A computational approach to politeness with application to social factors

Cristian Danescu-Niculescu-Mizil[*‡], **Moritz Sudhof**[†], **Dan Jurafsky**[†],
Jure Leskovec[*], and **Christopher Potts**[†]
[*]Computer Science Department, [†]Linguistics Department
[*†]Stanford University, [‡]Max Planck Institute SWS
cristiand|jure@cs.stanford.edu, sudhof|jurafsky|cgpotts@stanford.edu

Abstract

We propose a computational framework for identifying linguistic aspects of politeness. Our starting point is a new corpus of requests annotated for politeness, which we use to evaluate aspects of politeness theory and to uncover new interactions between politeness markers and context. These findings guide our construction of a classifier with domain-independent lexical and syntactic features operationalizing key components of politeness theory, such as *indirection*, *deference*, *impersonalization* and *modality*. Our classifier achieves close to human performance and is effective across domains. We use our framework to study the relationship between politeness and social power, showing that polite Wikipedia editors are more likely to achieve high status through elections, but, once elevated, they become less polite. We see a similar negative correlation between politeness and power on Stack Exchange, where users at the top of the reputation scale are less polite than those at the bottom. Finally, we apply our classifier to a preliminary analysis of politeness variation by gender and community.

1 Introduction

Politeness is a central force in communication, arguably as basic as the pressure to be truthful, informative, relevant, and clear (Grice, 1975; Leech, 1983; Brown and Levinson, 1978). Natural languages provide numerous and diverse means for encoding politeness and, in conversation, we constantly make choices about where and how to use these devices. Kaplan (1999) observes that "people desire to be *paid* respect" and identifies honorifics and other politeness markers, like *please*, as "the coin of that payment". In turn, politeness markers are intimately related to the power dynamics of social interactions and are often a decisive factor in whether those interactions go well or poorly (Gyasi Obeng, 1997; Chilton, 1990; Andersson and Pearson, 1999; Rogers and Lee-Wong, 2003; Holmes and Stubbe, 2005).

The present paper develops a computational framework for identifying and characterizing politeness marking in requests. We focus on requests because they involve the speaker imposing on the addressee, making them ideal for exploring the social value of politeness strategies (Clark and Schunk, 1980; Francik and Clark, 1985). Requests also stimulate extensive use of what Brown and Levinson (1987) call *negative politeness*: speaker strategies for minimizing (or appearing to minimize) the imposition on the addressee, for example, by being indirect (*Would you mind*) or apologizing for the imposition (*I'm terribly sorry, but*) (Lakoff, 1973; Lakoff, 1977; Brown and Levinson, 1978).

Our investigation is guided by a new corpus of requests annotated for politeness. The data come from two large online communities in which members frequently make requests of other members: Wikipedia, where the requests involve editing and other administrative functions, and Stack Exchange, where the requests center around a diverse range of topics (e.g., programming, gardening, cycling). The corpus confirms the broad outlines of linguistic theories of politeness pioneered by Brown and Levinson (1987), but it also reveals new interactions between politeness markings and the morphosyntactic context. For example, the politeness of *please* depends on its syntactic position and the politeness markers it co-occurs with.

Using this corpus, we construct a politeness classifier with a wide range of domain-independent lexical, sentiment, and dependency features operationalizing key components of po-

liteness theory, including not only the negative politeness markers mentioned above but also elements of *positive politeness* (gratitude, positive and optimistic sentiment, solidarity, and inclusiveness). The classifier achieves near human-level accuracy across domains, which highlights the consistent nature of politeness strategies and paves the way to using the classifier to study new data.

Politeness theory predicts a negative correlation between politeness and the power of the requester, where power is broadly construed to include social status, authority, and autonomy (Brown and Levinson, 1987). The greater the speaker's power relative to her addressee, the less polite her requests are expected to be: there is no need for her to incur the expense of paying respect, and failing to make such payments can invoke, and hence reinforce, her power. We support this prediction by applying our politeness framework to Wikipedia and Stack Exchange, both of which provide independent measures of social status. We show that polite Wikipedia editors are more likely to achieve high status through elections; however, once elected, they become less polite. Similarly, on Stack Exchange, we find that users at the top of the reputation scale are less polite than those at the bottom.

Finally, we briefly address the question of how politeness norms vary across communities and social groups. Our findings confirm established results about the relationship between politeness and gender, and they identify substantial variation in politeness across different programming language subcommunities on Stack Exchange.

2 Politeness data

Requests involve an imposition on the addressee, making them a natural domain for studying the inter-connections between linguistic aspects of politeness and social variables.

