# Social Power in Interactions: Computational Analysis and Detection of Power Relations

**Vinodkumar Prabhakaran**

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

**COLUMBIA UNIVERSITY**

2015

# ABSTRACT

## Social Power in Interactions: Computational Analysis and Detection of Power Relations


## Vinodkumar Prabhakaran



In this thesis, I investigate whether social power relations between individuals are manifested in the language and structure of their social interactions, and if so, in what ways, and whether we can use the insights gained from this study to build computational systems that can automatically identify these power relations. I analyze eleven different linguistic and/or structural aspects of interactions to study manifestations of power. To further understand these manifestations of power, I extend this study in two ways. First, I investigate whether a person's gender and the gender makeup of an interaction (e.g., are most participants female?) affect the manifestations of his/her power (or lack of it) and whether the gender information can help improve the predictive performance of an automatic power prediction system. Second, I investigate whether different types of power manifest differently in interactions, and whether they exhibit different but predictable patterns in the aspects of interactions we analyze. I perform this study on interactions from two different genres: organizational emails, which contain task oriented written interactions, and political debates, which contain discursive spoken interactions.

# Table of Contents

# List of Figures

# List of Tables

# Acknowledgments

Graduate school at Columbia has been an exciting journey, thanks to New York City, the wonderful friends it afforded me, the numerous distractions it offered that kept me sane through the years, and oh yeah, its numerous coffee shops that let me read/write/think on their premises for entire days. Beyond that, I would like to give a shout out here to some of the people who played important roles in my journey this far.

I was fortunate to have an excellent PhD thesis committee (Owen Rambow, Mona Diab, Kathy McKeown, Julia Hirschberg, and Lyn Walker). I could not have asked for a better PhD advisor than Owen. He always gave me the freedom to pursue my research in the directions I wanted to take it to and brainstormed with me to iron out the details. Thanks Owen for your constant guidance as an advisor and for also being a friend at the same time. Owen made getting the PhD a delightful experience, contrary to popular belief. It was a part-time job I took up with Mona in the spring of 2009 that led me to this PhD. Thanks Mona for believing in me and offering me this wonderful opportunity. Thanks Kathy for triggering my interest in NLP as a first year graduate student, and for your continued guidance. Thanks Julia for the encouragement over the years, and the feedback on earlier versions of this thesis. Thanks Lyn for being so very interested in my work, and for giving me valuable suggestions from the proposal stage onwards.

As part of my research, I worked with a set of brilliant graduate and undergraduate students, whose contributions were integral to this thesis. I thank Tucker Kuman for his help in carefully annotating power relations in email, Wisdom Omuya for his dedication in identifying dialog acts in email, Ashima Arora for her enthusiasm in modeling topic shifts in debates, Emily Reid for passionately studying gender variations in manifestations of power, Yanxi Pan for the numerous experiments she performed in record time, Prem Ganeshkumar for exploring how expressions of belief relates to power, and Michael Saltzman for making an elegant software that showcases a major part of this thesis.

Beyond my committee and the students I worked with, some individuals have had great impact on me as a researcher, and on this thesis. First, I thank Janet Kayfetz for teaching me how to present my work in writing and how to talk about it persuasively. The Academic Writing and Great Presentations courses I took with her have had immense impact on my research career. Next, I would like to thank Ajita John, who was an excellent mentor during my internship at Avaya Labs. I thank Ajita for letting me work on political debates, which became a major part of my thesis, and for all the encouragement and guidance she continued to give me even after the internship. Thanks Dorée Seligman for showing me how to place my research in the bigger industrial context. I am also thankful to the three other industrial internships I did, each of which were extremely beneficial. Thanks Swapna Somasundaran for giving very helpful feedback during my internship at Siemens Corporate Research, Chang Wang and Bran Boguraev for being great mentors at IBM Research and Hila Becker for helping me have a wonderful time at Google.

The vibrant NLP research environment at Columbia, both at the center for computational learning systems and at the computer science department, has enriched my experience immensely. I thank Becky Passonneau, Smaranda Muresan, and Nizar Habash for their support over the years. I also learned from some of the most eminent professors at Columbia. I thank Rocco Servedio and Anargyros Papageorgiou for showing what perfection in teaching looks like and how to make even the most complex concepts seem obvious to a beginner. Thanks Michael Collins for taking me on as a teaching assistant and showing me how to deal with a class full of brilliant students.

I am also grateful for the many friends who knowingly or unknowingly aided me in making critical decisions over the years. I thank Mayur Lodha for making me rethink my decision to turn down the Masters admission from Columbia. I thank Pravin Bhutada for introducing me to Mona. I thank Snehit Prabhu and Shreeharsh Kelkar for encouraging me to take up the PhD offer when I was apprehensive. I thank Pallika Kanani for the off-hand comment at a conference that triggered me to actually start building my professional network. I thank Yuval Marton for the numerous career tips and Daniel Bauer for the many chats over coffee about work and life in general. I also thank Apoorv Agarwal, Mohamed Altantawy, Yassine Benajiba, Or Biran, Heba Elfardy, Joshua Gordon, Ahmed El Kholy, Ramy Eskander, Noura Farra, Weiwei Guo, Faiza Khattak, Yves Petinot, Mohammad Rasooli, Hooshmand Razaghi, Sara Rosenthal, Wael Salloum, Kart Stratos, Swabha Swayamdipta, Kapil Thadani, Ilia Vovsha, and Boyi Xie for their support and friendship.

