Gender and Power: How Gender and Gender Environment Affect Manifestations of Power

Vinodkumar Prabhakaran
Dept. of Computer Science
Columbia University
New York, NY, USA
vinod@cs.columbia.edu

Emily E. Reid
Dept. of Computer Science
Columbia University
New York, NY, USA
eer2137@columbia.edu

Owen Rambow
CCLS
Columbia University
New York, NY, USA
rambow@ccls.columbia.edu

Abstract

We investigate the interaction of power, gender, and language use in the Enron email corpus. We present a freely available extension to the Enron corpus, with the gender of senders of 87% messages reliably identified. Using this data, we test two specific hypotheses drawn from the sociolinguistic literature pertaining to gender and power: women managers use face-saving communicative strategies, and women use language more explicitly than men to create and maintain social relations. We introduce the notion of "gender environment" to the computational study of written conversations; we interpret this notion as the gender makeup of an email thread, and show that some manifestations of power differ significantly between gender environments. Finally, we show the utility of gender information in the problem of automatically predicting the direction of power between pairs of participants in email interactions.

1 Introduction

It has long been observed that men and women communicate differently in different contexts. This phenomenon has been studied by sociolinguists, who typically rely on case studies or surveys. The availability of large corpora of naturally occurring social interactions has given us the opportunity to study language use at a broader level than before. In this paper, we use the Enron Corpus of work-related emails to examine written communication in a corporate setting. We investigate three factors that affect choices in communication: the writer's gender, the gender of his or her fellow discourse participants (what we call the "gender environment"), and the relations of organizational power he or she has to the discourse participants. We concentrate on modeling the writer's choices related to discourse structure, rather than lexical choice. Specifically, our goal is to show that gender, gender environment, and power all affect individuals' choices in complex ways, resulting in patterns in the discourse that reveal the underlying factors.

This paper makes three major contributions. First, we introduce an extension to the well-known Enron corpus of emails: we semi-automatically identify the sender's gender of 87% of email messages in the corpus. This extension will be made publicly available. Second, we use this enriched version of the corpus to investigate the interaction of hierarchical power and gender. We formalize the notion of "gender environment", which reflects the gender makeup of the discourse participants of a particular conversation. We study how gender, power, and gender environment influence discourse participants' choices in dialog. We investigate two specific hypotheses from the sociolinguistic literature, relating to face-saving use of language, and to the use of language to strengthen social relations. This contribution does not exhaust the possibilities of our corpus, but it shows how social science can benefit from advanced natural language processing techniques in analyzing corpora, allowing social scientists to tackle corpora such as the Enron corpus which cannot be examined in its entirety by hand. Third, we show that the gender information in the enriched corpus can be useful for computational tasks, specifically for training a system that predicts the direction of hierarchical power between participants in an interaction. Our use of the gender-based features boosts the accuracy of predicting the direction of power between pairs of email interactants from 68.9% to 70.2% on an unseen test set.

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1965–1976,
October 25-29, 2014, Doha, Qatar.

The paper is structured as follows. We review related work in Section 2. We present the Gender Identified Enron Corpus (our first contribution) in Section 3. Section 4 defines the problem of predicting power and the various dimensions of interaction we analyze. We turn to our second contribution, the analysis of the data, in Sections 5 and 6. Section 7 describes our third contribution, the machine learning experiments using gender-related features in the prediction of hierarchical power. We then conclude and discuss future work.

2 Related Work

There is much sociolinguistic background related to gender and language use, some of it specifically related to language use in the work environment (Kendall and Tannen, 1997; Holmes and Stubbe, 2003; Kendall, 2003; Herring, 2008). We do not provide a full discussion of this work for lack of space, but single out one paper which has particularly influenced our work. Holmes and Stubbe (2003) provide two case studies that do not look at the differences between male and female managers' communication, but at the difference between female managers' communication in more heavily female vs. more heavily male environments. They find that, while female managers tend to break many stereotypes of "feminine" communication, they have different strategies in connecting with employees and exhibiting power in the two gender environments. This work has inspired us to look at this phenomenon by including "Gender Environment" in our study. By finding the ratios of males to females on a thread, we can look at whether indicators change within a more heavily male or female thread. This notion of gender environment is supported by an idea in recent Twitter-based sociolinguistic research on gender identity and lexical variation (Bamman et al., 2014). One of the many insights from their work is that gendered linguistic behavior is oriented by a number of factors, one of which includes the speaker's audience. Their work looks at Twitter users whose linguistic style fails to identify their gender in classification experiments, and finds that the linguistic gender norms can be influenced by the style of their interlocutors.

Within the NLP community, there has been substantial research exploring language use and power. A large number of these studies are performed in the domain of organizational email where the notion of power is well defined in terms of organizational hierarchy. It is also aided by the availability of the moderately large Enron email corpus which captures email interactions in an organizational setting. Earlier approaches used simple lexical features alone (e.g. (Bramsen et al., 2011; Gilbert, 2012)) as a means to predict power. Later studies have used more complex linguistic and structural features, such as formality (Peterson et al., 2011), dialog acts (Prabhakaran and Rambow, 2013), and thread structure (Prabhakaran and Rambow, 2014). Our work is also on the Enron email corpus, and our baseline features are derived from some of this prior work. Researchers have also studied power and influence in other genres of interactions, such as online forums (Danescu-Niculescu-Mizil et al., 2012; Biran et al., 2012), multi-party chats (Strzalkowski et al., 2012) and off-line interactions such as presidential debates (Nguyen et al., 2013; Prabhakaran et al., 2013; Prabhakaran et al., 2014).

