Adil Soubki, Shyne E. Choi, and Owen Rambow. 2024. Examining Gender and Power on Wikipedia through Face and Politeness. In Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 40–50, Kyoto, Japan. Association for Computational Linguistics.

Examining Gender and Power on Wikipedia Through Face and Politeness

Adil Soubki[1,3], **Shyne Choi**[1], **Owen Rambow**[2,3]
[1]Department of Computer Science, [2]Department of Linguistics
[3]Institute for Advanced Computational Science, Stony Brook University
asoubki@cs.stonybrook.edu, {shyne.choi,owen.rambow}@stonybrook.edu

Abstract

We propose a framework for analyzing discourse by combining two interdependent concepts from sociolinguistic theory: face acts and politeness. While politeness has robust existing tools and data, face acts are less resourced. We introduce a new corpus created by annotating Wikipedia talk pages with face acts and we use this to train a face act tagger. We then employ our framework to study how face and politeness interact with gender and power in discussions between Wikipedia editors. Among other findings, we observe that female Wikipedians are not only more polite, which is consistent with prior studies, but that this difference corresponds with significantly more language directed at humbling aspects of their own face. Interestingly, the distinction nearly vanishes once limiting to editors with administrative power.

1 Introduction

Brown and Levinson (1987) (henceforth B&L) introduce an influential theory of politeness based on the concept of face, which they claim to be culturally universal. In this theory, face – i.e. the public image one seeks to claim – is a two-sided coin. Agents attend to their desire to have their wants appreciated, which they call positive face, as well as a complementary desire to act unimpeded and maintain freedom, which they call negative face. The face of every agent is ensnared with that of every other agent – agents cannot have their desires appreciated if they cannot appreciate the desires of others. As a result, utterances can raise (+) or threaten (-) the positive (Pos) or negative (Neg) face of the speaker (S) or hearer (H).

A face threat or face raising is not a property of particular linguistic choices, but of communicative intent. If I want to request information from you, then I necessarily need to threaten your negative face, since, if I am successful in communicating my request to you, I will oblige you to answer and thus I will restrict your choice of actions. In B&L's theory, discourse participants may choose among various strategies for minimizing threats to face. These strategies are linguistic strategies (for example, using hedges), and the choice of strategy depends on many factors such as cultural conventions and the discourse situation (who is talking to whom under what circumstances).

Work related to NLP has concentrated on studying linguistic manifestations of politeness (Walker et al., 1997; Danescu-Niculescu-Mizil et al., 2013) while largely disregarding the notion of face act. While B&L are frequently cited, the deep insight of their theory comes from a complexity which has been ignored. Their theory is not simply about politeness, but about how politeness, situated in the context of rational action, manifests from a combination of performing face acts to achieve certain goals and using mitigation strategies to lessen the impact of face-threatening acts. Danescu-Niculescu-Mizil et al. (2013) use politeness markers inspired by B&L strategies as features for a system which predicts perceived politeness without modeling face acts. Dutt et al. (2020) predict face acts in isolation from perceived politeness. In this paper, we re-examine the Wikipedia Talk Pages Corpus (Danescu-Niculescu-Mizil et al., 2012; Chang et al., 2020) and demonstrate how bringing face acts and politeness together provides deeper insight.

We do this by producing an annotation of face acts on the corpus and training a new model to label utterances. We then use this tool, along with prior systems which produce judgements of perceived politeness, to label roughly 1.3 million sentences from Wikipedia talk pages. To our knowledge, we are the first to apply an annotation grounded in politeness theory to a text corpus of this scale.

The paper is structured as follows. We start with a review of relevant literature (§2) and present our

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 40–50
September 18–20, 2024.

theoretical framework (§3). We then turn to producing an annotation of face acts on the Wikipedia Talk Pages Corpus and building a tagger using this new dataset (§4). Our framework is then applied by bringing this new tagger together with existing tools to re-analyze the corpus, paying special attention to gender and power (§5). We end by reporting our conclusions along with a discussion of future work (§6).

All of the code written, datasets prepared, and experimental observations made in the course of this research will be made available on GitHub.[1]

2 Related Work

The theory of politeness of B&L has found applications in many fields including sociology, psychology, and linguistics. Google Scholar lists nearly 38,000 citations. Curiously, in NLP there has not been much work building explicitly on B&L. Danescu-Niculescu-Mizil et al. (2013) concentrate on one type of face-threatening act (FTA), namely the negative face-threatening act of a request, and investigate the strategies used for doing this FTA. To do this, they use crowd sourcing to rate the requests on a politeness scale. They develop a model which predicts the politeness of these requests and use it to study the interactions between users on Wikipedia and StackExchange. Ziems et al. (2023) show that fine-tuning on the data of Danescu-Niculescu-Mizil et al. (2013) substantially outperforms zero-shot approaches.