Requests in online communities We base our analysis on two online communities where requests have an important role: the Wikipedia community of editors and the Stack Exchange question-answer community.[1] On Wikipedia, to coordinate on the creation and maintenance of the collaborative encyclopedia, editors can interact with each other on user talk-pages;[2] requests posted on a user talk-page, although public, are generally directed to the owner of the talk-page. On Stack Exchange, users often comment on existing posts requesting further information or proposing edits; these requests are generally directed to the authors of the original posts.

Both communities are not only rich in user-to-user requests, but these requests are also part of consequential conversations, not empty social banter; they solicit specific information or concrete actions, and they expect a response.

Politeness annotation Computational studies of politeness, or indeed any aspect of linguistic pragmatics, demand richly labeled data. We therefore label a large portion of our request data (over 10,000 utterances) using Amazon Mechanical Turk (AMT), creating the largest corpus with politeness annotations (see Table 1 for details).[3]

We choose to annotate requests containing exactly two sentences, where the second sentence is the actual request (and ends with a question mark). This provides enough context to the annotators while also controlling for length effects. Each annotator was instructed to read a batch of 13 requests and consider them as originating from a co-worker by email. For each request, the annotator had to indicate how polite she perceived the request to be by using a slider with values ranging from "very impolite" to "very polite".[4] Each request was labeled by five different annotators.

We vetted annotators by restricting their residence to be in the U.S. and by conducting a linguistic background questionnaire. We also gave them a paraphrasing task shown to be effective for verifying and eliciting linguistic attentiveness (Munro et al., 2010), and we monitored the annotation job and manually filtered out annotators who submitted uniform or seemingly random annotations.

Because politeness is highly subjective and annotators may have inconsistent scales, we applied the standard z-score normalization to each worker's scores. Finally, we define the politeness score (henceforth *politeness*) of a request as the average of the five scores assigned by the annotators. The distribution of resulting request scores (shown in Figure 1) has an average of 0 and stan-

[1] http://stackexchange.com/about

[2] http://en.wikipedia.org/wiki/Wikipedia:User_pages

[3] Publicly available at http://www.mpi-sws.org/~cristian/Politeness.html

[4] We used non-categorical ratings for finer granularity and to help account for annotators' different perception scales.

domain	#requests	#annotated	#annotators
Wiki	35,661	4,353	219
SE	373,519	6,604	212

Table 1: Summary of the request data and its politeness annotations.

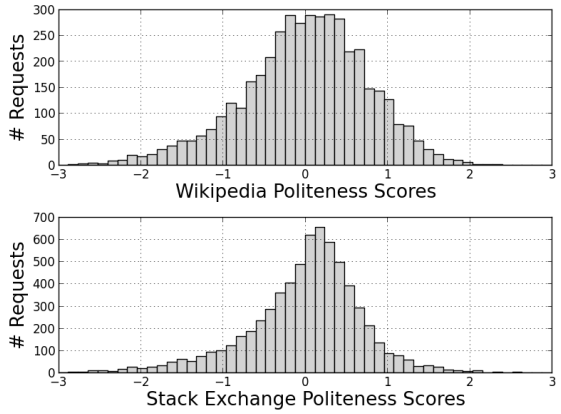

Figure 1: Distribution of politeness scores. Positive scores indicate requests perceived as polite.

dard deviation of 0.7 for both domains; positive values correspond to polite requests (i.e., requests with normalized annotations towards the "very polite" extreme) and negative values to impolite requests. A summary of all our request data is shown in Table 1.

Inter-annotator agreement To evaluate the reliability of the annotations we measure the inter-annotator agreement by computing, for each batch of 13 documents that were annotated by the same set of 5 users, the mean pairwise correlation of the respective scores. For reference, we compute the same quantities after randomizing the scores by sampling from the observed distribution of politeness scores. As shown in Figure 2, the labels are coherent and significantly different from the randomized procedure ($p < 0.0001$ according to a Wilcoxon signed rank test).[5]

Binary perception Although we did not impose a discrete categorization of politeness, we acknowledge an implicit binary perception of the phenomenon: whenever an annotator moved a slider in one direction or the other, she made a binary politeness judgment. However, the boundary between somewhat polite and somewhat impolite requests can be blurry. To test this intuition, we break the set of annotated requests into four groups, each corresponding to a politeness score quartile. For each quartile, we compute the percentage of requests for which all five annotators made the same binary politeness judgment. As shown in Table 2, full agreement is much more common in the 1st (bottom) and 4th (top) quartiles than in the middle quartiles. This suggests that the politeness scores assigned to requests that are only somewhat polite or somewhat impolite are less reliable and less tied to an intuitive notion of binary politeness. This discrepancy motivates our choice of classes in the prediction experiments (Section 4) and our use of the top politeness quartile (the 25% most polite requests) as a reference in our subsequent discussion.