There is also some work within the NLP field on analyzing language use in relation to gender. Mohammad and Yang (2011) analyzed the way gender affects the expression of sentiments in text, while we are interested in how gender relates to manifestations of organizational power. For their study, they assigned gender for the core employees in the Enron email corpus based on whether the first name of the person was easily gender identifiable or not. If the person had an unfamiliar name or a name that could be of either gender, they marked his/her gender as *unknown* and excluded them from their study.[1] For example, the gender of the employee Kay Mann was marked as *unknown* in their gender assignment. However, in our work, we manually research and determine the gender of every core employee.

Researchers have also attempted to automatically predict the gender of email senders using supervised learning techniques based on linguistic features (Corney et al., 2002; Cheng et al., 2011; Deitrick et al., 2012), a task we do not address in this paper. These studies use datasets that are relatively smaller in size. Corney et al. (2002) use around 4K emails from 325 gender identified authors. Cheng et al. (2011) use around 9K emails from 108 gender identified authors. Deitrick et al. (2012) use around 18K emails from 144 gender

[1] http://www.saifmohammad.com/WebDocs/dir-email-gender.txt

identified authors. The dataset we offer is much larger in size, with around 97K emails whose authors are gender identified. We believe that our resource will aid further research in this area.

3 Gender Identified Enron Corpus

3.1 Enron Corpus

In our work, we use the version of Enron email corpus released by Yeh and Harnly (2006). The corpus contains emails from the mailboxes of 145 core employees who held top managerial positions within Enron at the time of bankruptcy. Yeh and Harnly (2006) preprocessed the corpus to combine multiple email addresses belonging to the same entity and identify each entity in the corpus with a unique identifier. The corpus contains a total of 111,933 messages. This version of the corpus has been enriched later by Agarwal et al. (2012) with gold organizational power relations, manually determined using information from Enron organizational charts. It includes relations of 1,518 employees and captures dominance relations between 13,724 pairs of them. This information enables us to study the manifestations of power in these interactions, in relation to gender.

In this version of the corpus, the thread structure of email messages is reconstructed, with the missing messages restored from other emails in which they were quoted. This allows us to go beyond isolated messages and study the dialog structure within email threads. There were 34,156 unique discourse participants across all the email threads present in the corpus. Manually determining the gender of all the discourse participants in the corpus is not feasible. Hence, we adopt a two-step approach through which we reliably identify the gender of a large majority of entities in the email threads within the corpus. We manually determine the gender of the 145 core employees who have a bigger representation in the corpus, and we systemically determine the gender of the rest of the discourse participants using the Social Security Administration's baby names database. We adopt a conservative approach so that we assign a gender only when the name of the participant meets a very low ambiguity threshold.

3.2 Manual Gender Assignment

We researched each of the 145 core employees using web search and found public records about them or articles referring to them. In order to make sure that the results are about the same person we want, we added the word 'enron' to the search queries. Within the public records returned for each core employee, we looked for instances in which they were being referred to either using a gender revealing pronoun (*he*/*him*/*his* vs. *she*/*her*) or using a gender revealing addressing form (*Mr.* vs. *Mrs.*/*Ms.*/*Miss*). Since these employees held top managerial positions within Enron at the time of bankruptcy, it was fairly easy to find public records or articles referring to them. For example, the page we found for Kay Mann clearly identifies her gender.[2] We were able to correctly determine the gender of each of the 145 core employees in this manner. A benefit of manually determining the gender of these core employees is that it ensures a high coverage of 100% confident gender assignments in the corpus.

3.3 Automatic Gender Assignment

As mentioned in Section 3.1, our corpus contains a large number of discourse participants in addition to the 145 core employees for which we manually identified the gender. To attempt to find the gender of these other discourse participants, we first determine their first names and then find how ambiguous the names are by querying the Social Security Administration's (SSA) baby names dataset. We first describe how we calculate an ambiguity score for a name using the SSA dataset and then describe how we use it to determine the gender of discourse participants in our corpus.

3.3.1 SSA Names and Gender Dataset

The US Social Security Administration maintains a dataset of baby names, gender, and name count for each year starting with the 1880s, for names with at least five counts.[3] We used this dataset in order to determine the gender ambiguity of a name. The Enron data set contains emails from 1998 to 2001. We estimate the common age range for a large, corporate firm like Enron at 24-67,[4] so we used the SSA data from 1931-1977 to calculate ambiguity scores for our purposes.

For each name n in the database, let $mp(n)$ and $fp(n)$ denote the percentages of males and females with the name n. Then, we calculate the ambiguity score $AS(n)$ as $100 - |mp(n) - fp(n)|$.

