bar{s}_{\text{corpus}}$. Applying all these filters yielded our first dataset. We also created a smaller set composed of articles with single-gender authorship. This was done by further discarding articles for which the weighted mean gender score of all authors was $0.05 \le \bar{s} \le 0.95$. The two datasets are described as **D1** and **D2** in Table 1. Observations across all gender ratios (Sec. 4.1) are drawn from **D1**, and the remainder uses **D2**.

4 Linguistic Indicators of Gender

Given the domain and genre of our dataset, including lexical features would reveal the topics discussed in an article. In other words, lexical features are associated with topics, which are in turn associated with the authors. To control for the topics, we focus exclusively on syntactic features. Unlike previous work on language and gender, the instances in our data have multiple authors. We therefore extended the idea of gender being represented by a numerical value in $[0, 1]$. This was done taking a weighted average of the gender scores of all the authors. If an article d was written by $a_1, \ldots a_{k_d}$ (in order of authorship), the article's gender was represented by $(\sum_{i \in [k_d]} w_i s_i)/k_d$, where s_i is the gender of the i^{th} name. The weights were decided on held-out data, and the results presented here used $w_1 = 0.5$, $w_{k_d} = 1$, and $w_i = 0.25$ for all $2 \le i \le k_d$.

The common approach taken by computational work in language and gender has been to first produce a predictive model and then perform *a posteriori* analysis of the discriminatory power of its features. Instead, we start by looking at the widely used ‘involved’ and ‘informational’ features.

4.1 Syntactic Features

The involved dimension comprises frequent use of 1^{st}- and 2^{nd}-person pronouns (Hirschman, 1994; Argamon et al., 2003), present and past

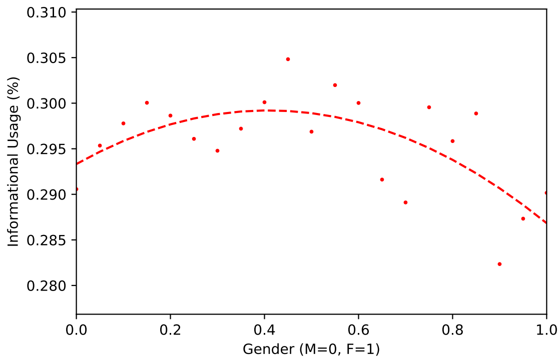

Figure 1: Variational use of 'informational' language.

tense (Biber et al., 1998; Newman et al., 2008), intensive adverbs (Mulac et al., 2000), personal pronouns (Newman et al., 2008), and conjunctions (Ireland and Pennebaker, 2010). Informational language, on the other hand, is characterized by a predominance of nouns, prepositions, and 3[rd]-person pronouns (Biber et al., 1998; Hirschman, 1994). Fig. 1 shows that this standard characterization of male and female language use varies with gender demographics. In course of this analysis, we also investigated other parts of speech, and found that women's language exhibited higher usage of *wh*-adverbs, possessive pronouns, and subjective and objective pronouns. Since multiple features were correlated with gender, we applied Bonferroni correction (Dunn, 1961). After the correction, however, they were not significant. In spite of the above observation, due to some promising results with deep syntactic features presented by Sarawgi et al. (2011) and Bergsma et al. (2012), our next step was to investigate whether or not complex stylometric aspects of language vary with changing gender demographics.

To control for topic, we focused on *interpretable* stylometry that has not seen much usage in computational methods. To this end, we used features from rhetorical theory explored by Feng et al. (2012) since their work distilled deep syntactic features from the same dataset (albeit for author identification). Upon investigating the distribution of different sentence structures as well as tree topological features, we found that like shallow syntax, these were correlated with demographic changes too, but were not statistically significant after Bonferroni correction.

4.2 A Predictive Model for Gender

After exploring rhetorical structures of syntax and the standard characterization, and failing to obtain statistically significant correlation with gender, we developed two classification models beyond the baseline: a feed-forward neural network (NN), and a support vector machine (SVM) classifier. Here, we were careful to model gender as the independent variable in order to avoid the kind of biases posited recently by Koolen and van Cranenburgh (2017). This approach is along the lines of the large-scale analyses of gender and language in social media undertaken by Bamman et al. (2014).