[5]The commonly used Cohen/Fleiss Kappa agreement measures are not suitable for this type of annotation, in which labels are continuous rather than categorical.

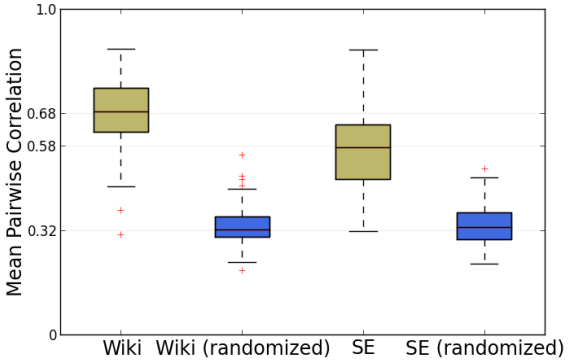

Figure 2: Inter-annotator pairwise correlation, compared to the same measure after randomizing the scores.

Quartile:	1st	2nd	3rd	4th
Wiki	62%	8%	3%	51%
SE	37%	4%	6%	46%

Table 2: The percentage of requests for which all five annotators agree on binary politeness. The 4th quartile contains the requests with the top 25% politeness scores in the data. (For reference, randomized scoring yields agreement percentages of <20% for all quartiles.)

3 Politeness strategies

As we mentioned earlier, requests impose on the addressee, potentially placing her in social peril if she is unwilling or unable to comply. Requests therefore naturally give rise to the negative po-

liteness strategies of Brown and Levinson (1987), which are attempts to mitigate these social threats. These strategies are prominent in Table 3, which describes the core politeness markers we analyzed in our corpus of Wikipedia requests. We do not include the Stack Exchange data in this analysis, reserving it as a "test community" for our prediction task (Section 4).

Requests exhibiting politeness markers are automatically extracted using regular expression matching on the dependency parse obtained by the Stanford Dependency Parser (de Marneffe et al., 2006), together with specialized lexicons. For example, for the hedges marker (Table 3, line 19), we match all requests containing a nominal subject dependency edge pointing out from a hedge verb from the hedge list created by Hyland (2005). For each politeness strategy, Table 3 shows the average **politeness** score of the respective requests (as described in Section 2; positive numbers indicate polite requests), and their **top politeness quartile** membership (i.e., what percentage fall within the top quartile of politeness scores). As discussed at the end of Section 2, the top politeness quartile gives a more robust and more intuitive measure of politeness. For reference, a random sample of requests will have a 0 politeness score and a 25% top quartile membership; in both cases, larger numbers indicate higher politeness.

Gratitude and deference (lines 1–2) are ways for the speaker to incur a social cost, helping to balance out the burden the request places on the addressee. Adopting Kaplan (1999)'s metaphor, these are the coin of the realm when it comes to paying the addressee respect. Thus, they are indicators of positive politeness.

Terms from the sentiment lexicon (Liu et al., 2005) are also tools for positive politeness, either by emphasizing a positive relationship with the addressee (line 4), or being impolite by using negative sentiment that damages this positive relationship (line 5). Greetings (line 3) are another way to build a positive relationship with the addressee.

The remainder of the cues in Table 3 are negative politeness strategies, serving the purpose of minimizing, at least in appearance, the imposition on the addressee. Apologizing (line 6) deflects the social threat of the request by attuning to the imposition itself. Being indirect (line 9) is another way to minimize social threat. This strategy allows the speaker to avoid words and phrases conventionally associated with requests. First-person plural forms like *we* and *our* (line 15) are also ways of being indirect, as they create the sense that the burden of the request is shared between speaker and addressee (*We really should* . . .). Though indirectness is not invariably interpreted as politeness marking (Blum-Kulka, 2003), it is nonetheless a reliable marker of it, as our scores indicate. What's more, direct variants (imperatives, statements about the addressee's obligations) are less polite (lines 10–11).

Indirect strategies also combine with hedges (line 19) conveying that the addressee is unlikely to accept the burden (*Would you by any chance* . . . *?*, *Would it be at all possible* . . . *?*). These too serve to provide the addressee with a face-saving way to deny the request. We even see subtle effects of modality at work here: the irrealis, counterfactual forms *would* and *could* are more polite than their ability (dispositional) or future-oriented variants *can* and *will*; compare lines 12 and 13. This parallels the contrast between factuality markers (impolite; line 20) and hedging (polite; line 19).

Many of these features are correlated with each other, in keeping with the insight of Brown and Levinson (1987) that politeness markers are often combined to create a cumulative effect of increased politeness. Our corpora also highlight interactions that are unexpected (or at least unaccounted for) on existing theories of politeness. For example, sentence-medial *please* is polite (line 7), presumably because of its freedom to combine with other negative politeness strategies (*Could you please* . . .). In contrast, sentence-initial *please* is impolite (line 8), because it typically signals a more direct strategy (*Please do this*), which can make the politeness marker itself seem insincere. We see similar interactions between pronominal forms and syntactic structure: sentence-initial *you* is impolite (*You need to* . . .), whereas sentence-medial *you* is often part of the indirect strategies we discussed above (*Would/Could you* . . .).

