Bold Tweets: Female Politicians Shed Classical Language Norms

Davey Proctor

May 3, 2018

1 Abstract

I examine the interaction between gender-based language norms and political success. I analyzed 50 tweets from each of 107 female and 109 male members of the U.S. congress using IBM's Tone Analyzer. Contrary to the standard hypothesis that women are less forceful in their speech than men, I find that female politicians tweet with as or more confident and analytical tones than male congressmen ($p < .0002$), an effect not explained away by party affiliation or age. Additionally, these women are not overly agreeable in the sense of tweeting with vanilla joyful tones: women's tones are at least as sad and no more joyful ($p < .01$) visa-vie men. Instead, joyfulness and agreeableness seem to have more to do with party affiliation and being in power (Democrats tweet with less joy than Republicans ($p < .0005$)). Finally, while women are actively encouraged by Twitter users to be confident ($p < .0005$), this encouragement is less strong than what men receive ($p \approx .05$). This data suggests that female politicians, in attaining their positions, have shed much of the gendered language norms that may exist; meanwhile, Twitter followers are becoming increasingly (but not yet completely) equally supportive of women's powerful speech patterns.

2 Introduction

How do gender-based language norms mediate women's success in politics? First, I will review research on how language is a gendered phenomenon: women are encouraged to use "indoor voices," express greater uncertainty in their opinions, be supportive rather than assertive, avoid controversy, and perhaps even refrain from "real world" discussions. Second, I narrow in the question on politics: what makes women successful / unsuccessful in politics? What implicit biases do voters have? How do the words you use affect your electability?

Putting the two together, in this section I will examine theoretical arguments for the *standard hypothesis*: voter bias and social pressures uphold gendered language norms that makes political life harder for women than men. In subsequent sections, we'll criticize this hypothesis using Twitter data.

2.1 Language Norms

Foundations. Lakoff's [1] is regarded as foundational in the study of the potentially pernicious effects of language on gender equality. Lakoff argues that women are meant to be "marginal to the serious concerns of life" both based on the way women are expected to speak and in the way women are spoken of. We are primarily interested in the first: how women do speak, and how that relates to how women are expected to speak. On this question, Lakoff makes several observations. On the one hand, "If a little girl 'talks rough' like a boy, she will normally be ostracized, scolded, or made fun of." On the other hand, "If the little girl learns her lesson well, [then by adulthood she] will be accused of being unable to speak precisely or to express herself forcefully." In this way, we recognize the emergence of a double bind in which "a girl is damned if she does, damned if she doesn't."

Lakoff's methods are primarily hypothetical and anecdotal. Her goal is not so much to prove her thesis as to lay the groundwork for future studies: "The data on which I am basing my claims have been

gathered mainly by introspection ... [neither] the methodology [nor] the results are final, or perfect." That said, the linguistic thought experiments used are interesting and have been influential to later research. She posits a study in which participants are asked to choose which of the following was said by a woman, which by a man:

1. Oh dear, you've put the peanut butter in the refrigerator again.

2. Shit, you've put the peanut butter in the refrigerator again.

Notwithstanding the 1970s anachronism of "Oh dear," the reader can perhaps predict how a majority of participants would answer that question.

Responses. We now move on to reactions to and scientific formalizations of Lakoff's work. Still at the theoretical level, several criticisms have been made of Lakoff's fundamental assumptions and treatment of the gendered language. As summarized in Leaper & Robnet [2], one potential issue is the comparison to a masculine norm by default. For instance, Spender [3] argues that even if women's speech differs from men, that does not necessarily mean they are less equiped to express their opinions, or that the male style is superior. This concern is particularly salient to our discussion. Suppose we find that female politicians tend to use less angry language than male politicians. This does not necessarily prove on its own that language is a barrier to women in politics. Suppose additionally that voters have an unconcious bias for women speaking a certain way, perhaps causing female politician's use of kinder or tentative language. The essential question would become whether voters on balance favor more powerful language from politicians in general; if not, women are not necessarily disadvantaged relative to men in politics.

A second concern about Lakoff's argument is that any differences in speech between the genders might be mostly or entirely correlative, and actually caused by differences in relative status. As Henley [4] points out, tentative language might have more to do with being in subordinate positions than being female. Hall [5] also takes up this argument, concerned with how gender-based differences in nonverbal communication might uphold social hierarchies, and vice-versa. Again, our focusing on the language of political leaders–those citizens essentially *not* in subordinate positions–should shed light on this debate. If Henley is right, we would expect women in power to use as forceful of language as men in power. Less concerned about the ability of women to lead once elected, we would still be concerned–perhaps additionally so–about women's ability to navigate the political pipeline and move up the ranks, imagining situations in which political superiors expect tentative or deferential language from potential candidates.