The dependent variables formed the feature space in which articles from the sub-dataset **D2** were represented. In addition to the syntactic features discussed in the earlier sections, we included parent-child bigrams from non-leaf production rules from constituency parse trees, dependency-label bigrams, part-of-speech bigrams, tree depth, mean sentence length, and the number of sentences. The constituency and dependency parse trees were generated using the Stanford CoreNLP Toolkit (Manning et al., 2014). As Table 1 shows, the dataset is highly imbalanced in favor of male authorship, so before training, we oversampled the articles written by women. We explored a perceptron classifier as a baseline, a feed-forward neural network, and a support vector machine (SVM) classifier. The perceptron model yielded a 54% accuracy. For the neural network model, we used an input layer of the size of the feature vector, and two hidden layers of 100 and 10 nodes, respectively. This model was trained with backpropagation, and with rectifiers (relU) as the activation function. With 10-fold cross-validation, this model achieved an accuracy of 76.75%.

Our best classifier was the SVM model. We used L2-loss. The feature vectors were built with `TF-IDF` encoding, and normalized to unit length. To avoid overfitting, we used L2 regularization with the parameter selected by the "warm-start" algorithm (Chu et al., 2015) recently added to LIBLINEAR (Fan et al., 2008). The final model was selected by 10-fold cross-validation, which achieved 90.02% accuracy. The complete confusion matrix is shown in Table 2. Since there is no prior work for gender attribution on multi-author documents, we do not include external baselines.

	Precision	Recall
Women	0.81	1.0
Men	1.0	0.76
Total	0.91	0.88

Table 2: Confusion matrix of the SVM classification model.

PCFG Segments		Dependency-label Bigrams	
Women	**Men**	**Women**	**Men**
VBZ →:	VP → JJ	dep → neg	acl → cop
JJR → TO	ADJP → RB	conj → csubj	ccomp → cc:preconj
JJR → RB	RBR → NNS	ccomp → case	nmod:tmod → nmod
VBG → VBD	VBG → RBR	nmod:tmod → nummod	dep → parataxis
PRP → DT	NP → ADJP	advmod → neg	acl:relcl → det
VBD → PRP$	NP → VB	cc → dep	cc:preconj → neg
VBP → RBR	PRP → IN	csubj → mark	dobj → nmod:npmod
JJ → WDT	CD → NNP	dobj → neg	appos → advmod
PRP → JJ	JJS → VBN	nmod → dobj	parataxis → advmod
NN → JJR	WP → NN	acl:relcl → nmod:tmod	dep → csubjpass

Table 3: Top 10 PCFG production segments and dependency-label bigrams for distinguishing women's and men's writing.

Since in linear SVM, feature weights indicate significance (Chang and Lin, 2008), we were able to extract significant features for both genders (Table 3). Even though possessive pronouns (PRP$) were not significantly associated with women's writing after applying Bonferroni correction, it appears in one of the most significant production rules indicative of female authorship. Similarly, intensive adjectives (JJR and JJS) are also associated with women's writing. It is worth noting that both have been regarded as components of 'involved' language in prior qualitative work. Analogously, nouns and cardinals appear in five of the top ten PCFG segments associated with men's writing. Similar observations may be made regarding the dependency features. For instance, men's writing seems to favor complex structures like relative clauses, parataxis, and appositives.

4.3 Longitudinal Analysis

Thus far, our analyses showed that (a) standard characterization of language and gender in 'involved' and 'informational' terms does not fit collaborative use of language in scientific writing. But, similar features may still be found if we explore deep syntactic distinctions between all-male and all-female writings. This, however, accounted only for the demographic changes *within* each document's authorship, not for any shifts in the gender demographics of the community as a whole. In this section, we study whether the gender-ratio in the community as a whole has any effect on how much the writing style of men and women changes. To answer this, we consider three probability distributions *per year*, from 1980 to 2015. These distributions are formed over the same feature space used by the SVM model. They are obtained by computing, for each year, the mean feature vectors of (i) all articles, (ii) articles with all male authors, and (iii) articles with all female authors. For a year y, let us denote these by P_y, $P_{y,f}$, and $P_{y,m}$. We would like to see if the change in gender demographics between the years y and $y+1$ correlates with changes in these distributions. To measure such changes, we compute the Kullback-Liebler (KL) divergence between P_y and P_{y+1} (and similarly for $P_{y,f}$ and $P_{y,m}$).

The null hypothesis is that male and female writing does not change any differently than that of the whole community. Fig. 2 shows this is not true. We observe a negative correlation of -0.34 between the (i) change in the number of women writers and the (ii) magnitude of change in their language. Changes in men's language use (not shown) was far less conspicuous.