4 Predicting politeness

We now show how our linguistic analysis can be used in a machine learning model for automatically classifying requests according to politeness. A classifier can help verify the predictive power, robustness, and domain-independent generality of the linguistic strategies of Section 3. Also, by providing automatic politeness judgments for large

Strategy	Politeness	In top quartile	Example
1. Gratitude	0.87^{***}	$78\%^{***}$	**I** really **appreciate** that you've done them.
2. Deference	0.78^{***}	$70\%^{***}$	**Nice work** so far on your rewrite.
3. Greeting	0.43^{***}	$45\%^{***}$	**Hey**, I just tried to ...
4. Positive lexicon	0.12^{***}	$32\%^{***}$	**Wow**! / This is a **great** way to deal...
5. Negative lexicon	-0.13^{***}	$22\%^{**}$	If you're going to **accuse** me ...
6. Apologizing	0.36^{***}	$53\%^{***}$	**Sorry** to bother you ...
7. Please	0.49^{***}	$57\%^{***}$	Could you **please** say more...
8. Please start	-0.30^{*}	22%	**Please** do not remove warnings ...
9. Indirect (btw)	0.63^{***}	$58\%^{**}$	**By the way**, where did you find ...
10. Direct question	-0.27^{***}	$15\%^{***}$	**What** is your native language?
11. Direct start	-0.43^{***}	$9\%^{***}$	**So** can you retrieve it or not?
12. Counterfactual modal	0.47^{***}	$52\%^{***}$	**Could/Would** you ...
13. Indicative modal	0.09	27%	**Can/Will** you ...
14. 1st person start	0.12^{***}	$29\%^{**}$	**I** have just put the article ...
15. 1st person pl.	0.08^{*}	27%	Could **we** find a less complex name ...
16. 1st person	0.08^{***}	$28\%^{***}$	It is **my** view that ...
17. 2nd person	0.05^{***}	$30\%^{***}$	But what's the good source **you** have in mind?
18. 2nd person start	-0.30^{***}	$17\%^{**}$	**You**'ve reverted yourself ...
19. Hedges	0.14^{***}	28%	I **suggest** we start with ...
20. Factuality	-0.38^{***}	$13\%^{***}$	**In fact** you did link, ...

Table 3: Positive (1-5) and negative (6–20) politeness strategies and their relation to human perception of politeness. For each strategy we show the average (human annotated) **politeness** scores for the requests exhibiting that strategy (compare with 0 for a random sample of requests; a positive number indicates the strategy is perceived as being polite), as well as the percentage of requests exhibiting the respective strategy that fall in the **top quartile** of politeness scores (compare with 25% for a random sample of requests). Throughout the paper: for politeness scores, statistical significance is calculated by comparing the set of requests exhibiting the strategy with the rest using a Mann-Whitney-Wilcoxon U test; for top quartile membership a binomial test is used.

amounts of new data on a scale unfeasible for human annotation, it can also enable a detailed analysis of the relation between politeness and social factors (Section 5).

Task setup To evaluate the robustness and domain-independence of the analysis from Section 3, we run our prediction experiments on two very different domains. We treat Wikipedia as a "development domain" since we used it for developing and identifying features and for training our models. Stack Exchange is our "test domain" since it was not used for identifying features. We take the model (features and weights) trained on Wikipedia and use them to classify requests from Stack Exchange.

We consider two classes of requests: **polite** and **impolite**, defined as the top and, respectively, bottom quartile of requests when sorted by their politeness score (based on the binary notion of politeness discussed in Section 2). The classes are therefore balanced, with each class consisting of 1,089 requests for the Wikipedia domain and 1,651 requests for the Stack Exchange domain.

We compare two classifiers — a bag of words classifier (BOW) and a linguistically informed classifier (Ling.) — and use human labelers as a reference point. The BOW classifier is an SVM using a unigram feature representation.[6] We consider this to be a strong baseline for this new

[6]Unigrams appearing less than 10 times are excluded.

classification task, especially considering the large amount of training data available. The linguistically informed classifier (Ling.) is an SVM using the linguistic features listed in Table 3 in addition to the unigram features. Finally, to obtain a reference point for the prediction task we also collect three new politeness annotations for each of the requests in our dataset using the same methodology described in Section 2. We then calculate human performance on the task (Human) as the percentage of requests for which the average score from the additional annotations matches the binary politeness class of the original annotations (e.g., a positive score corresponds to the polite class).