Finally, I am forever indebted to my close friends and family for their unending love and support. I thank Mayur for patiently helping me navigate the first few years of graduate school. I thank Sam for his love and support that helped me survive the ups and downs of the final years of my PhD. I thank my brother Pramod for his love, care and support through the years. Finally, I thank my parents to whom I owe all my accomplishments. Both my parents had to endure many financial and cultural struggles to pursue their education. The value I give for education is a result of growing up hearing their success stories. I am extremely fortunate that they made it easy for me to dream freely and to pursue those dreams.

To the stranger
03.11.2002

# Chapter 1

# Introduction

In recent years, there has been a rapid increase in the amount of social interaction that are captured and stored electronically in various digital repositories. In addition to those interactions that are inherently online such as emails, discussion forums, and social networking websites, various offline interactions such as debates, speeches, and teleconferences are also captured in real time and stored online in repositories such as YouTube and news media outlets. This growing mass of data representing various modes of interactions enables researchers to computationally analyze social interactions at a scale which was not feasible previously. Researchers have studied different dimensions of social interactions such as network structures of interactants (Diesner and Carley 2005, Rowe et al. 2007), propagation of opinions and influence through these networks (Bakshy et al. 2011), and the dynamics between participants of those interactions and their inter-relations (Peterson et al. 2011, Danescu-Niculescu-Mizil et al. 2012).

When people interact with one another, there is often a power differential that affects the way they interact. This differential may be derived from a multitude of factors such as social status, authority, experience, age, gender etc. One of the primary ways power is manifested in interactions is in the manner in which people participate in them. Power relations can sometimes constrain a dialog participant in his/her dialog behavior, whereas in some other cases, power relations enable him/her to constrain someone else's behavior. And in some cases, the dialog behavior becomes a tool to express, maintain and even pursue power. By dialog behavior, we mean the choices a dialog participant makes while engaging in interactions. It includes choices with respect to the message content, such as lexical choices, degree of politeness or overt displays of power such as

orders and commands. It also includes choices participants make in terms of dialog structure, such as the choices of when to participate with how much and what sort of contribution, whether to ask questions or respond to others' questions, and whether to stay on topic when responding. Within the field of computationally analyzing social interactions, there is growing interest in understanding how social power relations between participants are reflected in various facets of interactions, and if they can be detected using computational means (e.g., (Rowe et al. 2007, Bramsen et al. 2011, Gilbert 2012, Danescu-Niculescu-Mizil et al. 2012)).

## 1.1 Why Computationally Analyze Power?

Studying manifestations of power using computational means enables researchers to perform large-scale sociolinguistics studies which in turn help answer some of the fundamental questions in social sciences about power and how it affects the way we interact with one another. While the insights gained from this line of research are by themselves valuable, they can also aid in building computational systems that can automatically detect power relations. Identifying the powerful participants of an interaction through such computational means has various practical applications.

Power analysis can aid law enforcement and intelligence agencies to detect leaders and influential members in suspicious online communities. In recent years, there has been an exponential growth in the proliferation of websites and online communities that disseminate extremist propaganda (Janbek and Prado 2012). While monitoring such websites, the intelligence agencies would greatly benefit from being able to automatically infer the power structures that exist within the online communities those websites cater to. This capability is especially useful since the real identities of the members of such communities are often not revealed and the hierarchies of such communities are not be available to the intelligence agencies.

Power analysis also has many business applications. For example, it can help maximize the reach of advertisements in an online community by targeting them to its powerful members who have influence on others in the community. Market research also stands to gain from power analysis. The increasing impact of online forums on shaping consumers' purchase intentions has already been established (e.g., (Prendergast et al. 2010, Ludwig et al. 2013, King et al. 2014)). Identifying opinion leaders in such online communities and focusing on their needs and preferences will benefit

businesses in the long run, since obtaining favorable opinions from the leaders can influence their followers' future purchase intentions.

There are also many technology applications of power analysis. For example, conversations summarization systems (e.g., email summarization (Rambow et al. 2004), blog summarization (Hu et al. 2007) etc.) can benefit from knowing the power relations between the participants, since it allows us to make more informed choices on which parts of the conversation should be included in a summary. Similarly, information retrieval can also benefit from power analysis. Revealing power dynamics in online forums can help determine relevance for a user with information needs. For example, users in knowledge sharing forums may want to limit their search to posts by authors with higher power.

Power analysis can also have an impact in the efficacy of Massive Open Online Courses (MOOC), which are revolutionizing the field of higher education. The size of such online classrooms makes it a harder task for instructors to provide sufficient and timely feedback to all students. In this context, student leaders who voluntarily help other students in the forums play an important role in improving student engagement and reducing attrition, which is one of the major pain points in online classrooms (Moon et al. 2014). Moon et al. (2014) show that automatically identifying such student leaders from the online discussions can enable better tracking of leadership shown by students and incorporating that factor into the multi-dimensional student evaluations that are followed in online courses, which in turn encourages more student to become leaders.