[2] http://www.prnewswire.com/news-releases/kay-mann-joins-noble-as-general-counsel-57073687.html

[3] http://www.ssa.gov/oact/babynames/limits.html

[4] http://www.bls.gov/cps/demographics.htm

The value of $AS(n)$ varies between 0 and 100. A name that is 'perfectly unambiguous' would have an ambiguity score of 0, while a 'perfectly ambiguous' name (i.e., 50%/50% split between genders) would have an ambiguity score of 100. We assign the likely gender of the name to be the one with the higher percentage, if the ambiguity score is below a threshold AS_T.

$$G(n) = \begin{cases} M, & \text{if } AS(n) \leq AS_T \text{ and } mp(n) > fp(n) \\ F, & \text{if } AS(n) \leq AS_T \text{ and } mp(n) \leq fp(n) \\ I, & \text{if } AS(n) > AS_T \end{cases}$$

Around 88% of the names in the SSA dataset have $AS(n) = 0$. We choose a very conservative threshold of $AS_T = 10$ for our gender assignments, which assigns gender to around 93% names in the SSA dataset.[5]

3.3.2 Identifying the First Name

Each discourse participant in our corpus has at least one email address and zero or more names associated with it. The name field is automatically assembled by Yeh and Harnly (2006), where they captured the different names from email headers, which are populated from individual email clients and do not follow a standard format. Not all discourse participants are human; some may refer to organizational groups (e.g., HR Department) or anonymous corporate email accounts (e.g., a webmaster account, do-not-reply address etc.). The name field may sometimes be empty, contain multiple names, contain an email address, or show other irregularities. Hence, it is nontrivial to determine the first name of our discourse participants. We used the heuristics below to extract the most likely first name for each discourse participant.

- If the name field contains two words, pick the second or first word, depending on whether a comma separates them or not.
- If the name field contains three words and a comma, choose the second and third words (a likely first and middle name, respectively). If the name field contains three words but no comma, choose the first and second words (again, a likely first and middle name).
- If the name field contains an email address, pick the portion from the beginning of the string to a '.','_' or '-'; if the email address is in camel case, take portion from the beginning of the string to the first upper case letter.
- If the name field is empty, apply the above rule to the email address field to pick a name.

The above heuristics create a list of candidate names for each discourse participant which we then query for an ambiguity score (Section 3.3.1) and the likely gender. We find the candidate name with the lowest ambiguity score that passes the threshold and assign the associated gender to the discourse participant. If none of the candidate names for a discourse participant passes the threshold, we assign the gender to be 'I' (Indeterminate). We also assign the gender to be 'I', if none of the candidate names is present in the SSA dataset. This will occur if the name is a first name that is not in the database (an unusual or international name; e.g., *Vladi*), or if no true first name was found (e.g., the name field was empty and the email address was only a pseudonym). This will also include most of the cases where the discourse participant is not a human.

3.3.3 Coverage and Accuracy

We evaluated the coverage and accuracy of our gender assignment system on the manually assigned gender data of the 145 core people. We obtained a coverage of 90.3%, i.e., for 14 of the 145 core people, the ambiguity score was higher than the threshold. Of the 131 people the system assigned a gender to, we obtained an accuracy of 89.3% in correctly identifying the gender. We investigated the errors and found that all errors were caused due to incorrectly identifying the first name. These errors arise because the name fields are automatically populated and sometimes the core discourse participants' name fields include their secretaries. While this is common for people in higher managerial positions, we expect this not to happen in the middle management and below, to which most of the automatically gender-assigned discourse participants belong.

3.4 Corpus Statistics and Divisions

We apply the gender assignment system described above to all discourse participants of all email threads in the entire Enron corpus described in Section 3.1. Table 1 shows the coverage of gender assignment in our corpus at different levels: unique discourse participants, messages and threads. In Table 2, we show the male/female percentage split of all unique discourse participants, as well as the split at the level of messages (i.e., messages sent by males vs. females).

[5] In the corpus that will be released, we retain the $AS(n)$ of each name, so that the users of this resource can decide the threshold that suit their needs.

	Count (%)
Total unique discourse participants	34,156
- gender identified	23,009 (67.3%)
Total messages	111,933
- senders gender identified	97,255 (86.9%)
Total threads	36,615
- all senders gender identified	26,015 (71.1%)
- all participants gender identified	18,030 (49.2%)

Table 1: Coverage of Gender Identification at various level: unique discourse participants, messages and threads

	Male	Female
Unique Discourse Participants	66.1%	33.9%
Message Senders	58.2%	41.8%

Table 2: Male/Female split across a) all unique participants who were gender identified, b) all messages whose senders were gender identified

We divide the entire corpus into Train, Dev and Test sets at the thread level, through random sampling, with a distribution of 50%, 25% and 25% each. The number of threads and messages in each subdivision is shown in Table 3.

	Total	**Train**	**Dev**	**Test**
Threads	36,615	18,498	8,973	9,144
Messages	111,933	56,447	27,565	27,921

Table 3: Train/Test/Dev breakup of the entire corpus

We also create a sub-corpus of the threads called *All Participants Gender Identified* (APGI), containing the 18,030 threads for which the gender assignment system succeeded in assigning the genders of all participants, including senders and all recipients (To and CC). For the analysis and experiments presented in the rest of this paper, we use 17,788 threads from this APGI subset, excluding the remaining 242 threads that were used for previous manual annotation efforts.

4 Manifestations of Power

We use the gender information of the participants to investigate how the gender of the sender and recipients affect the manifestations of hierarchical power in interactions. In order to do this, we use the interaction analysis framework from our prior work (Prabhakaran and Rambow, 2014). In this section, we give a brief overview of the problem formulation and the structural features we used.