Classification results We evaluate the classifiers both in an in-domain setting, with a standard leave-one-out cross validation procedure, and in a cross-domain setting, where we train on one domain and test on the other (Table 4). For both our development and our test domains, and in both the in-domain and cross-domain settings, the linguistically informed features give 3-4% absolute improvement over the bag of words model. While the in-domain results are within 3% of human performance, the greater room for improvement in the cross-domain setting motivates further research on linguistic cues of politeness.

The experiments in this section confirm that our theory-inspired features are indeed effective in practice, and generalize well to new domains. In the next section we exploit this insight to automatically annotate a much larger set of requests (about 400,000) with politeness labels, enabling us to relate politeness to several social variables and outcomes. For new requests, we use class probability estimates obtained by fitting a logistic regression model to the output of the SVM (Witten and Frank, 2005) as *predicted politeness scores* (with values between 0 and 1; henceforth *politeness*, by abuse of language).

5 Relation to social factors

We now apply our framework to studying the relationship between politeness and social variables, focussing on social power dynamics. Encouraged by the close-to-human performance of our in-domain classifiers, we use them to assign politeness labels to our full dataset and then compare these labels to independent measures of power and status in our data. The results closely match those obtained with human-labeled data alone, thereby supporting the use of computational methods to pursue questions about social variables.

	In-domain		Cross-domain	
Train	Wiki	SE	Wiki	SE
Test	Wiki	SE	SE	Wiki
BOW	79.84%	74.47%	64.23%	72.17%
Ling.	83.79%	78.19%	67.53%	75.43%
Human	86.72%	80.89%	80.89%	86.72%

Table 4: Accuracies of our two classifiers for Wikipedia (Wiki) and Stack Exchange (SE), for in-domain and cross-domain settings. Human performance is included as a reference point. The random baseline performance is 50%.

5.1 Relation to social outcome

Earlier, we characterized politeness markings as currency used to pay respect. Such language is therefore costly in a social sense, and, relatedly, tends to incur costs in terms of communicative efficiency (Van Rooy, 2003). Are these costs worth paying? We now address this question by studying politeness in the context of the electoral system of the Wikipedia community of editors.

Among Wikipedia editors, status is a salient social variable (Anderson et al., 2012). Administrators (*admins*) are editors who have been granted certain rights, including the ability to block other editors and to protect or delete articles.[7] Admins have a higher status than common editors (*non-admins*), and this distinction seems to be widely acknowledged by the community (Burke and Kraut, 2008b; Leskovec et al., 2010; Danescu-Niculescu-Mizil et al., 2012). Aspiring editors become admins through public elections,[8] so we know when the status change from non-admin to admins occurred and can study users' language use in relation to that time.

To see whether politeness correlates with eventual high status, we compare, in Table 5, the politeness levels of requests made by users who will eventually succeed in becoming administrators (Eventual status: Admins) with requests made by users who are not admins (Non-admins).[9] We observe that admins-to-be are significantly more po-

[7] http://en.wikipedia.org/wiki/Wikipedia:Administrators

[8] http://en.wikipedia.org/wiki/Wikipedia:Requests_for_adminship

[9] We consider only requests made up to one month before the election, to avoid confusion with pre-election behavior.

Eventual status	Politeness	Top quart.
Admins	0.46^{**}	$30\%^{***}$
Non-admins	0.39^{***}	25%
Failed	0.37^{**}	22%

Table 5: Politeness and status. Editors who will eventually become admins are more polite than non-admins ($p<0.001$ according to a Mann-Whitney-Wilcoxon U test) and than editors who will eventually fail to become admins ($p<0.001$). Out of their requests, 30% are rated in the top politeness quartile (significantly more than the 25% of a random sample; $p<0.001$ according to a binomial test). This analysis was conducted on 31k requests (1.4k for Admins, 28.9k for Non-admins, 652 for Failed).

lite than non-admins. One might wonder whether this merely reflects the fact that not all users aspire to become admins, and those that do are more polite. To address this, we also consider users who ran for adminship but did not earn community approval (Eventual status: Failed). These users are also significantly less polite than their successful counterparts, indicating that politeness indeed correlates with a positive social outcome here.

5.2 Politeness and power

We expect a rise in status to correlate with a decline in politeness (as predicted by politeness theory, and discussed in Section 1). The previous section does not test this hypothesis, since all editors compared in Table 5 had the same (non-admin) status when writing the requests. However, our data does provide three ways of testing this hypothesis.

First, after the adminship elections, successful editors get a boost in power by receiving admin privileges. Figure 3 shows that this boost is mirrored by a significant decrease in politeness (blue, diamond markers). Losing an election has the opposite effect on politeness (red, circle markers), perhaps as a consequence of reinforced low status.