## 1.2 Computational Analysis of Power: A Brief Review

A significant number of early studies in the field of analyzing or detecting power relations have been performed in the domain of organizational email using the Enron email corpus (Klimt and Yang 2004) where there is a well defined notion of power (i.e., organizational hierarchy). Early computational approaches relied on social network analysis on collections of interactions (e.g., (Diesner and Carley 2005, Rowe et al. 2007)) using information only from the meta-data of emails. A limitation of this line of research is that it relies solely on the availability of large collections of interactions between the same set of people. In addition, network analysis based approaches often ignore the content of the interaction, thereby ignoring clues within the language used by the participants. Re-

cently, Natural Language Processing (NLP) techniques have been applied to predict power relations between people in the Enron email corpus using lexical features extracted from *all* the messages exchanged between them (e.g., (Bramsen et al. 2011, Gilbert 2012)). These approaches also require fairly large number of messages exchanged between the pairs of people since they rely solely on lexical cues. However, by taking the messages out of the context of the interactions in which they were exchanged, they fail to utilize the structure of those interactions, which may hold important clues about power relations.

Sociolinguists have long argued the importance of fundamental personal attributes such as age and gender in studying correlates of social aspects in language. However, within the computational approaches towards analyzing manifestations of power in interactions, such personal attributes have been largely ignored. There has not been much research on understanding how the manifestations of power differ based on personal demographic attributes such as age and gender, and whether these personal attributes could be helpful in automatically predicting power relations.

Another limitation with most of the early computational approaches in analyzing power is that they rely solely on static power structures such as corporate hierarchies as the source of the power differential between participants. However, many interactions happen outside the context of a pre-defined static power structure or hierarchy. Examples for such interactions include political debates, online discussions, and email interactions outside organizational boundaries. Although the participants of these interactions may not be part of an established power structure, there is often different types of power differentials between them drawn from various factors such as experience, knowledge, popularity etc. In such situations, the interaction itself plays an important role as a medium for the interactants to pursue, gain and maintain power over others. Consequently, the manifestations of power in such interactions will also inherently differ from the cases where a hierarchy is present. However, not much work has been done to understand how different types of power differ in the ways they affect how people interact in dialog.

In this thesis, we address some of these shortcomings in the existing research in this field. In Section 1.3, we describe an overview of the thesis. While this overview does not directly translate to the way the chapters in this thesis are structured, it gives a fair idea of the overall scope of this thesis, which serves as the background for discussing the specific research questions we state in Section 1.4 and for summarizing the contributions in Section 1.5.

## 1.3 Thesis Overview

In this thesis, we investigate whether social power relations are manifested in both *the language and the structure* of social interactions, and if so, in what ways, and whether we can use these insights to build computational systems that can automatically identify these power relations by analyzing social interactions. The focus of the thesis is on being able to make these predictions based solely on *single threads of interactions*, rather than requiring large collections of interactions. To further understand the manifestations of power, we deepen our research in three ways. First, we investigate whether a person's *gender* and the *gender environment of an interaction* (e.g., are most participants female?) affect the manifestations of his/her power (or lack of it) and whether gender information can help improve the predictive performance of an automatic power prediction system. Second, we study whether the *levels of beliefs* expressed by participants (i.e., whether participants are committed to the beliefs they express, non-committed to them, or express beliefs attributed to someone else?) correlate with power relations and whether we can use use them to improve power prediction performance. Third, we investigate whether *different types of power* manifest differently in interactions, and whether they exhibit different but predictable patterns. We study the manifestations of power across two genres: organizational emails, which contains task oriented written interactions, and political debates, which contains discursive spoken interactions.

The research presented in this thesis is not driven by the objective to build power prediction systems with the best accuracy, but to build accurate power prediction systems that also help understand the underlying social phenomena of power and its manifestations. The rest of this section gives a bird's eye view of the analysis presented in this thesis in terms of a) aspects of interaction we analyze, b) demographics of participants we study, c) types of power we study, and d) the genres of interactions we study.

### 1.3.1 Aspects of Interactions Analyzed

In this thesis, we analyze eleven different aspects of interactions. These aspects are collections of features that relate to different facets of an interaction. By teasing them apart, we hope to gain a better understanding of the manifestations of power in interactions. These aspects fall into three categories in terms of the level of NLP processing required to extract them — NonNLP, BasicNLP,

and DeepNLP. NonNLP aspects capture the structure of interactions without analyzing the textual content of messages (e.g., how often the participants speak, how do they take turns etc.). BasicNLP aspects require basic NLP processing, such as tokenization and part-of-speech tagging. For DeepNLP aspects, we need a deeper analysis of the content exchanged in the interactions. Our work makes considerable contributions in this area by developing automatic systems to analyze these deep aspects of interactions. Some of the DeepNLP aspects capture the structure of interactions within the content (e.g., who issues requests, whose requests get responses, who shifts topics etc.). Some others are extracted solely from the content of messages without considering the discourse structure or the context in which those messages were exchanged (e.g., how committed were the dialog participants to what they said). We briefly describe each aspect below.