4.1 Hierarchically Related Interacting Pairs

Let t denote an email thread and M_t denote the set of all messages in t. Also, let P_t be the set of all participants in t, i.e., the union of senders and recipients (*To* and *CC*) of all messages in M_t. We are interested in analyzing the power relations between pairs of participants who interact within a given email thread. Not every pair of participants $(p_1, p_2) \in P_t \times P_t$ interact with one another within t. Let $IM_t(p_1, p_2)$ denote the set of *Interaction Messages* — non-empty messages in t in which either p_1 is the sender and p_2 is one of the recipients or vice versa. We call the set of (p_1, p_2) such that $|IM_t(p_1, p_2)| > 0$ the *interacting participant pairs* of t (IPP_t). For every $(p_1, p_2) \in IPP_t$, we query the set of dominance relations in the gold hierarchy and assign their hierarchical power relation ($HP(p_1, p_2)$) to be *superior* if p_1 dominates p_2, and *subordinate* if p_2 dominates p_1. We exclude pairs that do not exist in the gold hierarchy from our analysis and call the remaining set *related interacting participant pairs* ($RIPP_t$). Table 4 shows the total number of pairs in IPP_t and $RIPP_t$ from all the threads in the APGI subset of our corpus and across Train, Dev and Test sets.

Description	**Total**	**Train**	**Dev**	**Test**
# of threads	17,788	8,911	4,328	4,549
$\sum_t \lvert IPP_t \rvert$	74,523	36,528	18,540	19,455
$\sum_t \lvert RIPP_t \rvert$	4,649	2,260	1,080	1,309

Table 4: Data Statistics
Row 1 presents the total number of threads in different subsets of the corpus. Row 2 and 3 present the number of interacting participant pairs (IPP) and related interacting participant pairs ($RIPP$) in those subsets.

4.2 Structural Features

Now, we describe various features that capture the structure of interaction between the pairs of participants in a thread. Each feature f is extracted with respect to a person p over a reference set of messages M (denoted f^p_M). For a pair (p_1, p_2), we extract 4 versions of each feature f: $f^{p_1}_{IM_t(p_1,p_2)}$, $f^{p_2}_{IM_t(p_1,p_2)}$, $f^{p_1}_{M_t}$ and $f^{p_2}_{M_t}$. The first two capture behavior of each person of the pair in interactions between themselves, while the third and fourth capture their overall behavior in the entire thread. We group our features into three categories — THR^{STR}, THR^{META} and DIA. THR^{STR} captures the thread structure in terms of verbosity and

positional features of messages (e.g., how many emails did a person send). THR$^{\text{META}}$ contain email header meta-data based features that capture the thread structure (e.g., how many recipients were there). Both sets of features do not perform any NLP analysis on the the content of the emails. DIA captures the pragmatics of the dialog and requires a deeper analysis of the email content (e.g., did they issue any requests).

THR$^{\text{STR}}$: This feature set includes two kinds of features — positional and verbosity. The positional features are a boolean feature to denote whether p sent the first message (Initiate), and the relative positions of p's first and last messages (FirstMsgPos and LastMsgPos) in M. The verbosity features are p's message count (MsgCount), message ratio (MsgRatio), token count (TokenCount), token ratio (TokenRato) and tokens per message (TokenPerMsg), all calculated over M.

THR$^{\text{META}}$: This feature set includes the average number of recipients (AvgRecipients) and *To* recipients (AvgToRecipients) in emails sent by p, the percentage of emails p received in which he/she was in the *To* list (InToList%), boolean features denoting whether p added or removed people when responding to a message (AddPerson and RemovePerson), average number of replies received per message sent by p (ReplyRate) and average number of replies received from the other person of the pair to messages where he/she was a *To* recipient (ReplyRateWithinPair). ReplyRateWithinPair applies only to $IM_t(p_1, p_2)$.

DIA: We use dialog acts (DA) and overt displays of power (ODP) tags to model the structure of interactions within the message content. We obtain DA and ODP tags using automatic taggers trained on manual annotations. The DA tagger (Omuya et al., 2013) obtained an accuracy of 92%. The ODP tagger (Prabhakaran et al., 2012) obtained an accuracy of 96% and F-measure of 54%. The DA tagger labels each sentence to be one of the 4 dialog acts: Request Action, Request Information, Inform, and Conventional. The ODP Tagger identifies sentences (mostly requests) that express additional constraints on their addressee, beyond those introduced by the dialog act. For example, the sentence "Please come to my office right now" is considered as an ODP, while "It would be great if you could come to my office now" is not, even though both issue the same request. For more details on ODP, we refer the reader to (Prabhakaran et al., 2012). We use 5 features: ReqAction, ReqInform, Inform, Conventional, and ODPCount to capture the number of sentences in messages sent by p that have each of these labels. We also use a feature to capture the number of p's messages with a request that did not get a reply, i.e., dangling request percentage (DanglingReq%), over all messages sent by p.

Feature Name	**Mean**$(f^X_{IM_t})\|X =$			
	F_{sub}	F_{sup}	M_{sub}	M_{sup}
THR$^{\text{META}}$				
AvgRecipients***	4.76	5.74	5.58	4.98
AvgToRecipients***	3.63	4.73	3.84	3.80
InToList%$^{\cdot}$	0.83	0.86	0.84	0.83
ReplyRate***	0.72	0.86	0.70	0.61
AddPerson	0.58	0.66	0.59	0.68
RemovePerson	0.55	0.60	0.54	0.65
THR$^{\text{STR}}$				
Initiate	0.38	0.24	0.39	0.30
FirstMsgPos*	0.18	0.25	0.19	0.22
LastMsgPos**	0.34	0.33	0.34	0.39
MsgCount***	0.92	0.61	0.93	0.91
MsgRatio***	0.33	0.23	0.33	0.32
TokenCount	76.5	41.0	102.0	54.3
TokenRatio	0.38	0.23	0.40	0.27
TokenPerMsg***	90.2	67.9	118.2	53.2
DIA^{PR}				
Conventional	0.55	0.43	0.64	0.56
Inform	3.50	1.96	4.51	2.53
ReqAction**	0.07	0.06	0.05	0.10
ReqInform	0.29	0.21	0.20	0.16
DanglingReq%	0.06	0.12	0.07	0.18
ODPCount***	0.10	0.07	0.09	0.13