Second, Stack Exchange allows us to test more situational power effects.[10] On the site, users request, from the community, information they are lacking. This informational asymmetry between the question-asker and his audience puts him at

[10]We restrict all experiments in this section to the largest subcommunity of Stack Exchange, namely Stack Overflow.

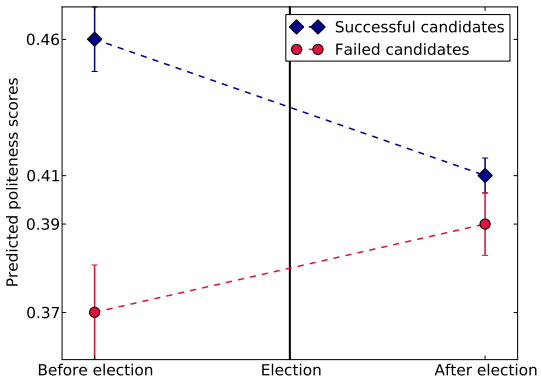

Figure 3: Successful and failed candidates before and after elections. Editors that will eventually succeed (diamond marker) are significantly more polite than those that will fail (circle markers). Following the elections, successful editors become less polite while unsuccessful editors become more polite.

a social disadvantage. We therefore expect the question-asker to be more polite than the people who respond. Table 6 shows that this expectation is born out: comments posted to a thread by the original question-asker are more polite than those posted by other users.

Role	Politeness	Top quart.
Question-asker	0.65^{***}	$32\%^{***}$
Answer-givers	0.52^{***}	$20\%^{***}$

Table 6: Politeness and dependence. Requests made in comments posted by the question-asker are significantly more polite than the other requests. Analysis conducted on 181k requests (106k for question-askers, 75k for answer-givers).

Third, Stack Exchange allows us to examine power in the form of authority, through the community's reputation system. Again, we see a negative correlation between politeness and power, even after controlling for the role of the user making the requests (i.e., Question-asker or Answer-giver). Table 7 summarizes the results.[11]

Human validation The above analyses are based on predicted politeness from our classifier. This allows us to use the entire request data cor-

[11]Since our data does not contain time stamps for reputation scores, we only consider requests that were issued in the six months prior to the available snapshot.

Reputation level	Politeness	Top quart.
Low reputation	0.68^{***}	$27\%^{***}$
Middle reputation	0.66^{***}	25%
High reputation	0.64^{***}	$23\%^{***}$

Table 7: Politeness and Stack Exchange reputation (texts by question-askers only). High-reputation users are less polite. Analysis conducted on 25k requests (4.5k low, 12.5k middle, 8.4k high).

PL name	Politeness	Top quartile
Python	0.47^{***}	23%
Perl	0.49	24%
PHP	0.51	24%
Javascript	0.53^{**}	$26\%^{**}$
Ruby	0.59^{***}	$28\%^{*}$

Table 8: Politeness of requests from different language communities on Stack Exchange.

pus to test our hypotheses and to apply precise controls to our experiments (such as restricting our analysis to question-askers in the reputation experiment). In order to validate this methodology, we turned again to human annotation: we collected additional politeness annotation for the types of requests involved in the newly designed experiments. When we re-ran our experiments on human-labeled data alone we obtained the same qualitative results, with statistical significance always lower than 0.01.[12]

Prediction-based interactions The human validation of classifier-based results suggests that our prediction framework can be used to explore differences in politeness levels across factors of interest, such as communities, geographical regions and gender, even where gathering sufficient human-annotated data is infeasible. We mention just a few such preliminary results here: (i) Wikipedians from the U.S. Midwest are most polite (when compared to other census-defined regions), (ii) female Wikipedians are generally more polite (consistent with prior studies in which women are more polite in a variety of domains; (Herring, 1994)), and (iii) programming language communities on Stack Exchange vary significantly by politeness (Table 8; full disclosure: our analyses were conducted in Python).

6 Related work

Politeness has been a central concern of modern pragmatic theory since its inception (Grice, 1975; Lakoff, 1973; Lakoff, 1977; Leech, 1983; Brown and Levinson, 1978), because it is a source of pragmatic enrichment, social meaning, and cultural variation (Harada, 1976; Matsumoto, 1988; Ide, 1989; Blum-Kulka and Kasper, 1990; Blum-Kulka, 2003; Watts, 2003; Byon, 2006). The starting point for most research is the theory of Brown and Levinson (1987). Aspects of this theory have been explored from game-theoretic perspectives (Van Rooy, 2003) and implemented in language generation systems for interactive narratives (Walker et al., 1997), cooking instructions, (Gupta et al., 2007), translation (Faruqui and Pado, 2012), spoken dialog (Wang et al., 2012), and subjectivity analysis (Abdul-Mageed and Diab, 2012), among others.