#### 1.3.1.1 NonNLP Aspects

**POSITIONAL (PST):** In this aspect, we capture positional features such as whether the person initiated the conversation, and at what point in the interaction he/she started and stopped participating. We believe that these features reveal instances of participants taking initiative, or having the final word in the interactions. We obtain these features using only the author attribution information (i.e., who contributed which utterance), which is often captured in the representation of interaction data.

**VERBOSITY (VRB):** In this aspect, we capture how verbose the participant was within the interactions. This includes features such as how often they contributed to the interaction, how long their contributions were and how much they contributed overall to the interaction. To obtain these features, we look at the content of the interaction just to count number of words, without performing any NLP analysis.

**THREAD STRUCTURE (THR):** In this aspect, we capture the conversational thread structure as captured in the meta data of interactions such as who replies to whom, how many recipients did the messages have, and how many replies did they get. These features are obtained without looking at the textual content of the messages exchanged.

**TURN TAKING (TT):** Sociolinguistics studies have found that turn-taking and interruption patterns reveal social relations such as power and influence (Ng and Bradac 1993, Ng et al. 1993, Reid and Ng 2000). In this aspect, we capture the turn-taking patterns between participants. More specifically, we look at instances where participants speak out of turn, possibly interrupting others. We do not explore this aspect in written interactions, since the asynchronous nature of written interactions allows for simultaneous contributions without interrupting each other.

#### 1.3.1.2 BasicNLP Aspects

**LEXICAL (LEX):** Lexical features have previously been shown to be valuable in predicting power relations (Bramsen et al. 2011, Gilbert 2012). We also use lexical features in our study. We capture word-lemma and part of speech ngrams, along with a new formulation of mixed ngrams we introduce, in which open class words are replaced with part-of-speech tags. We obtain features in this aspect through basic NLP processing (tokenization and part-of-speech tagging).

**MENTIONS (MNT):** In this aspect, we look at how often a participant of an interaction is being mentioned within the interaction. We believe that being mentioned often might reveal the importance of a person, which in turn might be correlated with power. Following socio-linguistic studies (Brown and Ford 1961, Dickey 1997), we also investigated if the form of addressing used has any correlation with the kind of power a person may possess. We obtain features in this aspect also through basic NLP preprocessing.

#### 1.3.1.3 DeepNLP Aspects

**DIALOG ACTS (DA):** Dialog Act analysis, inspired by the speech act theory of Austin (1975) and Searle (1976), has been used in the NLP community to understand and model the structure of dialog. A dialog act is a high-level categorization of the pragmatic meaning of the utterance in an interaction. This is one of the aspects with which we capture the structure of interactions. We use both manual annotations and automatically generated dialog acts to perform this analysis. We build a dialog act tagger, which we describe in Chapter 5, in which we use a novel multi-class

classification algorithm to improve the minority class performance of a previously existed dialog act tagger.

**DIALOG LINKS (DL):** Dialog Acts only assign the pragmatic meaning of utterances. However, in the context of interactions, especially written asynchronous interactions, each utterance may be linked to any previous utterances (i.e., not necessarily the immediate previous one). Understanding these links may help us understand whose utterances were referred to later in the conversation, and whose requests were responded to. We perform this analysis using manual annotations for dialog links in Chapter 10.

**OVERT DISPLAY OF POWER (ODP):** We introduce a notion called "Overt Display of Power" (ODP) to capture instances where a speaker uses linguistic forms to signal that she/he is creating additional constraints on its response beyond those imposed by the general dialog act. We obtain manual annotations for instances of overt display of power and built an automatic tagger, which we will describe in Chapter 6. We use both manually generated and automatically tagged ODP assignments in our analysis.

**COMMITTED BELIEFS (CB):** We investigate whether the level of committed belief expressed by a speaker/writer to his/her propositions have any correlations with power relations. We use the committed belief annotations from (Diab et al. 2009) and (Prabhakaran et al. 2015) to build an automatic tagger that labels each propositional head in text as committed belief (CB), non committed belief (NCB), reported belief (ROB), and non-belief (NA). We will describe this analysis and our findings in more detail in Chapter 9.

**TOPIC SHIFTS (TS):** In this aspect, we capture the topical dynamics within interactions. This is an alternate way of modeling the structure of interactions, through which we capture who tries to shift topic of discussion and who stays on topic when responding to questions. We introduce two new methods to assign topics in interactions — one building on the traditional LDA (Blei et al. 2003) approach, and another using a variation of the Speaker Identity for Topic Segmentation (Nguyen et al. 2012) approach. We describe them in Chapter 12.

### 1.3.2 Analysis of Gender and Power

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. However, most computational approaches have ignored the effect of gender in manifestations of power. In this thesis, we will investigate two factors that affect manifestations of social power in communication: the writer's gender, and the gender of his or her fellow discourse participants (what we call the "gender environment"). Our goal is to study the interplay between the gender, gender environment, and power relations, and how they affect an individual's choices in the discourse.