Table 5: ANOVA results and group means for Hierarchical Power and Gender
F_{sub}: Female subordinates; F_{sup}: Female superiors; M_{sub}: Male subordinates; M_{sup}: Male superiors; * ($p < .05$); ** ($p < .01$); *** ($p < .001$)

5 Gender and Power

In this subsection, we analyze the impact of gender on the expression of power in email. We perform an ANOVA test on all features described in Section 4.2 keeping both Hierarchical Power and Gender as independent variables. We perform this on the Train subset of the APGI subset of our corpus. Table 5 shows the results for thread level version of the features (we obtain similar significance results at the interaction level as well). As can be seen from the ANOVA results, the mean values of many features differ significantly for the factorial

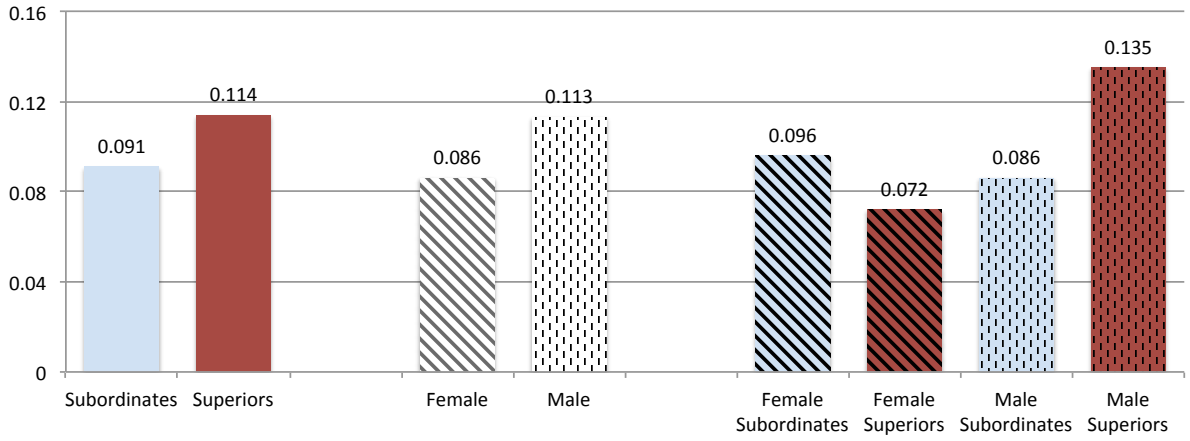

Figure 1: Mean values of ODPCounts in different groups: Subordinates vs. Superiors; Female vs. Male; across all combinations of Hierarchical Power and Gender.

groups of Hierarchical Power and Gender. For example, ReplyRate was highly significant; female superiors obtain the highest reply rate.

It is crucial to note that ANOVA only determines that there is a significant difference between groups, but does not tell which groups are significantly different. In order to ascertain that, we must use the Tukey's HSD (Honest Significant Difference) Test. We do not describe the analysis of all our features to that depth in this paper due to space limitations. Instead, we investigate specific hypotheses which we have derived from sociolinguistic literature. The first hypothesis we investigate is:

- **Hypothesis 1**: Female superiors tend to use "face-saving" strategies at work that include conventionally polite requests and impersonalized directives, and that avoid imperatives (Herring, 2008).

As a stand-in for a face-threatening communicative strategy, we use our "Overt Display of Power" feature (ODP). An ODP limits the addressee's range of possible responses, and thus threatens his or her (negative) face.[6] We thus reformulate our hypothesis as follows: the use of ODP by superiors changes when looking at the splits by gender, with female superiors using fewer ODPs than male superiors. We look further into the ANOVA analysis of the thread-level ODPCount treating Hierarchical Power and Gender as independent variables. Figure 1 shows the mean values of ODP counts in each group of participants. A summary of the results follows.

Hierarchical Power was significant. Subordinates had an average of 0.091 ODP counts and Superiors had an average of 0.114 ODP counts. Gender was also significant; Females had an average of 0.086 ODP counts and Males had an average of 0.113 ODP counts. When looking at the factorial groups of Hierarchical Power and Gender, however, several results were very highly significant. The significantly different pairs of groups, as per the Tukey's HSD test, are Male Superiors/Male Subordinates, Male Superiors/Female Superiors, and Male Superiors/Female Subordinates. Male Superiors used the most ODPs, with an average of 0.135 counts. Somewhat surprisingly, Female Superiors used the *least* of the entire group, with an average of 0.072 counts. Among Subordinates, Females actually used slightly more ODP, with an average of 0.096 counts. Male Subordinates had an average of 0.086 ODP counts. However, the differences among these three groups (Female Superiors, Female Subordinates, and Male Subordinates) are not significant.

The results confirm our hypothesis: female superiors use fewer ODPs than male superiors. However, we also see that among women, there is no significant difference between superiors and subordinates, and the difference between superiors and subordinates in general (which is significant) is entirely due to men. This in fact shows that a more specific (and more interesting) hypothesis than our original hypothesis is validated: only male superiors use more ODPs than subordinates.