In recent years, politeness has been studied in online settings. Researchers have identified variation in politeness marking across different contexts and media types (Herring, 1994; Brennan and Ohaeri, 1999; Duthler, 2006) and between different social groups (Burke and Kraut, 2008a). The present paper pursues similar goals using orders of magnitude more data, which facilitates a fuller survey of different politeness strategies.

Politeness marking is one aspect of the broader issue of how language relates to power and status, which has been studied in the context of workplace discourse (Bramsen et al., ; Diehl et al., 2007; Peterson et al., 2011; Prabhakaran et al., 2012; Gilbert, 2012; McCallum et al., 2007) and social networking (Scholand et al., 2010). However, this research focusses on domain-specific textual cues, whereas the present work seeks to leverage domain-independent politeness cues, building on the literature on how politeness affects worksplace social dynamics and power structures (Gyasi Obeng, 1997; Chilton, 1990; Andersson and Pearson, 1999; Rogers and Lee-Wong, 2003; Holmes and Stubbe, 2005). Burke and Kraut (2008b) study the question of how and why specific individuals rise to administrative positions on Wikipedia, and Danescu-Niculescu-Mizil et al. (2012) show that power differences on Wikipedia

[12] However, due to the limited size of the human-labeled data, we could not control for the role of the user in the Stack Exchange reputation experiment.

are revealed through aspects of linguistic accommodation. The present paper complements this work by revealing the role of politeness in social outcomes and power relations.

7 Conclusion

We construct and release a large collection of politeness-annotated requests and use it to evaluate key aspects of politeness theory. We build a politeness classifier that achieves near-human performance and use it to explore the relation between politeness and social factors such as power, status, gender, and community membership. We hope the publicly available collection of annotated requests enables further study of politeness and its relation to social factors, as this paper has only begun to explore this area.

Acknowledgments

We thank Jean Wu for running the AMT annotation task, and all the participating turkers. We thank Diana Minculescu and the anonymous reviewers for their helpful comments. This work was supported in part by NSF IIS-1016909, CNS-1010921, IIS-1149837, IIS-1159679, ARO MURI, DARPA SMISC, Okawa Foundation, Docomo, Boeing, Allyes, Volkswagen, Intel, Alfred P. Sloan Fellowship, the Microsoft Faculty Fellowship, the Gordon and Dailey Pattee Faculty Fellowship, and the Center for Advanced Study in the Behavioral Sciences at Stanford.

References

Muhammad Abdul-Mageed and Mona Diab. 2012. AWATIF: A multi-genre corpus for Modern Standard Arabic subjectivity and sentiment analysis. In *Proceedings of LREC*, pages 3907–3914.

Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2012. Effects of user similarity in social media. In *Proceedings of WSDM*, pages 703–712.

Lynne M. Andersson and Christine M. Pearson. 1999. Tit for tat? the spiraling effect of incivility in the workplace. *The Academy of Management Review*, 24(3):452–471.

Shoshana Blum-Kulka and Gabriele Kasper. 1990. Special issue on politeness. *Journal of Pragmatics*, 144(2).

Shoshana Blum-Kulka. 2003. Indirectness and politeness in requests: Same or different? *Journal of Pragmatics*, 11(2):131–146.

Philip Bramsen, Martha Escobar-Molana, Ami Patel, and Rafael Alonso. Extracting social power relationships from natural language. In *Proceedings of ACL*, pages 773–782.

Susan E Brennan and Justina O Ohaeri. 1999. Why do electronic conversations seem less polite? the costs and benefits of hedging. *SIGSOFT Softw. Eng. Notes*, 24(2):227–235.

Penelope Brown and Stephen C. Levinson. 1978. Universals in language use: Politeness phenomena. In Esther N. Goody, editor, *Questions and Politeness: Strategies in Social Interaction*, pages 56–311, Cambridge. Cambridge University Press.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: some universals in language usage*. Cambridge University Press.

Moira Burke and Robert Kraut. 2008a. Mind your Ps and Qs: the impact of politeness and rudeness in online communities. In *Proceedings of CSCW*, pages 281–284.

Moira Burke and Robert Kraut. 2008b. Taking up the mop: identifying future wikipedia administrators. In *CHI '08 extended abstracts on Human factors in computing systems*, pages 3441–3446.

Andrew Sangpil Byon. 2006. The role of linguistic indirectness and honorifics in achieving linguistic politeness in Korean requests. *Journal of Politeness Research*, 2(2):247–276.

Paul Chilton. 1990. Politeness, politics, and diplomacy. *Discourse and Society*, 1(2):201–224.

Herbert H. Clark and Dale H. Schunk. 1980. Polite responses to polite requests. *Cognition*, 8(1):111–143.