### 1.3.3 Types of Power Analyzed

In social sciences, different typologies of power have been proposed (e.g., (French and Raven 1959, Wartenberg 1990)). Wartenberg (1990) makes the distinction between power-over and power-to in the context of interactions. Power-over refers to relationships between interactants set by external power structures, while power-to refers to the ability an interactant possesses within the interaction. Computational studies of power have not explored how different types of power differ in the ways they manifest in interactions.

In this thesis, we will analyze five different types of power — hierarchical power, situational power, power over communication, influence, and power of confidence — within and across genres. Our notions of hierarchical power, influence, and power of confidence are special cases of *power-over* derived from different static and dynamic external power structures, Hierarchical power is determined by organizational hierarchy; influence is determined by knowledge, expertise etc.; power of confidence stems from many other external sources. Our notions of situational power and power over communication are special cases of power-to. Situational power applies to the situation or task at hand, whereas power over communication applies to the interaction itself.

#### 1.3.3.1 Hierarchical Power (HP)

Hierarchical power is the notion of power most commonly used in computational studies on power. Here, the power differential between participants is drawn from the statuses they hold in a static power structure or hierarchy external to the interaction. A typical example for this type of power

relation is the superior-subordinate relation within an organizational setting. Similar hierarchies may also be in place in many online settings, such as moderators in online discussion forums. One of the aspects of this type of power is that it may be latent in many interactions; i.e., bosses need not always act bossy.

#### 1.3.3.2 Situational Power (SP)

Situational power is the power or authority someone has by being in charge of a situation or task. They will have the power to direct and/or approve another person's actions in the given situation or while a particular task is being performed. It is a more active notion of power than hierarchical power. Situational power is independent of hierarchy. A boss does not have situational power all the time. Also, someone with situational power may or may not be the boss. A typical example of someone with situational power is an HR personnel in the context of enforcing an HR policy within an organization.

#### 1.3.3.3 Power over Communication (PC)

A person is said to have power over communication if he/she actively attempts to achieve the intended goals of the communication. People with power over communication ask questions, request others to take action, make sure the conversation stays on topic etc. They do not just respond to questions or perform actions when directed to do so. A typical example for someone with power over communication is the chair or moderator of a meeting.

#### 1.3.3.4 Influence (INFL)

Influence is the power that stems from having the credibility in a group to be able to influence other's actions and opinions.[1] Influence can stem from a multitude of factors such as expertise, knowledge or information. The affordance of credibility could be explicit (e.g., by asking the influencer for an opinion) or implicit (e.g., by adopting the influencer's ideas or language). One distinguishing factor of influence is that there is no expectation by the influencer that his ideas/opinions be accepted.

[1]We derive this definition from the IARPA Socio-Cultural Content in Language (SCIL) program, where many of the researchers participating in the SCIL program contributed to the scope and refinement of the definition of a person with influence

In contrast, a person with situational power does expect the other person to take his advice and opinions.

#### 1.3.3.5 Power of Confidence (CONF)

Power of confidence is the power someone has due to the confidence drawn from some source(s) external to the conversation. This can be thought of as the power someone possess from being the front runner among his/her peers. For example, in an election campaign, someone who is higher up in the polls has the power of confidence over others who are trailing.

### 1.3.4 Genres Analyzed

In this thesis, we perform our research on two different genres: organizational emails, which contains task oriented written interactions, and political debates, which contains discursive spoken interactions. We describe each genre briefly below.

**Organizational Emails** The genre of Organizational emails is a rich one in which to explore manifestations of power since there is often a strong notion of power, i.e., the organizational hierarchy. Most email conversations happening within an organization are task oriented in nature. It also is a genre where our other research questions — effects of gender in the manifestations of power, and variations in manifestations of different types of power — have salience. We use the Enron email corpus for this study. We present a sample email thread from our corpus in Table 1.1.

**Political Debates** A second genre on which we perform our study on is the genre of political debates. Specifically, we choose presidential debates held as part of the US Presidential election campaigns. Presidential debates serve as a forum for candidates to discuss their stances on policy issues and contrast them with other candidates' stances. In addition, it also serves as a medium for the candidates to pursue and maintain power over other candidates. This makes it an interesting genre to investigate how power dynamics between participants are manifested in interactions. We analyze the debates associated with the 2012 Republican Presidential Primary elections in this thesis. We present an excerpt from one of the debates in Table 1.2 in which the candidates are discussing their positions on the issue of gay marriage.

**M1 — 05 Oct 2001 2:59 PM; From: Kim S Ward; To: Sara Shackleton;**

Sara,

Believe it or not, we are very close getting our signed ISDA from the City of Glendale. Steve Lins, the City attorney had a couple of questions which I will attempt to relay without having a copy of the documents.

1) I am assuming that he obtained a for legal opinion letter or document of some sort. [...] What is your opinion regarding this?