The face acts (FAs) themselves are the object of Dutt et al. (2020). In addition to developing a dataset annotated with FAs, they present a FA classifier based on a neural architecture they devise on top of BERT, which achieves 69% F-measure (0.60 macro). As the data involves participants convincing others to donate to a charity, they also use this corpus to investigate the relationship between face acts and persuasion by predicting if a participant chose to donate. This corpus, which we refer to as the "CMU Face Acts Corpus" (or "CMU Corpus" for short) in this paper, is the direct inspiration for our annotation effort on the Wikipedia data. We differ from their annotation scheme in some important details; we present our annotation in §4. In prior work, we investigated the interaction of intention (through dialog act tagging) and face acts in the CMU Corpus (Soubki and Rambow, 2024).

There has been an explosion work in computational social science in general, in which NLP tools are used to extract relevant signals from large amounts of data in order to study a social phenomenon, such as changing attitudes towards certain topics as expressed on social media. For an overview, see (Edelmann et al., 2020). In the area of studying how gender and power shape written dialogs, there has been some work in NLP. Working with corporate emails, (Prabhakaran et al., 2014) find that gender differences become exaggerated when looking at individuals with greater social power; specifically, among people with power, women behave *more* differently from men than when comparing people without power.

Finally, turning to the study of politeness and gender outside of NLP, there have been some studies based on manual analysis of collected data, for example (Herring, 1994; Tannen, 1994; Kunsmann, 2013). For space reasons, we discuss only one example in more detail. Kendall (2005), using a framing approach following (Goffman, 1974), finds that women in power who "downplay status differences (...) are exercising and constituting their authority by speaking in ways that accomplish work-related goals while maintaining the faces of their interlocutors". In the terminology of B&L (which Kendall (2005) does not use), women perform similar face acts to men but use strategies to mitigate the effects, which results in women in power appearing more polite than men in power.

3 Theoretical Framework

In this section we provide a brief summary of relevant concepts from politeness theory as it relates to our work. Our goal in this paper is to explore how face acts contribute to the perception of politeness. For B&L, "face" refers to the public self-image of agents, and it is a universal component of human interaction. It consists of two complementary facets (Brown and Levinson, 1987, §3.1, p. 61). (1) negative face: "the basic claim to territories, personal preserves, rights to non-distraction – i.e. to freedom of action and freedom from imposition." (2) positive face: "the positive consistent self-image or 'personality' (crucially including the desire that this self-image be appreciated and approved of) claimed by interactants."

A face act is an intentional communicative act which inherently interacts with the face of the speaker and/or addressees (Brown and Levinson, 1987, §3.2, p. 65). Face acts can threaten (-) or

[1] https://github.com/cogstates/wikiface

Face Act	Mnemonic	Sample Discourse Goals
HNEG-	IMPOSITION	Requests, commands, questions, offers, promises, ...
HPOS-	DISAGREEMENT	Criticism, insults, disapproval, ...
HNEG+	PERMISSIVENESS	Granting permission, making exceptions, ...
HPOS+	AGREEMENT	Seeking common ground, group cohesion, ...
SNEG-	INDEBTEDNESS	Thanking, accepting offers or thanks, commitments, ...
SPOS-	APOLOGIES	Confessions, embarrassment, ...
SNEG+	AUTONOMY	Refusing requests, asserting freedoms, ...
SPOS+	CONFIDENCE	Self-promotion, signaling virtue, ...

Table 1: Face acts with mnemonic label and examples of discourse goals.

affirm (+) the face; they can be about the speaker's face (S) or the hearer's (H); and they can be about positive (Pos) or negative (Neg) face. This gives us eight possible face acts, shown in Table 1, where we also provide a short mnemonic names which we will use in this paper, as the terminology of B&L can be unintuitive.

Face acts are part of a larger sequence of choices a speaker makes. First, the speaker chooses a discourse goal or goals (which may form a hierarchy) which will be realized in a speech act (Austin, 1962); then they determine which face acts contribute to the discourse goals; they then choose a strategy to realize this face act, in conformance with the cultural norms of their community which are mutually known by them and the hearer in the communicative context (age, gender, power differential of the discourse participants); and finally, they produce the utterance, which the hearer will perceive as more or less polite, given the discourse goal of the speaker, the communicative context, and the mutually known cultural norms. We see that the notion of "strategy" plays a crucial role in the mediation between face act performance and perceived politeness, and B&L devote a large portion of their study to strategies. Unfortunately, there are no corpora annotated for face act strategies.[2]

We emphasize that face acts do not imply perceived politeness (§B). Consider the following examples from the Wikipedia corpus.