Finally, the lack of an empirical basis for Lakoff's assertions might lead to relying on and reinforcing potentially false stereotypes about women (Mulac & Bradac [6]).

Empirical findings. Next we'll consider empirical treatments of Lakoff's theses. The literature has been very divided on the empirical validity of Lakoff's claims that women, to a large degree, learn to use language differently from men. Additionally, the contexts in which language norms hold or fail to hold is an extremely important area of research. One recent meta-analysis by Leaper and Robnet [2] focused on the language norm of tentative speech. They found that while women do tend to use more tentative speech than men, they do so by a small margin. Instead, the circumstances of the studies in the meta-analysis (longer vs. shorter conversations, group vs. individual settings, undergraduates vs. other adults, etc.) created large variances in women's tentativeness. The authors suggest that women might use more tentative language due to interpersonal sensitivity rather than a lack of assertiveness. This finding bears directly on our question: is politics a setting in which women use tentative language? If so, why? Do they feel forced by voter bias to use that language, or are they perhaps more interpersonally sensitive?

Another meta-analysis by Leaper and Ayres [7] looked at gender differences in talkativeness, affiliative speech, and assertive speech. As defined by the authors, "Affiliative language functions to affirm or positively engage the other person." Contrary to stereotypes and expectation, men are slightly more talkative than women, but women do use slightly more affiliative and less assertive speech than men. This is a basis for expecting that female politicians might use less angry and controversial language than male politicians; they might speak favorably towards what they *do* believe rather than negatively against what they don't.

2.2 Women in Politics

The Problem. One in five representatives in national legislatures (and in the U.S. Congress) are women [8], [9]. There are several possible explanations. On the voter side, explanations include outright voter hostility, in which voters explicitly favor male leaders to female; implicit voter bias, in which subconscious preferences make women less electable; and demographic characteristics such as being married with children. On the candidacy side, gatekeeper bias suggests that those in power are biased against promoting female candidates. We will briefly explain each one, discuss the evidence, and relate the explanation to gender-based language norms.

Voters. In a recent survey of American voters by Teele, Kalla, & Rosenbluth [10], there was little evidence for outright voter hostility. If anything, women were slightly preferred to men, accounting for the possibility of partisan bias among survey respondents. Gender-based language norms, even if they apply to politicians, are not likely to exist at the explicit level. Explicit bias for certain language would include insults of the form "that politician talks like a woman" – such speech is fortunately not acceptable in modern politics, and is likely rare among voters.

And while the authors found little evidence for implicit voter bias in the form of double standards, they suggest double standards could be an issue for female politicians: "it seems unlikely that Hillary Clinton would be where she is in politics if she had mothered children with three different fathers. But this did not seem to pose an obstacle for Donald Trump." Additionally, Dolan [11] found that men and women differ in their support of female candidates, and the difference increases with the level of the office being sought. Finally, there are strong empirical grounds to suspect that implicit bias can influence voters on various issues. Young, Ratner, and Fazio [12] used neuroscientific techniques to conclude that one's partisan affiliation literally changes one's mental representation of a candidate's face. The authors suggest "[Citizens] may construct a political world in which they literally see candidates differently." Returning to language norms, the standard hypothesis might be that implicit voter bias for gendered speech patterns limits the success and effectiveness of female politicians. This can include less confident, analytical, and angry speech.

The last consideration of voter influence on the poor gender ratio of female leaders is in the form of demographic characteristics. Teele, Kalla, & Rosenbluth [10] found that American voters prefer candidates similar to the average adult: married with children. Because of the expectation and practice that women invest more in childcare, it becomes harder for women to satisfy this demographic constraint than men, while advancing in the political ranks. Endorsing this explanation, Rachel Silberman [13] found that women who live in districts far from the capital are much less likely to run for office, seeking instead to minimize travel time more than men. In terms of speech patterns, we might expect politicians without children to express a parental spirit to compensate for not being in that demographic. However, this hypothesis is outside the scope of what I'll be testing.