[6]For a discussion of the notion of "face", see (Brown and Levinson, 1987).

6 Gender Environment and Power

We now turn to gender environments and their relation to the expression of power in written dialogs. We again start with a hypothesis based on the sociolinguistic literature.

- **Hypothesis 2**: Women use language to create and maintain social relations, for example, they use more small talk (based on a reported "stereotype" in (Holmes and Stubbe, 2003)).

We first define more formally what we mean by "gender environment" (Section 6.1), and then investigate our hypothesis (Section 6.2).

6.1 The Notion of "Gender Environment"

The notion of "gender environment" refers to the gender composition of a group who are communicating. In the sociolinguistic studies we have consulted (Holmes and Stubbe, 2003; Herring, 2008), the notion refers to a stable work group who interact regularly. Since we are interested in studying email conversations (threads), we adapt the notion to refer to a single thread at a time. Furthermore, we assume that a discourse participant makes communicative decisions based on (among other factors) his or her own gender, and based on the genders of the people he or she is communicating with in a given conversation (i.e., email thread). We therefore consider the "gender environment" to be specific to each discourse participant and to describe the other participants from his or her point of view. Put differently, we use the notion of "gender environment" to model a discourse participant's (potential) audience in a conversation. For example, a conversation among five women and one man looks like an all-female audience from the man's point of view, but a majority-female audience from the women's points of view.

We define the gender environment of a discourse participant p in a thread t as follows. As discussed, we assume that the gender environment is a property of each discourse participant p in thread t. We take the set of all discourse participants of the thread t, P_t (see Section 4.1), and exclude p from it: $P_t \setminus \{p\}$. We then calculate the percentage of women in this set.[7] We obtain three groups by setting thresholds on these percentages. Finer-grained gender environments resulted in partitions of the data with very few instances, since most of our data involves fairly balanced gender ratios. The three gender environments we use are the following:

- **Female Environment**: if the percentage of women in $P_t \setminus \{p\}$ is above 66.7%.
- **Mixed Environment**: if the percentage of women in $P_t \setminus \{p\}$ is between 33.3% and 66.7%.
- **Male Environment**: if the percentage of women in $P_t \setminus \{p\}$ is below 33.3%

Across all threads and discourse participants in the threads, we have 791 female, 2087 mixed and 1642 male gender environments.

6.2 Gender Environment and Conventional Dialog Acts

We now turn to testing Hypothesis 2. We have at present no way of testing for "small talk" as opposed to work-related talk, so we instead test Hypothesis 2 by asking how many conventional dialog acts a person performs. Conventional dialog acts serve not to convey information or requests (both of which would typically be work-related in the Enron corpus), but to establish communication (greetings) and to manage communication (sign-offs); since communication is an important way of creating and maintaining social relations, we can say that conventional dialog acts serve the purpose of easing conversations and thus of maintaining social relations. Since this aspect of language is specifically dependent on a group of people (it is an inherently social function), we assume that the relevant feature is not simply Gender, but Gender Environment. Specifically, we make our Hypothesis 2 more precise by saying that a higher number of conventional dialog acts is used in Female Environments. We use the thread level version of the feature ConventionalCount.

Figure 2 shows the mean values of ConventionalCount in each sub-group of participants. Hierarchical Power was highly significant as per ANOVA results. Subordinates use conventional language more (0.60 counts) than Superiors (0.52). Gender is a very highly significant variable; Males use 0.60 counts on average, whereas

[7] We note that one could also define the notion of gender environment at the level of individual emails: not all emails in a thread involve the same set of participants. We leave this to future work.

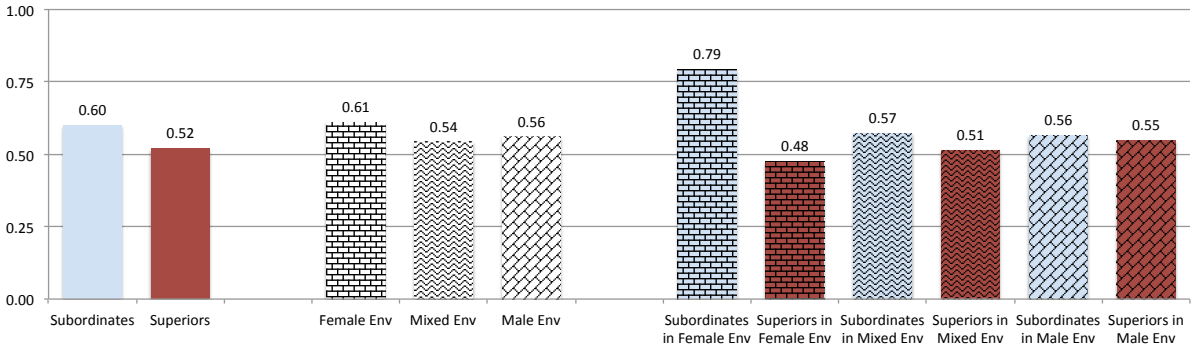

Figure 2: Mean values of Conventional Counts: Subordinates vs. Superiors; across all Gender Environments; across all combinations of Hierarchical Power and Gender Environments.