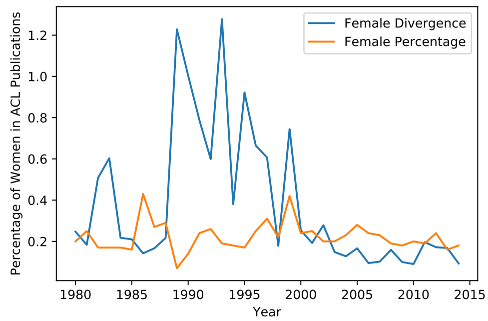

Figure 2: Annual changes in aggregate writing styles.

5 Conclusion

We have viewed gender as a social identity and analyzed a domain where gender attribution is much harder due to multi-author documents and the formal writing. We also showed that standard characterization of language and gender may not be stable, and change hand-in-hand with demographic changes. These changes affecting language use can be both within a small group or the larger community. The fact that the minority gropu exhibited larger changes may be due to out-group behavior (Tajfel and Turner, 2004). This, however, required further in-depth research into the matter.

References

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text – Interdisciplinary Journal for the Study of Discourse*, 23(3):321–346.

Paul Baker. 2014. *Using Corpora to Analyze Gender*. A&C Black.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Shane Bergsma, Matt Post, and David Yarowsky. 2012. Stylometric Analysis of Scientific Articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337. Association for Computational Linguistics.

Douglas Biber, Susan Conrad, and Randi Reppen. 1998. *Corpus linguistics: Investigating Language Structure and Use*. Cambridge University Press.

Steven Bird, Robert Dale, Bonnie J Dorr, Bryan Gibson, Mark Thomas Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proc. of Language Resources and Evaluation Conference (LREC 08)*. European Language Resources Association - ELRA.

Marilynn B Brewer. 1991. The Social Self: On Being the Same and Different at the Same Time. *Personality and Social Psychology Bulletin*, 17(5):475–482.

Judith Butler. 2011. *Bodies That Matter: On the Discursive Limits of Sex*. Taylor & Francis.

Yin-Wen Chang and Chih-Jen Lin. 2008. Feature ranking using linear svm. In *Causation and Prediction Challenge*, pages 53–64.

Bo-Yu Chu, Chia-Hua Ho, Cheng-Hao Tsai, Chieh-Yen Lin, and Chih-Jen Lin. 2015. Warm start for parameter selection of linear classifiers. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 149–158. ACM.

Kay Deaux. 1984. From individual differences to social categories: Analysis of a decade's research on gender. *American Psychologist*, 39(2):105.

Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Penelope Eckert. 2008. Variation and the Indexical Field. *Journal of Sociolinguistics*, 12(4):453–476.

David Ellis. 2009. Social (distributed) language modeling, clustering and dialectometry. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages 1–4. Association for Computational Linguistics.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Characterizing Stylistic Elements in Syntactic Structure. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1522–1533. Association for Computational Linguistics.

Katja Filippova. 2012. User demographics and language in an implicit social network. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1478–1488. Association for Computational Linguistics.

Susan C Herring and John C Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459.

Lynette Hirschman. 1994. Female–male differences in conversational interaction. *Language in Society*, 23(3):427–442.

Molly E Ireland and James W Pennebaker. 2010. Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology*, 99(3):549.

Corina Koolen and Andreas van Cranenburgh. 2017. These are not the Stereotypes You are Looking For: Bias and Fairness in Authorial Gender Attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22.

Brian N Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations.

Geoffrey J Leonardelli, Cynthia L Pickett, and Marilynn B Brewer. 2010. Optimal Distinctiveness Theory: A Framework for Social Identity, Social Cognition, and Intergroup Relations. In *Advances in experimental social psychology*, volume 43, pages 63–113. Elsevier.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in natural Language Processing*, pages 207–217. Association for Computational Linguistics.

Anthony Mulac, David R Seibold, and Jennifer Lee Farris. 2000. Female and male managers and professionals criticism giving: Differences in language use and effects. *Journal of Language and Social Psychology*, 19(4):389–415.

Matthew L Newman, Carla J Groom, Lori D Handelman, and James W Pennebaker. 2008. Gender Differences in Language Use: An Analysis of 14,000 Text Samples. *Discourse Processes*, 45(3):211–236.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, 54(1):547–577.

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender Attribution: Tracing Stylometric Evidence Beyond Topic and Genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 78–86. Association for Computational Linguistics.

Carolyn Wood Sherif. 1982. Needed Concepts in the Study of Gender Identity. *Psychology of Women Quarterly*, 6(4):375–398.

Henri Tajfel and John C Turner. 2004. The social identity theory of intergroup behavior.

Deborah Tannen. 1990. *You Just Don't Understand: Women and Men in Conversation*. New York: Ballantine Books.