Candidates. The second source of explanation of the gender ratio among our leaders focuses on candidacy side. Perhaps it is harder for women to become politicians than it is for men. Most of the literature on this topic focuses on the "Gatekeeper Bias" in which the party leaders, who are mostly men, might disproportionately promote the careers of men in politics. This discrimination can be either explicit or statistical: explicit if gatekeepers's taste favors men, statistical if gatekeepers choose candidates who are electable, and an electable candidate has to persuade potentially biased voters. In Kalla, Rosenbluth, and Teele [14], ficticious emails from political self-starters were sent to public officials, asking for advice on how to advance a beginning political career. Strikingly, women were just as encouraged as men in the responses, if not more so. Thus, the gatekeeper bias from public officials does not seem to be a primary source of concern. Whether a difference in the tone of the emails would influence these results is an area of future research.

3 Method

3.1 Research Design

Based on the literature, I will be testing several related hypotheses about how female politicians express sentiments in their tweets: I expect they use less angry, analytical, and confident tones; instead, they might be more joyful / agreeable, fearful, and tentative. I will also see how this speech correlates with positive / negative responses from twitter

users: I predict politicians receive positive feedback when they conform to expected language norms.

My method consisted of two phases. First, selecting the politicians and gathering the tweets. Second, tagging the tweets for various tones.

3.2 Phase 1: Tweets

Using the complilation `https://github.com/unitedstates/congress-legislators`, I collected 107 active Twitter accounts of currently serving congresswomen. To roughly match, I randomly selected (without replacement) 109 male members of congress. No special thought was given to other demographic information such as seniority or race.

Next, I used Tweepy, the Twitter Python API [15] to gather the 50 most recent tweets for each politician. The dates over which these tweets were posted varied based on politician, typically ranging over the past one to two years. In addition to the text of the tweet, I also used Tweepy to query basic meta-information about each tweet, including the tweet retweet count (retweeting a post allows a user's followers to see the material) and tweet favorites (a user can "favorite" a post to express agreement/joy with the post's content). Altogether, the resulting 10800-row table had column headers *politician_name, politician_username, gender, party, tweet_text, tweet_retweet_count, tweet_favorite_count*.

One systematic error I noticed in the data was the truncation of certain tweets. Until recently, Twitter would allow users to post above the 140 character limit (recently changed to 280), but Twitter would truncate the tweet at the first 140 characters, instead adding a '...' ellipsis at the end. The default behavior of Tweepy is to return the truncated text. We will keep this error in mind when analyzing the tones of the sometimes-truncated text.

3.3 Phase 2: Tones

Watson. I used IBM's Watson Tone Analyzer [16] to tag the tweets with a variety of sentiment scores. These are understood through three types of tones: emotions, such as anger, fear, joy, sadness, and disgust; personality traits such as openness, conscientiousness, extroversion, agreeableness, and emotional range; and writing styles such as confident, analytical, and tentative.

As an engineering consideration, in order to minimize the number of API calls to Watson (so it would run fast and remain free to use), it was advantageous to remove all sentence delimiters such as '.', '!', '?' and newlines from the tweets. I replaced these with '–' to best approximate the semantic "breath" that a sentence delimiters expresses. Presumably, the true sentence delimiters can help signal certain tones, and thus I recognize this might have introduced a slight systematic error in the results Watson returned. However, I applied the transformations uniformly to both men and women's tweets to minimize the negative impact.

In the results returned by Watson, only six sentiments came up in the tweets, each creating an additional column header in the table of tweets: *Joy, Analytical, Sadness, Confident, Fear, Anger, Tentative*, available from `https://github.com/daveyproctor/Polyspeech` [17] under the *more-data* directory. I imagine that the personality traits only get tagged on longer corpora; my analysis will thus be limited to emotions and writing styles. Still, those six variables are sufficient to speak forcefully on the hypothesis. Unlike the first half of the table involving tweets and their meta-data, not all the entries under the tone columns had non-none values. This makes sense, given that not all phrases in English express all tones. Instead, the table was somewhat sparse, Watson typically able to find between zero and two tones per tweet. These values took on decimal values between 0 and 1. I chose to fill with zero all the entries of the table that Watson had left empty.

Accuracy. How accurate do we expect the tagger to be? Watson is a proprietary product, and thus it's difficult to evaluate it on academic grounds. Our approach is three fold: one, to use the academic state of the art (SOTA) as a benchmark, arguing that private computational systems tend to match or exceed academic efforts; two, to use "sanity checks" on the output from Watson, ensuring sample tagged tweets conform with our own intuitions; three, to handle systematic error from Watson as random noise in the overall dataset.