[1] B: Why open a peer review when we are looking for someone to do the GA review?

A: Why request a second GA, 3 days after the first one failed?

[2] A: Hi Plange, any reason why this category is named differently to the others?

Both utterances are HNEG-/IMPOSITION face acts, because they impose on the hearer the obligation to respond. However, (1) rejects the previous question by B and challenges B, while (2) is just a request for information, so that (1) is perceived as more impolite than (2).

It is possible for a single utterance to perform multiple face acts at once. For example, (1) could also be seen as DISAGREEMENT, since it entails a critique of B's actions. However, Dutt et al. (2020) observed multi-labeled acts in only 2% of their data, leading them to consider a single label per utterance. We make this simplification as well in the work presented in this paper.

4 Face Act Tagging

In this section we outline the data, modeling techniques, and evaluation measures used in developing our face act tagger for Wikipedia talk pages.

4.1 Dataset

On Wikipedia, talk pages are used by editors to coordinate changes and improvements to the encyclopedia.[3] A variety of social and power dynamics are at play in these conversations which can range from discussions of bureaucratic process to heated, and sometimes personal, conflicts. The Wikipedia Talk Pages Corpus (Danescu-Niculescu-Mizil et al., 2012) collects 125,292 exchanges between 38,462 editors resulting in a total of 391,294 posts for analysis. Unlike the CMU Face Acts Corpus, where participants are on mostly level ground, editors can hold administrative privileges or greater notoriety

[2] Danescu-Niculescu-Mizil et al. (2012) use a notion of "strategy" which is defined by a grouping of lexical items that are assumed to affect the hearer's perception of politeness. They can be considered a simple approximation of the notion in B&L, and in fact helps in predicting politeness. We have chosen not to use these "stratgeies" (though they are straightforward to determine, as they are based exclusively on word matching), since we would like to address the issue in a more principled manner in the future.

[3] https://en.wikipedia.org/wiki/Wikipedia:Talk_page_guidelines

within the community, resulting in interactions with large social distance. Additionally, some editors self-identify gender on their user page.[4] This is desirable in our case as it allows us to study how these social factors interact with face and politeness.

There can be nested replies in talk pages which allow for situations where an utterance is not a reply to the preceding utterance. We do not attempt to correct for these cases and sort first to preserve reply structure and then by the time of the post.

4.2 Annotation

Similar to the CMU Corpus, we use the criteria outlined by B&L, which serves as our reference. The CMU Corpus annotation guidelines, as the authors noted, contain some departures from politeness theory. In particular, the CMU Corpus annotates both thanking and complimenting as AGREEMENT. In contrast, B&L analyze thanking and complimenting as INDEBTEDNESS and IMPOSITION, respectively. We choose to remain faithful to B&L, and in fact assert this to be a critical piece of the theory. Consider a compliment such as *you have a lovely smile*. How is it that a compliment can be taken so poorly by the addressee if the speaker is not risking anything? They are often very risky social acts because the speaker assumes they are among the people their addressee wishes to be complimented by; a very imposing assumption. Thanking, on the other hand, can be seen as an exchange of currency. Similar to writing an IOU, the speaker offers a token of their freedom to the addressee. We note that we expect future versions of face act annotations to annotate multiple face acts at once, which may resolve this difference between the CMU Corpus annotation style and ours.

We randomly selected 200 conversations from the WikiTalks data for manual annotation. As the posts contain multiple sentences, each with the possibility of their own face act, we segment the sentences prior to annotation using spaCy (Honnibal and Johnson, 2015). To reduce errors in segmentation, we scrubbed hypertext tags and masked any remaining urls. This resulted in 1850 sentences. We will refer to these basic units of annotation as "utterances" in the following sections. Two of the authors annotated the 1850 utterances for face acts. We examined 100 utterances labeled by both annotators and computed a Cohen's Kappa of 0.69 which indicates moderate to substantial agreement.

[4] https://en.wikipedia.org/wiki/Wikipedia:User_pages

4.3 Modeling

We model face act tagging as a text classification task. Given a sequence of n utterances $S = [t_1, t_2, \ldots, t_n]$, we wish to assign a label $y \in Y$ where Y represents a set containing the 8 possible face acts and one additional label for no face act. Recently, many classification tasks have achieved stronger results using parameter efficient fine-tuning methods of larger models rather than full fine-tuning smaller ones (Hu et al., 2022; Dettmers et al., 2024). We adopt this approach and use Llama-3-8B (AI@Meta, 2024) and LoRA with Int8 quantization (Dettmers et al., 2022) for fine-tuning.[5] Details of the configuration are given in Appendix A.

4.4 Data Representation

While fine-tuning approaches unify many aspects of the model design, they present challenges when it comes to determining effective input and output representations.