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.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, pages 449–454.

Christopher P. Diehl, Galileo Namata, and Lise Getoor. 2007. Relationship identification for social network discovery. In *Proceedings of the AAAI Workshop on Enhanced Messaging*, pages 546–552.

Kirk W Duthler. 2006. The Politeness of Requests Made Via Email and Voicemail: Support for the Hyperpersonal Model. *Journal of Computer-Mediated Communication*, 11(2):500–521.

Manaal Faruqui and Sebastian Pado. 2012. Towards a model of formal and informal address in english. In *Proceedings of EACL*, pages 623–633.

Elen P. Francik and Herbert H. Clark. 1985. How to make requests that overcome obstacles to compliance. *Journal of Memory and Language*, 24:560–568.

Eric Gilbert. 2012. Phrases that signal workplace hierarchy. In *Proceedings of CSCW*, pages 1037–1046.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics*, volume 3: Speech Acts, pages 43–58. Academic Press, New York.

S Gupta, M Walker, and D Romano. 2007. How rude are you?: Evaluating politeness and affect in interaction. *Affective Computing and Intelligent Interaction*, pages 203–217.

Samuel Gyasi Obeng. 1997. Language and politics: Indirectness in political discourse. *Discourse and Society*, 8(1):49–83.

S. I. Harada. 1976. Honorifics. In Masayoshi Shibatani, editor, *Syntax and Semantics*, volume 5: Japanese Generative Grammar, pages 499–561. Academic Press, New York.

Susan Herring. 1994. Politeness in computer culture: Why women thank and men flame. In *Cultural performances: Proceedings of the third Berkeley women and language conference*, volume 278, page 94.

Janet Holmes and Maria Stubbe. 2005. *Power and Politeness in the Workplace: A Sociolinguistic Analysis of Talk at Work*. Longman, London.

Ken Hyland. 2005. *Metadiscourse: Exploring Interaction in Writing*. Continuum, London and New York.

Sachiko Ide. 1989. Formal forms and discernment: Two neglected aspects of universals of linguistic politeness. *Multilingua*, 8(2–3):223–248.

David Kaplan. 1999. What is meaning? Explorations in the theory of *Meaning as Use*. Brief version — draft 1. Ms., UCLA.

Robin Lakoff. 1973. The logic of politeness; or, miding your P's and Q's. In *Proceedings of the 9th Meeting of the Chicago Linguistic Society*, pages 292–305.

Robin Lakoff. 1977. What you can do with words: Politeness, pragmatics and performatives. In *Proceedings of the Texas Conference on Performatives, Presuppositions and Implicatures*, pages 79–106.

Geoffrey N. Leech. 1983. *Principles of Pragmatics*. Longman, London and New York.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Governance in Social Media: A case study of the Wikipedia promotion process. In *Proceedings of ICWSM*, pages 98–105.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion Observer: analyzing and comparing opinions on the Web. In *Proceedings of WWW*, pages 342–351.

Yoshiko Matsumoto. 1988. Reexamination of the universality of face: Politeness phenomena in Japanese. *Journal of Pragmatics*, 12(4):403–426.

Andrew McCallum, Xuerui Wang, and Andr'es Corrada-Emmanuel. 2007. Topic and role discovery in social networks with experiments on Enron and academic email. *Journal of Artificial Intelligence Research*, 30(1):249–272.

Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130.

Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the enron corpus. In *Proceedings of the ACL Workshop on Language in Social Media*, pages 86–95.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2012. Predicting Overt Display of Power in Written Dialogs. In *Proceedings of NAACL-HLT*, pages 518–522.

Priscilla S. Rogers and Song Mei Lee-Wong. 2003. Reconceptualizing politeness to accommodate dynamic tensions in subordinate-to-superior reporting. *Journal of Business and Technical Communication*, 17(4):379–412.

Andrew J. Scholand, Yla R. Tausczik, and James W. Pennebaker. 2010. Social language network analysis. In *Proceedings of CSCW*, pages 23–26.

Robert Van Rooy. 2003. Being polite is a handicap: Towards a game theoretical analysis of polite linguistic behavior. In *Proceedings of TARK*, pages 45–58.

Marilyn A Walker, Janet E Cahn, and Stephen J Whittaker. 1997. Improvising linguistic style: social and affective bases for agent personality. In *Proceedings of AGENTS*, pages 96–105.

William Yang Wang, Samantha Finkelstein, Amy Ogan, Alan W. Black, and Justine Cassell. 2012. "love ya, jerkface": Using sparse log-linear models to build positive and impolite relationships with teens. In *Proceedings of SIGDIAL*, pages 20–29.

Richard J. Watts. 2003. *Politeness*. Cambridge University Press, Cambridge.

Ian H Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.