SOTA. The current state of the art for sentiment analysis was recently achieved by Socher et al. [18] with 85% accuracy on positive / negative tone classification. Their method uses a hierarchical sentence structure on which they train a recursive neural network. As motivation for the complexity

of the task, consider the following two sentences:

1. It's just incredibly dull.
2. It's definitely not dull.

Although a trivial algorithm can figure out that "dull" expresses a negative sentiment, a multi-word understanding of the sentences is needed to correctly classify both of the above two sentences, given that "definitely not" negates the negative sentiment expressed by "dull". Even in a multi-word model, in which "defintely not dull" is understood as a positive sentiment, the following example remains problematic:

1. This film doesn't care about cleverness, wit or any other kind of intelligent humor.

Despite many positive words, and even longer positive phrases, the sentence is indisputably very negative in tone. Thus, we adopt 85% as a rough benchmark as accuracy in the area of sentiment analysis, understanding all the while the inherent difficulty in the task.

Sanity Checks. It is always good to sample the output from a "black box" computational system such as Watson to roughly understand its strengths and weaknesses. In Table 1, eight samples are shown from the dataset. Tweets expressing "Joy" and "Analytical" tones, seem to be correctly tagged. However, Watson incorrectly classified one tweet as expressing "Sadness," perhaps incorrectly understanding "passed" as in a death. A full-sentence example of anger returned by Watson was the tweet

1. "RT @EPWBoxer: Senator Barbara Boxer speaks on the #Senate Floor Exposing #WebOfDenial Blocking Action on Climate https://t–co/gdSqqRDkfY #——."

3.4 Analysis of the Design

Similar Studies. My study is similar in the analysis of gendered speech tones to the Leaper and Ayres [7] study which looked at gender differences in talkativeness, affiliative speech, and assertive speech. Unlike Leaper and Ayres, I am not gauging talkativeness, given that Twitter is an inherently concise and non-talkative medium. However, gauging confidence in the tones of choice will allow us to comment on the results that Leaper and Ayres found regarding assertive speech.

My approach will also directly comment on Leaper and Robnett [2], in which the question of tentative tones between men and women in a variety of contexts is discussed.

Strengths. With over $300M$ monthly followers [19], Twitter is one of the most frequent media of communication between individuals, particularly at the political level. In terms of sentiments, the brevity of a tweet (140 or 280 characters) encourages users to pack content into minimal space. We can naturally gauge the public receipt of a tweet by looking at the retweet count and favorite count. Additionally, the model is scalable, and builds off a tradition of collecting and analyzing large social media data sets to determine facts about politicians (see for example Anderson's [20], in which Trump's tweeting habits reflect on his day-to-day thinking).

Limitations. There are several issues with my methodology as a way to gauge the influence of gendered language norms on women's success in politics. In addition to Watson's potential errors in judgement, tweets are somewhat unreliable as a metric for how these politicians actually speak. For one, many politicians might employ a part or full-time social media expert to "ghost write" tweets. In this way, while it is still very interesting what sentiments receive a positive response from Twitter users, the tweets themselves, and tones therein, might reflect what the social media expert has determined will most positively represent the politician rather than the politician's true beliefs / speech tones. As concrete evidence, David Robinson [21] found that leading up to the 2016 election, Trump himself was tweeting roughly half the tweets from his account, owing to the clear tonal differences between sets of tweets from an iPhone and an Android.

Another problem relates to the Twitter users that can favorite / retweet politicians' posts. Colleoni et al. [22] found that followers on Twitter exhibit significant political homophily, implying that the people who see a politician's post are more likely to be favorable towards it than the average American. Thus, even if users respond favorably to a female politician's tweets, that politician might still exprience expectations of certain language norms in more politically heterogenous contexts (outside the "echo chamber").

Table 1: Watson-tagged Tweets

Sentiment	Score	Abbreviated Tweet
Analytical	.53	[...] families shouldn't have to pay [...]
Anger	.52	[...] innocent until proven guilty [...] judge needs to hit a quota.
Confident	.61	All Americans who are fit and willing [...] should be allowed to do so.
Fear	.66	[...] the opioid crisis, [...] overcome addiction.
Joy	.86	Wonderful to join Alaskans [...] We are so proud [...]
Joy	.72	[...] working very hard [...] - very exciting to see [...]
Sadness	.52	The recently passed funding includes [...]
Sadness	.79	After $65+ billion of damage [...] tragedy [...] disaster [...]

4 Results

4.1 Tone differences

The Normal Model. To explain the model, we first focus on just one tone: confidence. We took the mean over fifty tweets for each woman and man. This gave a total of 216 data points, 107 women and 109 men. Call the confidence scores $x_1, \dots x_m$ for the women and $y_1, \dots y_n$ for the men.