Females use 0.50. This result is somewhat surprising, but does not invalidate our Hypothesis 2, since our hypothesis is not formulated in terms of Gender, but in terms of Gender Environment. The analysis of Gender Environment at first appears to be a negative result: while the averages by Gender Environment differ, the differences are not significant. However, the groups defined by both Hierarchical Power *and* Gender Environment have highly significant differences. Subordinates in Female Environments use the most conventional language of all six groups, with an average of 0.79. Superiors in Female Environments use the least, with an average of 0.48. Mixed Environments and Male Environments differ, but are more similar to each other than to Female Environments. In fact, in the Tukey HSD test, the only significant pairs are exactly the set of subordinates in Female Environments paired with each other group (Superiors in Female Environments, and Subordinates and Superiors in Mixed Environments and Male Environments). That is, Subordinates in Female environments use significantly more conventional language than any other group, but the remaining groups do not differ significantly from each other.

Our hypothesis is thus only partially verified: while gender environment is a crucial aspect of the use of conventional DAs, we also need to look at the power status of the writer. In fact only subordinates in female environments use more conventional DAs than any other group (as defined by power status and gender environment). While our hypothesis is not fully verified, we interpret the results to mean that subordinates are more comfortable in female environments to use a style of communication which includes more conventional DAs than outside the female environments.

7 Predicting Power in Participant Pairs

In this section, we use the formulation of the power prediction problem presented in our prior work (Prabhakaran and Rambow, 2014). Given a thread t and a pair of participants $(p_1, p_2) \in RIPP_t$, we want to automatically detect $HP(p_1, p_2)$. We use the SVM-based supervised learning system from (Prabhakaran and Rambow, 2014) that can predict $HP(p_1, p_2)$ to be either *superior* or *subordinate* based on the interaction within a thread t for any pair of participants $(p_1, p_2) \in RIPP_t$. The order of participants in (p_1, p_2) is fixed such that p_1 is the sender of the first message in $IM_t(p_1, p_2)$. The power prediction system is built using the ClearTK (Ogren et al., 2008) wrapper for SVMLight (Joachims, 1999) package. It uses a quadratic kernel to capture feature-feature interactions, which is very important as we see in Section 5 and 6. We use the Train, Dev and Test subsets of the APGI subset of our corpus for our experiments. We use the related interacting participant pairs in threads from the Train set to train our models and optimize our performance on those from the Dev set. We report results on both Dev and Test sets.

In addition to the features described in Section 4.2, the power prediction system presented in (Prabhakaran and Rambow, 2014) uses a lexical feature set (LEX) that captures word ngrams, POS (part of speech) ngrams and mixed ngrams, since lexical features have been established to be very useful for power prediction. Mixed ngrams are word ngrams where words belonging to open classes are replaced with their POS tags. We add two gender-based feature sets: GEN containing the gender of both persons of the pair and ENV containing the gender environment feature.

Table 6 presents the results obtained using various feature combinations. We experimented using all subsets of {LEX, THR^{STR}, THR^{META}, DIA, GEN, ENV } on the Dev set; we report the most interesting results here. The majority baseline (*subordinate*) obtains an accuracy of 55.8%. Using the gender-based features alone performs only slightly better than the majority baseline. We use the best performing feature subset from (Prabhakaran and Rambow, 2014) (LEX + THR^{META}) as another baseline, which obtains an accuracy of 68.2%. Adding the GEN features improves the performance to 70.6%. Further adding the ENV features improves the performance, but only marginally to 70.7% (our overall best result, an improvment of 2.4% points). The best performing feature set without using LEX was the combination of DIA, THR^{META} and GEN (67.3%). Removing the gender features from this reduced the performance to 64.6%. Similarly, the best performing feature set which do not use the content of emails at all was THR^{STR} + THR^{META} + GEN (66.6). Removing the gender features decreases the accuracy by a larger margin (5.4% accuracy reduction to 63.0).

We interpret the differences in absolute improvement as follows: the gender-based features on their own are not very useful, and gain predictive value only when paired with other features. This is because the other features in fact make quite different predictions depending on gender and/or gender environment. However, the content features (and in particular the lexical features) are so powerful on their own that the relative contribution of the gender-based features decreases again. Nonetheless, we take these results as validation of the claim that gender-based features enhance the value of other features in the task of predicting power relations.

We performed another experiment where we partitioned the data into two subsets according to the gender of the first person of the pair and trained two separate models to predict power. At test time, we chose the appropriate model based on the gender of the first person of the pair. However, this did not improve the performance.

On our blind test set, the majority baseline obtains an accuracy of 57.9% and the (Prabhakaran and Rambow, 2014) baseline obtains an accuracy of 68.9%. On adding the gender-based features, the accuracy of the system improves to 70.2%.

Description	**Accuracy**
Majority (Always Subordinate)	55.83
GEN	57.59
GEN + ENV	57.59
Baseline (LEX + THR^{META})	68.24
Baseline (LEX + THR^{META}) + GEN	70.56
Baseline (LEX + THR^{META}) + GEN + ENV	**70.74**
DIA + THR^{META} + GEN	67.31
DIA + THR^{META}	64.63
THR^{STR} + THR^{META} + GEN	66.57
THR^{STR} + THR^{META}	62.96

Table 6: Accuracies on feature subsets (Dev set). THR^{META}: meta-data; THR^{STR}: structural; DIA: dialog-act; GEN: gender; ENV: gender environment; LEX: ngrams;

8 Conclusion

We presented a new, freely available resource: the Gender Identified Enron Corpus, and explored the relation between power, gender, and language using this resource. We also introduced the notion of gender environment, and showed that the manifestations of power differ significantly between gender environments. We also showed that the gender-related features helps in improving power prediction. In future work, we will explore machine learning algorithms which capture the interactions between features better than our SVM with quadratic kernel.