2) We sent him a couple of form documents to facilitate the documents required under the ISDA. [...] Will this suffice?

When you return, I may try to do one last conference call [...]

Thanks for your help,

**M2 — 08 Oct 2001 9:02 AM; From: Sara Shackleton; To: Kim S Ward; CC: Marie Heard;**

Kim: Can you obtain the name of Glendale's bond counsel (lawyer's name, phone number, email, etc.)?

Thanks. SS

**M3 — 08 Oct 2001 9:26 AM; From: Kim S Ward; To: Sara Shackleton;**

Glendale's City Attorney is Steve Lins. His phone number is 818-548-2080 and his email is slins@ci.glendale.ca.us. Please let me know if you need anything else. I will be in their offices on Wednesday.

**M4 — 08 Oct 2001 9:27 AM; From: Sara Shackleton; To: Kim S Ward;**

I need the city's bond counsel (outside counsel).

**M5 — 08 Oct 2001 10:03 AM; From: Kim S Ward; To: Sara Shackleton;**

Is this to obtain outside opinion? I thought we were going to do that at our own expense.

**M7 — 08 Oct 2001 10:38 AM; From: Sara Shackleton; To: Kim S Ward;**

We are going to do this at our own expense. But we would like to hire Glendale's bond counsel.

I don't know the name of Glendale's bond counsel or how to get in touch with them.

**M8 — 08 Oct 2001 11:43 AM; From: Kim S Ward; To: Sara Shackleton;**

I will work on this for you - and will be in touch. Thanks!

**M6 — 08 Oct 2001 10:15 AM; From: Marie Heard; To: Sara Shackleton;**

Sara: I do not see a copy of an opinion in the file nor have we received one since I sent the execution copies of the ISDA to Steve Lins.

**M9 — 08 Oct 2001 4:18 PM; From: Kim S Ward; To: Steve Lins;**

Steve, could you provide the name, phone number, etc. of your bond council for our attorney, Sara Shackleton? Thanks,

Table 1.1: Example email thread from the Enron email corpus. Subject line: "City of Glendale".

| Turn # | Turn Speaker and Text |
|---|---|
| 223 | PAWLENTY: I support a constitutional amendment to define marriage between a man and woman. I was the co-author of the state – a law in Minnesota to define it and now we have courts jumping over this. |
| 224 | KING (M): OK. Let's just go through this. |
| 225 | PAUL: The federal government shouldn't be involved. I wouldn't support an amendment. But let me suggest – one of the ways to solve this ongoing debate about marriage, look up in the dictionary. We know what marriage is all about. But then, get the government out of it. Why doesn't it go to the church? And why doesn't it to go to the individuals? I don't think government should give us a license to get married. It should be in the church. |
| 226 | KING (M): Governor Romney, constitutional amendment or state decision? |
| 227 | ROMNEY: Constitutional. |
| 228 | KING (M): Mr. Speaker? |
| 229 | GINGRICH: Well, I helped author the Defense of Marriage Act which the Obama administration should be frankly protecting in court. I think if that fails, at that point, you have no choice except to constitutional amendment. |
| 230 | KING (M): We heard the congresswoman's answer, Senator. |
| 231 | SANTORUM: Constitutional amendment. Look, the constitutional amendment includes the states. Three-quarters of the states have to – have to ratify it. So the states will be involved in this process. We should have one law in the country with respect to marriage. There needs to be consistency on something as foundational as what marriage is. |
| 232 | KING (M): Very quickly? |
| 233 | BACHMANN: John, I do support a constitutional amendment on – on marriage between a man and a woman, but I would not be going into the states to overturn their state law. |

Table 1.2: Excerpt from 2012 GOP primary debates.
Discussion of marriage equality. Goffstown, NH (06/13/11).

| | Organizational Email | Political Debates |
|---|---|---|
| POSITIONAL | ✓ | n/a |
| VERBOSITY | ✓ | ✓ |
| THREAD STRUCTURE | ✓ | n/a |
| TURN TAKING | n/a | ✓ |
| LEXICAL | ✓ | ✓ |
| MENTIONS | — | ✓ |
| DIALOG ACTS | ✓ | — |
| DIALOG LINKS | ✓ | — |
| OVERT DISPLAY OF POWER | ✓ | — |
| TOPIC SHIFTS | — | ✓ |
| COMMITTED BELIEFS | ✓ | — |

Table 1.3: Analysis overview: Interaction aspect analyzed across genres.
✓ indicates analysis that is done; n/a indicates analysis that is not applicable;
— denotes analysis that has not been included in this thesis

### 1.3.5 Summary

Thesis summary: Interaction aspect analyzed across genres

We summarize the aspects of interaction we analyze in each genre in Table 1.3. We do not explore POSITIONAL features in political debates since moderators always initiate the debates, and they decide the order in which they ask questions to candidates. We also do not investigate THREAD STRUCTURE in the debates since the thread structure is implicit in synchronous spoken conversations. Similarly, we do not investigate TURN TAKING in organizational email since its asynchronous nature allows for simultaneous contributions without interrupting others. In our initial investigation of dialog act based features in political debates, we found that the dialog structure almost always follows the pattern of the moderator asking questions and candidates answering them. Hence we excluded DIALOG ACTS, DIALOG LINKS or OVERT DISPLAY OF POWER features from our analysis in that genre. Similarly, we do not investigate TOPIC SHIFTS in organizational email since the email threads are relatively short (around three messages on average). There is potentially value in investigating MENTIONS in organizational email (e.g., (Agarwal et al. 2014)) and COMMITTED

BELIEFS in political debates. However, we do not analyze them in this thesis.