We provide the models an input which contains an utterance prefixed with the Wikipedia username of the discourse participants,[6] along with previous utterances as context. Each utterance is followed by a newline character. We give an example with two lines of context, though in our experiments we use more, as discussed just below.

```
[Input]
  Jossi: I will.
  Jossi: Just play nice, that is all I ask.
  Kelly: What's that supposed to mean?
[Output]
  hpos-
```

The target output is a distribution where the highest probability is given to the correct label for the final utterance of the input text, in this case HPOS-(DISAGREEMENT). We experimented with different output formats, and found they do not make much of a difference. In our experiments we noticed context to be a critical factor with the optimal size varying by model. Llama 3 performed best with a size of four, for a total of five utterances. As there are no previous turns for the first four turns in each dialog, those examples are provided in a similar format containing only three, two, one or no lines of context.

[5] Our choice of Llama-3 was informed by a preliminary set of experiments in which a variety of pre-trained models and methods were were examined on single seed runs.

[6] We note that the Wikipedia usernames shield the actual identity of the discourse participant, and that the Wikipedia username is public.

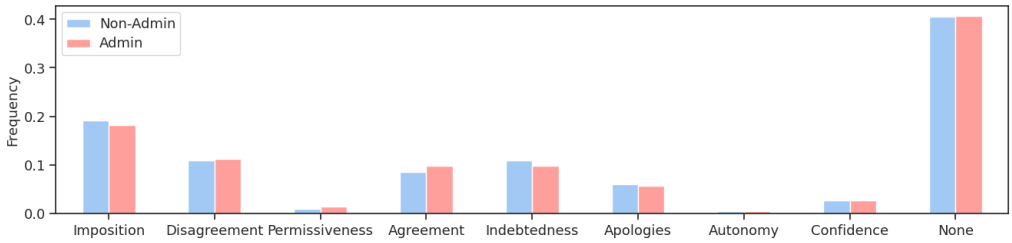

Figure 1: Frequency of face acts for admins and non-admins.

4.5 Experimental Setup And Evaluation

We perform all experiments using five-fold cross validation and the evaluation metrics are averaged across all five folds. We evaluate model performance using F-measure for each of the nine classes as well as micro and macro F-measure aggregated over all labels. We performed hyperparameter tuning, and report metrics only for the best model.

4.6 Results

The results of these experiments are reported in Table 2. We achieve a micro-averaged F1 of 0.68 (average across five folds). Since the task is, with the exception of some nuances (§4.2), identical to the CMU Face Acts Corpus we also tried continued training on the CMU Face Acts Corpus, but this did not improve performance. We suspect this is due to the difference in genre and slight change in annotation procedure, which results in a different distribution of labels between the two datasets.

5 Application and Analysis

We apply our new face act tagger along with the politeness scores provided by ConvoKit (Chang et al., 2020) to study the interactions of face and politeness over the entire Wikipedia Talk Pages Corpus. Our face act tagger is trained using our entire annotation (§4.2) before applying it to the Wikipedia data. This produces roughly 1.3 million sentences labeled with face acts and perceived politeness. We note that the politeness scores are obtained for the entire turn, as this is what the perceived politeness model is trained on, while face acts are tagged by sentence to allow for greater granularity.

Micro	0.68
Macro	0.51
IMPOSITION	0.73
DISAGREEMENT	0.56
PERMISSIVENESS	0.40
AGREEMENT	0.58
INDEBTEDNESS	0.80
APOLOGIES	0.56
AUTONOMY	0.04
CONFIDENCE	0.14
NONE	0.76

Table 2: Mean F1 across all folds of our annotation.

In our analysis of politeness we investigate how polite (magnitude) editors are perceived to be by looking at their scores and how often that occurs (frequency) by considering the proportion of utterances in the top 25% of politeness scores. For face acts, we compare the overall distribution (frequency) of labels. Statistical significance is calculated using the Mann-Whitney U test. This analysis was also performed on only the human annotated portion of the data and the trends remained consistent. We report results on the entire corpus.

5.1 Admin Differences

On Wikipedia, editors with administrative status wield significant power in the community including the ability to block or unblock users by IP address and delete or restore pages. This increased status is known to be recognized in the community (Danescu-Niculescu-Mizil et al., 2012; Burke and Kraut, 2008; Leskovec et al., 2010) which endows editors with these powers through public elections. We note that politeness theory anticipates speakers with greater social power than their addressee to more often select strategies that reduce ambiguity and lengthiness. This means opting to perform face threatening acts more often (as opposed to avoiding them all together) and mitigating them through the trade-offs of strategies less often, which one would expect to correspond with a perception of being less polite overall.