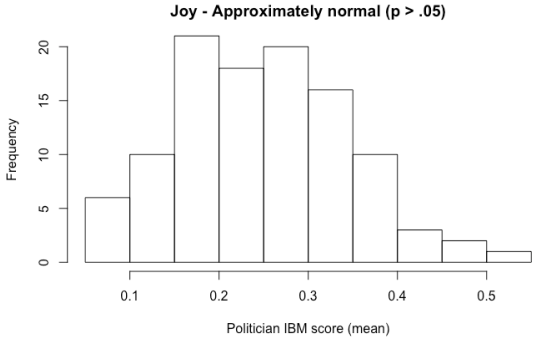
In our most naive model (which will nonetheless carry the majority of the analysis), we assume that these IBM scores for a certain tone are independently distributed:

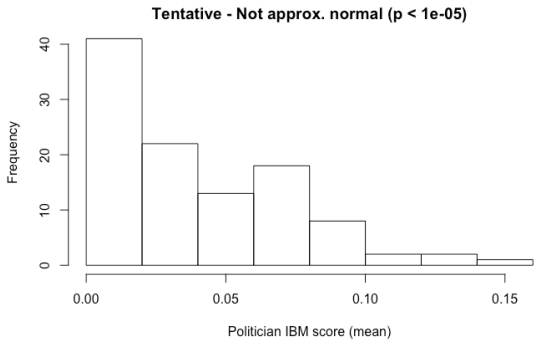
$$x_i \sim Normal(\mu_w, \sigma_w^2),$$

$$y_i \sim Normal(\mu_m, \sigma_m^2)$$

where μ_w, μ_m are the true mean score across a population of women, men, respectively. The σ_w^2, σ_m^2 are the variances of the populations of women and men - how much do individuals in the population differ in how they express that tone on Twitter?

This assumption of normality was reasonable for some tones such as sadness, joy, analytical, and confidence. For tentative, fear, and anger, the IBM tagger tagged the tweets with such sparseness that even fifty data points per user did not make the normality assumption reasonable. To see this, consider the following two plots, in the first, mean *joy* scores per politician approximate a normal distribution via the central limit theorem; in the second, extremely sparse *tentative* scores makes the histogram zero-heavy. We use the Shapiro-Wilk test to reject the normality of tentativeness while keeping that of joy.

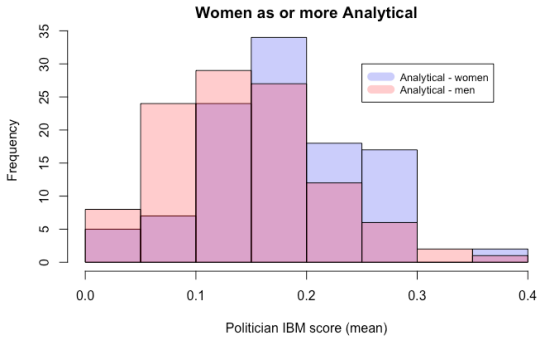




Given this model, the goal is to estimate the μ_i, σ_i for men and women for the various tones, and test the hypothesis that $\mu_w \neq \mu_m$ (there is a systematic difference between the tones by gender), and also determine what that difference is if so.

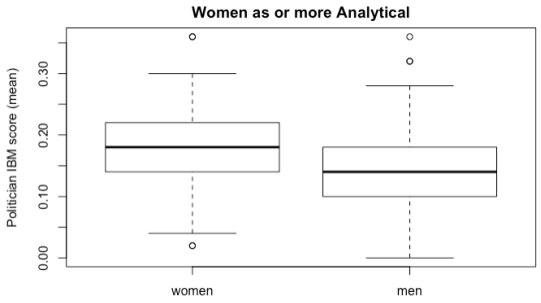
Forcefulness. The standard hypothesis from the literature review claimed that women are not as forceful as men in expressing their opinions - we thought this would be reflected in female politicians using less analytical, less confident tones in their Twitter posts. This was not the case. There are several ways to visualize that women had as or more

of these forceful tones than men. First, consider this joint histogram of analytical tone:



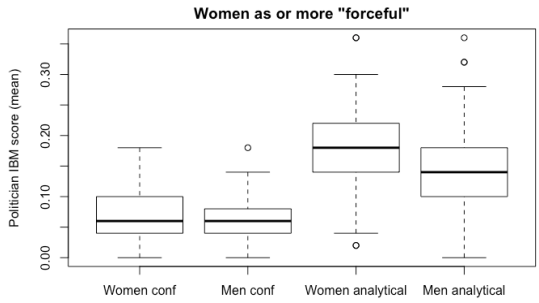
From this we can see that the mean IBM analytical score for men was tending to be lower than that of women, since the above histogram is weighted heavily on the left for men and on the right for women.