Table 1.4 summarizes the analysis of demographic attributes of participants. We study how gender of a participant and gender environment (i.e., the gender of other participants of an interaction) affect the manifestations of power. We perform this analysis only in the organizational email genre. There was only one female candidate in the 2012 Republican presidential primary campaign, the debates during which we use as our data. Hence we did not perform any analysis of gender in the political debates genre.

| | Organizational Email | Political Debates |
|---|---|---|
| Gender | ✓ | n/a |
| Gender Environment | ✓ | n/a |

Table 1.4: Analysis overview: Personal attributes analyzed across genres.
✓ indicates analysis that has been completed; n/a indicates analysis that is not applicable

In Table 1.5 we summarize which types of power we analyze in each genre. We study four types of power — hierarchical power, situational power, power over communication, and influence — in the organizational email genre. We study the power of confidence in the political debates genre.

| | Organizational Email | Political Debates |
|---|---|---|
| Hierarchical Power | ✓ | n/a |
| Situational Power | ✓ | n/a |
| Power over Communication | ✓ | n/a |
| Influence | ✓ | n/a |
| Power of Confidence | n/a | ✓ |

Table 1.5: Analysis overview: Power types analyzed across genres.
✓ indicates analysis that has been completed; n/a indicates analysis that is not applicable

## 1.4 Research Questions

In this section, we formally state the major research questions we address in this thesis. Table 1.6 lists the genres we investigate each research question in. At a very high level, we are asking the following four questions:

RQ 1 Are social power relations manifested in the language and structure of social interactions, and if so, in what ways?

RQ 2 Can we use the insights about these manifestations to build a computational system that can automatically identify power relations by analyzing social interactions?

RQ 3 Do a person's gender and the gender makeup of an interaction (e.g., are most participants female?) affect the manifestations of their power (or lack of it) and can gender information help improve the predictive performance of an automatic power prediction system?

RQ 4 Do different types of power manifest differently in interactions, and do they exhibit different but predictable patterns?

| | Organizational Email | Political Debates |
|---|---|---|
| RQ 1: Are power relations manifested in interactions? | ✓ | ✓ |
| RQ 2: Can these manifestations help us detect power? | ✓ | ✓ |
| RQ 3: Does gender affect manifestations of power? | ✓ | n/a |
| RQ 4: Do types of power differ in how they are manifested? | ✓ | n/a |

Table 1.6: Research questions investigated across genres.

✓ indicates analysis that has been completed; n/a indicates analysis that is not applicable

## 1.5 Contributions

The contributions of this thesis include both the insights gained from the various statistical analyses performed on the manifestations of power along different aspects of interaction, as well as the different datasets and computational systems built to analyze different aspects of interaction. Many of these resources have relevance to NLP problems beyond the research questions we ask in this particular thesis (for example, the Gender Identified Enron Corpus and the committed belief tagger). Table 1.7 lists the contributions, along with the associated publication (Table 1.8). We describe the main contributions in terms of datasets, algorithms, and systems below.

### 1.5.1 Datasets

**Overt Display of Power Annotations:** As part of this thesis, we built a corpus of 122 email threads from the Enron email corpus collection that are annotated with instances of overt display of power at an utterance level. The details of these annotations are described in Chapter 6. The annotated corpora has been made publicly available.[2]

**Power Types Annotations:** We also built a corpus of email threads annotated with different types of power relations between participants. The annotations capture instances of situational power, influence, power over communication, as well as perceived hierarchical power. The details of these annotations are described in detail in Chapter 10. The annotations are obtained on the same corpus in which the instances of overt display of power are captured and is included in the annotated corpus that has been made publicly available.

**Gender Identified Enron Corpus:** We released an extension to the Enron email corpus in which we have assigned the gender of authors of 87% of the emails. The procedure followed to perform the gender assignment is described in detail in Chapter 8. The Gender Identified Enron Corpus has been made publicly available.[3]