We divide utterances by their politeness score into the polite utterances (top 25%), neutral (next 50%) and impolite (bottom 25%). When compar-

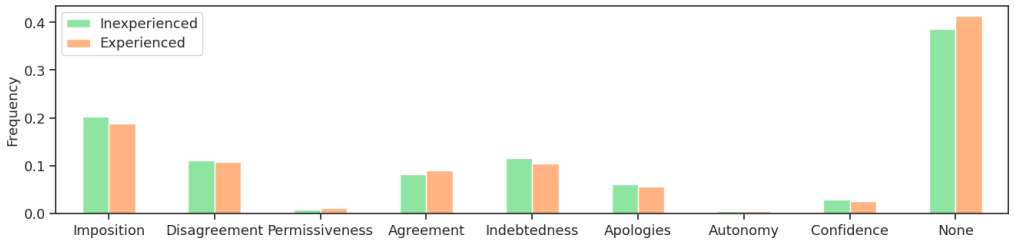

Figure 2: Frequency of face acts by editor experience

ing politeness between admins and non-admins we see the same trend as observed by Danescu-Niculescu-Mizil et al. (2013). Utterances produced by editors with administrative privileges ("admins") are not more often impolite, however they are significantly ($p < 0.001$ using the Mann-Whitney U test) less frequently polite, with a mean score difference of 3. Additionally the frequency by which admins produce polite posts is also significantly ($p < 0.001$) lower resulting in messages which are deemed polite 5% less often compared to non-admin editors.

When looking at the distributional differences in face acts by adminship (Figure 1) this decrease in politeness corresponds with small, but salient variations. Admins are significantly ($p < 0.001$) less likely to express INDEBTEDNESS (e.g. thanking, accepting offers) and APOLOGIES (e.g., admitting mistakes, confessions). Though admins produce more utterances labeled AGREEMENT (e.g. appreciation, seeking common ground, group cohesion), their AGREEMENT utterances are significantly ($p < 0.001$) less often (-4% absolute) perceived as polite compared to AGREEMENT utterances by non-admins. Similarly, while non-admins do more IMPOSITION (e.g. issuing commands, making requests), their IMPOSITION utterances are significantly ($p < 0.05$) more often ($+3\%$ absolute) taken politely compared to IMPOSITION utterances by admins.. This shows, as we anticipated, that face acts do not imply politeness, contrary to possible intuition.

5.2 Experience Differences

We explore whether the experience and productivity of the editor is another means to achieve increased social power without the explicit additional privileges the "admin" title confers. To investigate this we categorize users by the number of edits they have made and label users in the top and bottom quartiles "experienced" and "inexperienced", respectively.

	Politeness
Experienced Admin	$0.34^{\dagger}$
Experienced Non-Admin	$0.36^{\dagger}$
Inexperienced Admin	$0.38^{\dagger}$
Inexperienced Non-Admin	$0.40^{\dagger}$

Table 3: Mean politeness scores for difference admin types. All differences are found to be significant using the Mann-Whitney U test with $p < 0.001$.

	Inexperienced	**Experienced**
Impolite	0.07	0.07
Polite	$0.35^{\ddagger}$	$0.28^{\ddagger}$

Table 4: Proportion of turns classified as (im)polite by editor experience level. ‡ indicates significance with $p < 0.0001$ using the Mann-Whitney U test.

We observe similar trends in politeness among experienced editors (Table 4) to that of admins, with turns by experienced editors being labeled polite 7% less often relative to inexperienced editors. When looking at the differences in face acts (Figure 2) we note that there are ways in which newcomers behave like experienced Wikipedians such as a willingness to the face act DISAGREEMENT. However, like admins, experienced users are significantly ($p < 0.001$) less likely to express INDEBTEDNESS or APOLOGIES. Unlike when comparing by admin status, we find that experienced admins are significantly ($p < 0.001$) less likely to interact with face all together (more labeled NONE).

We now investigate how experience interacts with admin status. As expected, experience is correlated ($r = 0.37$) with adminship with nearly half of all admins landing in the top quartile of editors by edit count. We find admins in the top quartile by edit count are significantly ($p < 0.001$) less polite than the bottom quartile. Additionally, intersecting experience with admin status (Table 3) finds a spectrum. Experienced admins are the least polite but

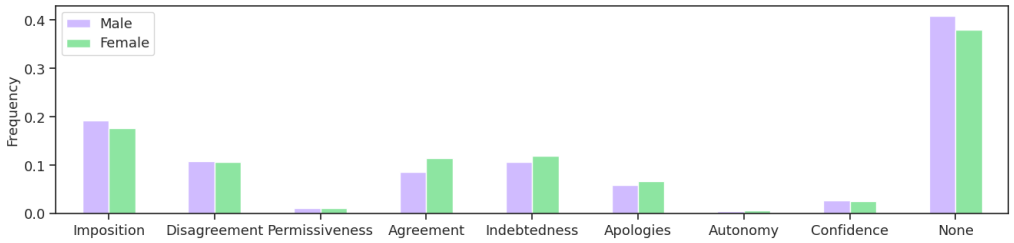

Figure 3: Frequency of face acts by gender.

experienced non-admins are less polite than inexperienced admins. This indicates that these factors are additive in their contribution to social power.