We can also view the same data a slightly different way, using boxplots.



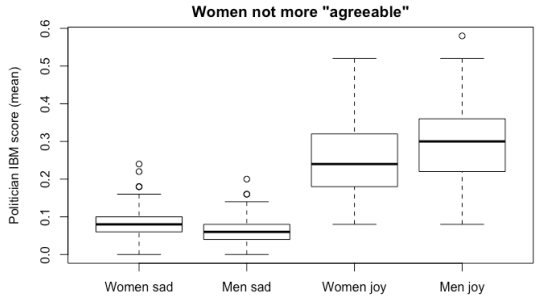
A boxplot is a way to visualize the spread of a set of points about its mean and median. The strong horizontal line in the middle is the median (the women's is higher than the men's), and the spread away from that corresponds to quantiles that capture standard deviations of the data. Outliers are also shown with dots. Clearly, the boxplot for women is as high or higher than that of the men.

A similar set of figures holds for confident tones. Although women are not clearly more confident than men, they are clearly NOT significanly less confident. We summarize female politician's forcefulness of tone with the following joint boxplots:



To formalize this process, we use the assumption of independently, normally distributed random tone scores with unknown variances to motivate a Welch Two Sample t-test to assess the viability of the null hypothesis that $\mu_w = \mu_m$. I find it conclusive that women are more analytical than men ($p < .0002$). There is also no reason to reject the null for a hypothesis that men are more confident than women ($p > .86$). This sheds strong doubt on extending to politics the standard hypothesis that women are less forceful in their use of language than men.

Agreeableness. The two relevant tones for agreeableness were sadness and joy - I claim that, given the example IBM taggings, a most agreeable person would tend to express vanilla joyful tones and stay away from sad tones. Recall the standard hypothesis that women are more agreeable than men. This hypothesis was criticized by the data. First, consider the following plot.



In this we see that men seem to be more joyful and less sad in their tones than men. The Welch two-sample t-test shows that women are express more sadness ($p < .001$) and less joy ($p < .01$) than men. Again, this criticizes the standard hypothesis that female politicians are more agreeable than men.

Other Models. After these initial surprising results, I entertained fuller models that might ex-

plain the data without requiring such unintuitive systematic gendered differences in language. Specifically, I looked at party affiliation and age as other explanatory factors in the differences in mean tone values.

Additive model. An additive model does what it sounds like - it adds two explanatory variables together to make a prediction. For concreteness, consider gender and party affiliation as the two explanatory variables. The means we are interested in are $\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}$ where μ_{ij} is the mean of the tones given the i^{th} gender (Female, Male) and j^{th} party affiliation (Democrat, Republican). Let α_i be the main effect due to gender, β_j be the main effect due to party affiliation. The additive model involves

$$\mu_{ij} = \mu + \alpha_i + \beta_j,$$

and within a group such as female ($i = 1$) Republicans ($j = 2$), the scores are independently distributed with corresponding $Normal(\mu_{ij}, \sigma^2)$. Again, the σ^2 are ideosyncratic differences among individuals in the population. For example, if $\alpha_1 < \alpha_2$ and $\beta_1 > \beta_2$, then women use the tone less strongly than men, and Democrats use it more strongly than Republicans.

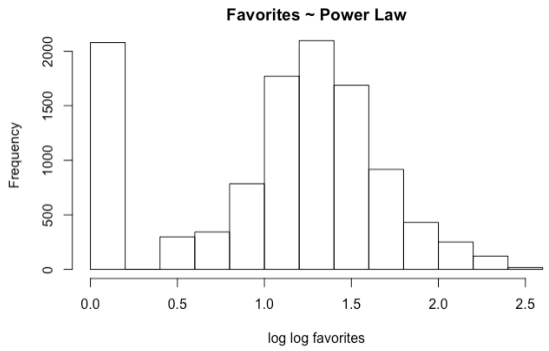
The additive model is a way to discover the true explainer of a phenomenon when when multiple explainers might be correlated with one another. For instance, more women are Democrat by proportion than men. Sure enough, under the additive model of gender and party affiliation, the earlier finding that men are more joyful than women was not robust. Instead, we find that *Republicans* are more joyful than Democrats ($p < .0005$), perhaps wanting to speak optimistically given their party's control of Congress and the presidency. However, the effect due to gender was no longer significant ($p > .1$).