We expect our corpus to be a rich resource for social scientists interested in the effect of power and gender on language use. We will investigate several other sociolinguistic-inspired research questions; for example, do the strategies managers use for "effectiveness" of communication differ based on gender environments?

While our findings pertain to the Enron data set, we believe that the insights and techniques from this study can be extended to other genres in which there is an independent notion of hierarchical power, such as moderated online forums.

Acknowledgments

This paper is based upon work supported by the DARPA DEFT Program. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. We thank several anonymous reviewers for their constructive feedback.

References

Apoorv Agarwal, Adinoyi Omuya, Aaron Harnly, and Owen Rambow. 2012. A comprehensive gold standard for the enron organizational hierarchy. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 161–165, Jeju Island, Korea, July. Association for Computational Linguistics.

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

Or Biran, Sara Rosenthal, Jacob Andreas, Kathleen McKeown, and Owen Rambow. 2012. Detecting influencers in written online conversations. In *Proceedings of the Second Workshop on Language in Social Media*, pages 37–45, Montréal, Canada, June. Association for Computational Linguistics.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *ACL*, pages 773–782. The Association for Computational Linguistics.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness : Some Universals in Language Usage (Studies in Interactional Sociolinguistics)*. Cambridge University Press, February.

Na Cheng, R. Chandramouli, and K. P. Subbalakshmi. 2011. Author gender identification from text. *Digit. Investig.*, 8(1):78–88, July.

Malcolm Corney, Olivier de Vel, Alison Anderson, and George Mohay. 2002. Gender-preferential text mining of e-mail discourse. In *Computer Security Applications Conference, 2002. Proceedings. 18th Annual*, pages 282–289. IEEE.

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 the 21st international conference on World Wide Web*, WWW '12, New York, NY, USA. ACM.

William Deitrick, Zachary Miller, Benjamin Valyou, Brian Dickinson, Timothy Munson, and Wei Hu. 2012. Author gender prediction in an email stream using neural networks. *Journal of Intelligent Learning Systems & Applications*, 4(3).

Eric Gilbert. 2012. Phrases that signal workplace hierarchy. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, pages 1037–1046, New York, NY, USA. ACM.

Susan C Herring. 2008. Gender and power in online communication. *The handbook of language and gender*, page 202.

Janet Holmes and Maria Stubbe. 2003. feminine workplaces: stereotype and reality. *The handbook of language and gender*, pages 572–599.

Thorsten Joachims. 1999. Making Large-Scale SVM Learning Practical. In Bernhard Schölkopf, Christopher J.C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, Cambridge, MA, USA. MIT Press.

Shari Kendall and Deborah Tannen. 1997. Gender and language in the workplace. In *Gender and Discourse*, pages 81–105. Sage, London.

Shari Kendall. 2003. Creating gendered demeanors of authority at work and at home. *The handbook of language and gender*, page 600.

Saif Mohammad and Tony Yang. 2011. Tracking sentiment in mail: How genders differ on emotional axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 70–79, Portland, Oregon, June. Association for Computational Linguistics.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, Deborah A. Cai, Jennifer E. Midberry, and Yuanxin Wang. 2013. Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*, pages 1–41.

Philip V. Ogren, Philipp G. Wetzler, and Steven Bethard. 2008. ClearTK: A UIMA toolkit for statistical natural language processing. In *Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP workshop at Language Resources and Evaluation Conference (LREC)*.

Adinoyi Omuya, Vinodkumar Prabhakaran, and Owen Rambow. 2013. Improving the quality of minority class identification in dialog act tagging. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 802–807, Atlanta, Georgia, June. Association for Computational Linguistics.

Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the enron corpus. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 86–95, Portland, Oregon, June. Association for Computational Linguistics.

Vinodkumar Prabhakaran and Owen Rambow. 2013. Written dialog and social power: Manifestations of different types of power in dialog behavior. In *Proceedings of the IJCNLP*, pages 216–224, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Vinodkumar Prabhakaran and Owen Rambow. 2014. Predicting power relations between participants in written dialog from a single thread. In *Proceedings of the 52nd Annual Meeting of the Association*

for Computational Linguistics (Volume 2: Short Papers), pages 339–344, Baltimore, Maryland, June. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2012. Predicting Overt Display of Power in Written Dialogs. In *Human Language Technologies: The 2012 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Montreal, Canada, June. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Ajita John, and Dorée D. Seligmann. 2013. Who had the upper hand? ranking participants of interactions based on their relative power. In *Proceedings of the IJCNLP*, pages 365–373, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Vinodkumar Prabhakaran, Ashima Arora, and Owen Rambow. 2014. Power of confidence: How poll scores impact topic dynamics in political debates. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, page 49, Baltimore, MD, USA, June. Association for Computational Linguistics.

Tomek Strzalkowski, Samira Shaikh, Ting Liu, George Aaron Broadwell, Jenny Stromer-Galley, Sarah Taylor, Umit Boz, Veena Ravishankar, and Xiaoai Ren. 2012. Modeling leadership and influence in multi-party online discourse. In *Proceedings of COLING*, pages 2535–2552, Mumbai, India, December. The COLING 2012 Organizing Committee.

Jen-Yuan Yeh and Aaron Harnly. 2006. Email thread reassembly using similarity matching. In *CEAS 2006 - The Third Conference on Email and Anti-Spam, July 27-28, 2006, Mountain View, California, USA*, Mountain View, California, USA, July.