5.3 Gender Differences

Some editors self-identify their gender on their user page allowing us to study communicative differences along this axis as well. Prior work found female Wikipedians to be generally more polite (Danescu-Niculescu-Mizil et al., 2013) which is consistent with studies in several domains. We also observe this, with utterances by women scoring more polite (+5, $p < 0.001$), more often (+7%, $p < 0.0001$).

When comparing the distribution of face acts (Figure 3) we see several disparities that the politeness scores alone do not convey. In general, the NONE category is lower for women, i.e. female Wikipedians are more likely, and perhaps more willing, to interact with face in their utterances. When doing so, they humble their own positive face (APOLOGIES, e.g. admitting mistakes, making confessions, accepting compliments) and their own negative face (INDEBTEDNESS, e.g. thanking, accepting apologies) more often than men. This self-deference is accompanied by fewer impositions on their addressee's face (IMPOSITION, e.g. requests, commands, insults, criticism) and more attention to the hearer's own wants (AGREEMENT, e.g. seeking common ground, showing respect). Unlike when looking at admins, these AGREEMENT utterances are less frequently judged to be impolite. These trends have been observed in various prior studies (Lakoff, 1973; Prabhakaran and Rambow, 2017; Herring, 1994).

5.4 Intersectional Differences

We have seen that male Wikipedians are less polite, more distant with regards to face, and more likely to express IMPOSITION (§5.3). Similarly, much of the same is true when comparing admins to non-admins (§5.1). How do these factors interact? As mentioned in §2, previous work in other domains has found gender differences to become exaggerated in the communication patterns of individuals with power. One might expect a similar trend to hold on Wikipedia.

	Male	Female
Non-Admin‡	0.37	0.43
Admin	0.34	0.35
Inexperienced‡	0.41	0.43
Experienced‡	0.34	0.42

Table 5: Mean politeness scores by experience and admin status compared across gender. ‡ indicates significance with $p < 0.0001$ using the Mann-Whitney U test when comparing across gender.

When comparing politeness across both gender and administrative status (Table 5), we find that this does not appear to be the case. While women admins are more polite (magnitude) than male admins, the difference is not significant ($p > 0.1$). Meanwhile, their non-admin counterparts are significantly more polite than non-admin men (+6, $p < 0.0001$). Among non-admin editors, women produce utterances in the top quartile of politeness 10% more often than men, while this reduces to just 1% when comparing admins across genders.

Overall the distribution of face acts (Figure 4) between male and female admins is similar to that of non-admins (the red lines for admins and blue lines for non-admins in Figure 4 are in the same direction), except that the difference between men and women is reduced (the red lines are shorter than the blue lines).There is one striking exceptions: among non-admins, men make many more IMPOSITION (e.g., making requests, issuing commands) face acts than women, but this difference disappears for admins (and in fact women perform IMPOSITION utterances slightly more frequently than men). We note that IMPOSITION is the face act that becoming an admin specifically entitles the

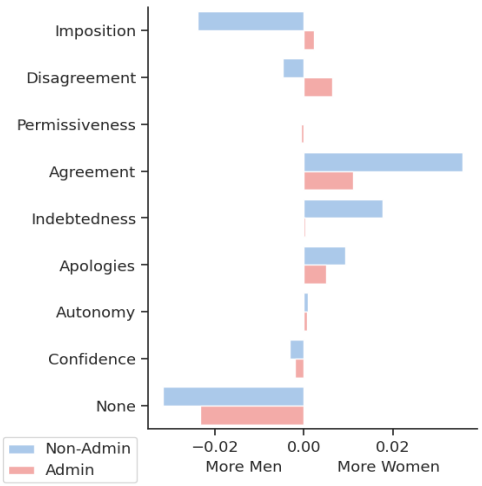

Figure 4: Differences between relative usage of face acts by gender, broken down by non-admins (blue) and admins (red); lines to the right (left) indicate that women (men) perform the face act more often