The other results remain robust to fuller models also involving party affiliation, age, and / or interactions among the variables. In an additive model with gender and party, there was little evidence that the Democrats are more sad in tone ($p > .1$), and the effect due to gender remained significant: women express more sad tones ($p < .01$). Although the Democrats, perhaps in criticism of the ruling Republican party, are more analytical than Republicans ($p < 5e{-}5$), the effect due to gender remains significant: women express more analytical tones ($p < .05$).

4.2 Twitter Norms

The second main thrust of the analysis focused on the reactions of Twitter users to the various tones that came up. This measures language norms in politics more broadly, asking not just what norms women have internalized, but the factors on the voter end that might push the norms one way or the other.

Data Tranformation. The best normal approximation of favorites per tweet used a log log transformation. This is because a few rare tweets go viral and garner a huge amount of favorites while the vast majority get just a few favorites. This phenomenon is called the power law, and is well studied in social media contexts by Rastogi [23]. The following plot shows the success of the transformation, creating roughly normally distributed counts of log log favorites. Note that tweets with zero favorites still weigh heavily on the distribution. I chose not to consider these tweets as outliers, however.



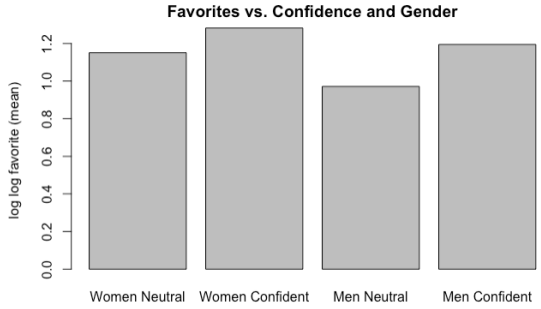
The Model. We attempt to use gender as an explanatory variable in interaction with tone. The standard hypothesis would expect that women are discouraged from using, for instance, confident tones, while men are encouraged to do so. This might manifest itself in an interaction model

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \delta_{ij}$$

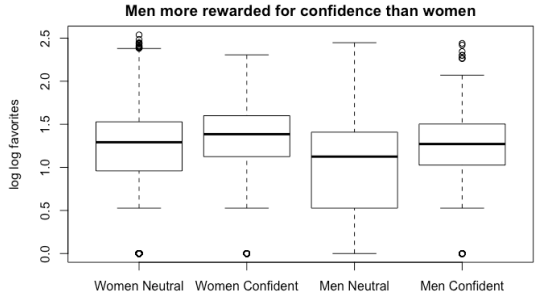
where μ_{ij} are the mean log log favorites under conditions $i \in \{1, Male; 2, Female\}$, $j \in \{1, \text{Not confident}; 2, \text{Confident}\}$. If the δ_{ij} are not zero, that might mean that (for example) while being female might increase favorites and being confident might increase favorites, being male and confident gets *more* of a boost than female and confident. And this is what we find.

Confidence. We motivate the analysis with the following plot. On the y axis is the explained

variable, our normally-distributed log log favorites per tweet. We've grouped the points along two explanatory variables: gender and confidence. We see that while women get more favorites than men, and while both do better when they're confident, men receive a bigger boost than the women when they're confident.



Is this difference statistically significant? Given the normality assumption, we construct box plots that give a notion of error on the estimates of means above.



From this diagram, it is easier to see that the boost on expected favorites that women get when confident is not quite as high as the one men receive.

Using analysis of variance, we reject the simpler additive model for the model with interactions ($p \approx .05$). I did not examine effects due to other tones.

5 Discussion

Based on the literature, we motivated the standard hypothesis that congresswomen would use systematically less forceful and more agreeable tone. To test forcefulness, we looked at confidence and analytical tones as judged by IBM's Watson Tone Analyzer. Female politicians are not less forceful. They are also not more agreeable, in the sense that they use no more straightforwardly joyful tones in their tweets than do men.

These results are surprising, so we considered other explanations. We looked at both party affiliation and age as additional explanatory variables, noticing that all but the joyfulness result was robust to deeper analysis. Another study could consider even more factors, such as seniority in Congress, race, state, etc.

Second, we turned to the users of Twitter, finding that while Twitter users favorite women more, and even more when they're confident, the boost is not as large for women as for men. When a man is more (less) confident, he is rewarded (hurt) disproportionately. This result could be symptomatic of slowly changing language norms. Another study could look into the effect over time of these norms. Obviously, Twitter as a medium is limited in its historical scope, but perhaps historical speeches and their reactions could be used.

Theory. What theory best explains these surprising results? It seems Henley's [4] thesis is substantiated, in which gendered language is a result of often being in subordinate positions. Once women attain power, they no longer need to use that stereotypical language. Another plausible explanation is the "Iron lady" hypothesis, in which women have to over compensate as politicians for being in a predominately male field.

Policy. Relating to politicians specifically, we should remain aware of any implicit biases that voters (and, generally, consumers of political media) may have related to gender. For instance, a moderator of a political debate between a man and a woman should be aware of how the same thing, if said / interjected by a woman, might not be as positively received as if by the man.

These findings, if we accept Henley's theory, also have far-reaching implications in other fields. Have female business leaders also shed these language norms? Is the use of tentative language solely caused by being in subordinate positions? What other observed differences between men and women might decrease as gender equality increases?

References

[1] Robin Lakoff. Language and woman's place. *Language in Society*, 2(1):45–80, 1973.

[2] Campbell Leaper and Rachael D. Robnett. Women are more likely than men to use tentative language, aren't they? a meta-analysis testing for gender differences and moderators. *Psychology of Women Quarterly*, 35(1):129–142, 03/01; 2018/03 2011. doi: 10.1177/0361684310392728; 04.

[3] D. Spender. *Man-made language*. Routledge & Kegan Paul, London, UK, 1984.

[4] Nancy Henley M. *Body Politics : Self and Society / N.M. Henley*. Blackwell Publishers, Malden, MA, 01/01 2001.

[5] Judith A. Hall. Nonverbal behavior, status, and gender: How do we understand their relations? *Psychology of Women Quarterly*, 30(4):384–391, 2006.

[6] Anthony Mulac and James J. Bradac. Women's style in problem solving interactions: Powerless, or simply feminine?, July 2013 1995. ID: 271777.

[7] Campbell Leaper and Melanie M. Ayres. A meta-analytic review of gender variations in adults' language use: Talkativeness, affiliative speech, and assertive speech. *Pers Soc Psychol Rev*, 11(4):328–363, 11/01; 2018/03 2007. doi: 10.1177/1088868307302221; 06.

[8] Frances Rosenbluth, Joshua Kalla, and Dawn Teele. The female political career. In *The world bank. Women in Parliaments Global Forum (WIP).*, 2015.

[9] Center for american women and politics. Technical report, Rutgers: Eagleton Institute of Politics, 2018.

[10] Dawn Teele, Joshua Kalla L., and Frances Rosenbluth. *Faces of Bias in Politics: Evidence from Elite and Voter Conjoint Experiments on Gender*. 01/01 2017.

[11] Kathleen Dolan. Gender differences in support for women candidates. *Women & Politics*, 17(2):27–41, 03/23 1997. doi: 10.1300/J014v17n02_02.

[12] Alison I. Young, Kyle G. Ratner, and Russell H. Fazio. Political attitudes bias the mental representation of a presidential candidate's face. *Psychol Sci*, 25(2):503–510, 02/01; 2018/03 2014. doi: 10.1177/0956797613510717; 04.

[13] Rachel Silbermann. Gender roles, work-life balance, and running for office. *Quarterly Journal of Political Science*, 10(2):123–153, 2015.

[14] Joshua Kalla, Frances Rosenbluth, and Dawn Langan Teele. Are you my mentor? a field experiment on gender, ethnicity, and political self-starters. *The Journal of Politics*, 80(1):337–341, 01/01; 2018/03 2018. doi: 10.1086/693984; 04.

[15] Inc Twitter. Search tweets, 2018.

[16] IBM. Watson tone analyzer, 2018.

[17] Davey Proctor. Polyspeech, 2018.

[18] Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank.

[19] Inc Twitter. Number of monthly active twitter users worldwide from 1st quarter 2010 to 4th quarter 2017 (in millions). Technical report, Statista, 2018.

[20] Bryan Anderson. Tweeter-in-chief: A content analysis of president trump's tweeting habits. *Elon Journal*, 8(2):36, 2017.

[21] David Robinson. Text analysis of trump's tweets confirms he writes only the (angrier) android half, 2018.

[22] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, 64(2):317–332, 2014.

[23] Tushti Rastogi. A power law approach to estimating fake social network accounts. *CoRR*, abs/1605.07984, 2016.