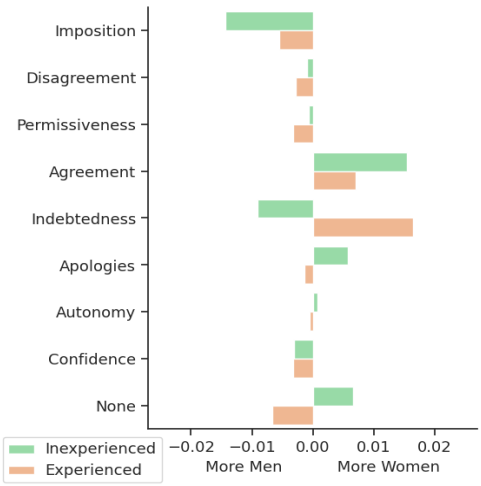

Figure 5: Differences between relative usage of face acts by gender, broken down by inexperienced (green) and experienced (orange) users; lines to the right (left) indicate that women (men) perform the face act more often

editor to perform: admins have the right to request changes (and that changes be undone). We speculate that female admins specifically make use of their socially sanctioned power, while men perform IMPOSITION acts even when having no specific admin authority. In summary, admin privileges maintain but substantially lessen the previously observed gender differences in politeness and face. Put differently, female admins behave more like men (whether admins or not), which we also saw in the politeness scores (Table 5).

We now turn to the intersection of gender and experience. Here, we see a strikingly different result. For all conditions (non-admin, admin, inexperienced, experienced), women are more polite. However, we see from Table 5 that men become more impolite as they become experienced, while this is not the case for women: there is no significant change in their politeness as they become experienced. The only exception is for women who become admins (who are, often, experienced), who behave as men do. Put differently, experience and the official power designator of "admin" do not function in the same way across gender: for men, both result in less politeness, but for women, only the "admin" title does.

When looking at face acts (Figure 5), we see that for some categories the differences between men and women are reduced with experience (the orange bars are shorter than the green bars). However, a notable exception is for INDEBTEDNESS, for which we see a large increase in the difference between men and women, and in fact a flip in which gender performs it more often. When looking at the absolute numbers (not shown in the table), we can see why: women do not change the frequency of their INDEBTEDNESS utterances at all as they gain experience, while men decrease their frequency of INDEBTEDNESS utterances from 12.3% to 9.8% of their utterances. This decrease is a major contributor to the decrease in politeness among experienced men (but not among experienced women). We extend our previous interpretation by speculating that experienced women do not feel they have a socially sanctioned position of power, and/or men experience a decrease in social distance towards other Wikipedians as they become more experienced, while women do not.

6 Conclusion and Future Work

We identify an optimized method for training face act taggers using fine-tuning on LLMs, contribute a new corpus annotated for face acts, and make available a pre-trained model for use on Wikipedia. Through several methods of analysis we demonstrate the usefulness of examining perceived politeness in combination with face acts by reporting a number of findings based on their interaction. In future work we plan to allow multiple face acts per utterance (including for the same segment), and to incorporate the strategy (as conceived of by B&L) more explicitly into our modeling framework.

Limitations

The principal scientific limitation of this work is that we could only consider three aspects of the larger model of B&L: face acts, the communicative setting (gender and power), and perceived politeness. The major missing elements in the full framework include intention, communicative intention, social norms, and strategies. We intend this paper to be a first step towards a fuller implementation of an explicit cognitive theory of communication which involves all of the mentioned elements.

The experiments for this work were performed using computational resources that are not, in general, freely available. In part due to these computational requirements, but also a result of minimal data, we were not able to evaluate the techniques on additional languages and acknowledge the limitations this places on extending our results to other cultures. We also note along similar lines that while Brown and Levinson (1987) claim their theory of politeness to be culturally universal, this claim has been contested – most notably for eastern cultures (Al-Duleimi et al., 2016). As discussed in detail above, taking utterances to have a single face act or intent is a critically limiting assumption which lends some uncertainty to our conclusions.

We note that while many of the linguistic differences observed were consistent across multiple rounds of analysis and significant using the Mann-Whitney U test, the effect sizes were generally small. The conclusions should be interpreted with that in mind.

Ethics Statement

Despite an analysis of the errors, we cannot verify the safety of this system in any user-oriented context and therefore do not recommend such uses without further study. While we do not produce any datasets directly from human annotations, we do use several datasets which were, to the best of our knowledge, compiled ethically. As the primary object of study in this work is the relationship between politeness and language, we do not anticipate broad risks to its application.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under the CCU (No. HR001120C0037, PR No. HR0011154158, No. HR001122C0034) program. Soubki has received additional support from the National Science Foundation (NSF) under No. 2125295 (NRT-HDR: Detecting and Addressing Bias in Data, Humans, and Institutions). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF or DARPA.

We thank both the Institute for Advanced Computational Science and the Institute for AI-Driven Discovery and Innovation at Stony Brook for access to the computing resources needed for this work. These resources were made possible by NSF grant No. 1531492 (SeaWulf HPC cluster maintained by Research Computing and Cyberinfrastructure) and NSF grant No. 1919752 (Major Research Infrastructure program), respectively.

We would also like to thank our anonymous reviewers for their perceptive comments, which improved this work.

References

AI@Meta. 2024. The llama 3 herd of models.

Hutheifa Y. Al-Duleimi, Sabariah Hj Md Rashid, and Ain Nadzimah Abdullah. 2016. A critical review of prominent theories of politeness. *Advances in Language and Literary Studies*, 7:262–270.

J. L. Austin. 1962. *How to do things with words*. Oxford University Press.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Moira Burke and Robert Kraut. 2008. Taking up the mop: Identifying future wikipedia administrators. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '08, page 3441–3446, New York, NY, USA. Association for Computing Machinery.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*, pages 699–708.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013.

A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332. Curran Associates, Inc.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Ritam Dutt, Rishabh Joshi, and Carolyn Rose. 2020. Keeping up appearances: Computational modeling of face acts in persuasion oriented discussions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7473–7485, Online. Association for Computational Linguistics.

Achim Edelmann, Tom Wolff, Danielle Montagne, and Christopher A. Bail. 2020. Computational social science and sociology. *Annual Review of Sociology*, 46(1):61–81.

Erving Goffman. 1974. *Frame Analysis: An Essay on the Organization of Experience*. Northeastern University Press, Boston, MA.

Susan C Herring. 1994. s. In *Cultural performances: Proceedings of the third Berkeley women and language conference*, pages 278–294.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with lora.

Shari Kendall. 2005. Creating gendered demeanors of authority at work and at home. In Janet Holmes and Miriam Meyerhoff, editors, *The Handbook of Language and Gender*. Blackwell.

Peter Kunsmann. 2013. Gender, status and power in discourse behavior of men and women. *Linguistik Online*, 5(1).

Robin Tolmach Lakoff. 1973. Language and woman's place. *Language in Society*, 2:45 – 79.

Jure Leskovec, Daniel P. Huttenlocher, and Jon M. Kleinberg. 2010. Governance in social media: A case study of the wikipedia promotion process. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Vinodkumar Prabhakaran and Owen Rambow. 2017. Dialog structure through the lens of gender, gender environment, and power. *Dialogue Discourse*, 8:21–55.

Vinodkumar Prabhakaran, Emily E. Reid, and Owen Rambow. 2014. Gender and power: How gender and gender environment affect manifestations of power. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1965–1976, Doha, Qatar. Association for Computational Linguistics.

Adil Soubki and Owen Rambow. 2024. Intention and face in dialog. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9143–9153, Torino, Italia. ELRA and ICCL.

Deborah Tannen. 1994. *Talking from 9 to 5: Women and Men in the Workplace: Language, Sex and Power*. Avon Books, New York.

Marilyn A. Walker, Janet E. Cahn, and Stephen J. Whittaker. 1997. Improvising linguistic style: Social and affective bases for agent personality. In *Proceedings of the First International Conference on Autonomous Agents*, AGENTS '97, page 96–105, New York, NY, USA. Association for Computing Machinery.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? ArXiv preprint arXiv:2305.03514.

A Configuration Details for Experiments

For all experiments we fine-tune Llama-3-8B on each of the five cross-validation folds with a batch size of 1 and no gradient accumulation steps. The AdamW optimizer is configured with a learning rate of 2e-5, weight decay of 0, and epsilon of 1e-8. As the cross-validation preparation does not contain a development set to conserve data, we train for a fixed 10 epochs. We configure LoRA with α of 16, dropout of 0.1, and r of 64. Since r is somewhat large, we observed slightly better results using rank-stabilization which scales adapters during forward passes by a factor of $\alpha/\sqrt{r}$, instead of the typical α/r (Kalajdzievski, 2023). These parameters were arrived at through a run of hyperparameter tuning experiments.

B Supplementary Correlation Analysis

This analysis was performed based our model (§4) output on the Wikipedia Talk Pages Corpus. Aside from INDEBTEDNESS (e.g. thanking, commitments, accepting offers), DISAGREEMENT (e.g. criticism, insults, disapproval), and NONE (avoiding face altogether) the correlations have fairly low magnitude (absolute value less than 0.1).

	Politeness	**Impoliteness**
IMPOSITION	0.01	0.05
DISAGREEMENT	-0.11	0.18
PERMISSIVENESS	-0.01	0.01
AGREEMENT	0.03	-0.04
INDEBTEDNESS	0.31	-0.25
APOLOGIES	0.04	-0.07
AUTONOMY	0.00	-0.01
CONFIDENCE	-0.01	-0.01
None	-0.17	0.06

Table 6: Pearson's correlation coefficients between politeness scores and face acts.