[2] http://www.cs.columbia.edu/~vinod/powerann.html

[3] http://www.cs.columbia.edu/~vinod/giec.html

| Contribution | Notion | Data/Annotations | Method | System | Chapter |
|---|---|---|---|---|---|
| Types of Power | | | | | |
| Hierarchical | | | | ACL14 | 7 |
| Situational | LREC12 | LREC12 | | COLING12 | 10 |
| Power over Comm. | LREC12 | LREC12 | | IJCNLP13a | 10 |
| Influence | | LREC12 | | IJCNLP13a | 10 |
| Power of Confidence | WWW13 | WWW13 | | IJCNLP13b | 11 |
| Extraction of Interaction Aspects | | | | | |
| DIALOG ACTS | | | NAACL13 | NAACL13 | 5 |
| OVERT DISPLAY OF POWER | NAACL12 | NAACL12 | NAACL12 | NAACL12 | 6 |
| COMMITTED BELIEFS | | | | COLING10 | 9 |
| TOPIC SHIFTS | | | NLPSD14 | | 12 |
| | | | EMNLP14b | EMNLP14b | 12 |
| User Attributes | | | | | |
| Gender | | EMNLP14a | EMNLP14a | EMNLP14a | 8 |
| Gender Environment | EMNLP14a | EMNLP14a | EMNLP14a | EMNLP14a | 8 |

Table 1.7: Contributions of this thesis.

| Reference | Venue Type | Mode | Citation |
|---|---|---|---|
| COLING10 | Conference Proceedings | Long Paper | (Prabhakaran et al. 2010) |
| LREC12 | Conference Proceedings | Long Paper | (Prabhakaran et al. 2012c) |
| NAACL12 | Conference Proceedings | Short Paper | (Prabhakaran et al. 2012b) |
| COLING12 | Conference Proceedings | Long Paper | (Prabhakaran et al. 2012d) |
| WWW13 | Conference Proceedings | Short Paper | (Prabhakaran et al. 2013b) |
| NAACL13 | Conference Proceedings | Short Paper | (Omuya et al. 2013) |
| IJCNLP13a | Conference Proceedings | Long Paper | (Prabhakaran and Rambow 2013) |
| IJCNLP13b | Conference Proceedings | Long Paper | (Prabhakaran et al. 2013a) |
| ACL14 | Conference Proceedings | Short Paper | (Prabhakaran and Rambow 2014) |
| NLPSD14 | Workshop Proceedings | Short Paper | (Prabhakaran et al. 2014b) |
| EMNLP14a | Conference Proceedings | Long Paper | (Prabhakaran et al. 2014c) |
| EMNLP14b | Conference Proceedings | Short Paper | (Prabhakaran et al. 2014a) |

Table 1.8: Descriptions of references to associated publications.

**Topical Non-substantivity Annotations:** In Chapter 12, we describe the annotations we obtained for topical non-substantivity of speaker turns in one of the presidential debates. We use these annotations to reliably detect instances of topic shifts in the debates.

### 1.5.2 Machine Learning Algorithms

**Minority Preference Multi-class SVM:** In Chapter 5, we introduce two new methods for SVM multi-class classification that improves the performance on minority class prediction — Divide and Conquer (DAC) and Cascaded Minority Preference (CMP). DAC is intertwined with the feature optimization experiments a researcher perform, whereas CMP is a generic modification to the original SVM one-vs-all prediction step. These approaches have already been applied to other problems by other researchers obtaining significant improvements (e.g., (Hou et al. 2013)). As part of this thesis research we have built the SVMlight-CMP package with a generic implementation of the CMP algorithm, that will be made publicly available.

### 1.5.3 Automatic NLP Systems

**A Direction of Power Predictor:** We built a supervised learning system that can automatically detect the direction of power between pairs of people in a conversation. We have built the gSPIN system — a Google Chrome browser extension — that seamlessly integrate this power prediction system with Gmail email threads. The direction-of-power predictor and the gSPIN plugin is explained in detail in Chapter 7.

**A Person with Power Detector:** We also built a person-with-power predictor to detect people with different types of power — situational power, hierarchical power, power over communication and influence — in written interactions. The details of each person-with-power predictor system is described in detail in Chapter 10.

**An Automatic Power Ranker:** In Chapter 11, we present an automatic power ranker, a supervised learning system that ranks participants of an interaction based on their relative power of confidence.

**An Improved Dialog Act Tagger:** Using the DAC and CMP methods for multi-class classification, we built a dialog act tagger with around 23% error reduction in minority class prediction

performance and an overall 10% accuracy error reduction. We have made this dialog act tagger available via the gSPIN browser extension that allows to invoke it on Gmail email threads.

**An Overt Display of Power Tagger:** Using the overt display of power annotations, we built an automatic tagger to detect instances of overt displays of power in interactions. This system is also made available via the gSPIN browser extension that allows to invoke the tagger on Gmail email threads.

**A New Committed Belief Tagger:** The first committed belief tagger (Prabhakaran et al. 2010) was built as part of this thesis. This tagger has since generated great research interest in applying it to other NLP tasks such as knowledge base population and sentiment/opinion analysis (Prabhakaran et al. 2015). We describe in detail the original tagger developed as part of this thesis in Chapter 9.

## 1.6 Thesis Outline

In this section, we give the outline of the thesis. We divide the thesis into four parts. The first part — Data and Methods — lays the foundation for the rest of the thesis, describing in detail the datasets we use in our study and the methods we use for analysis and to build our systems. The second part — Modeling Dialog Behavior — presents contributions that deal with modeling the dialog behavior of interactants. In the third part — Manifestations of Power in Dialog — we analyze the manifestations of power along different aspects of interactions, including the aspects we introduce in the second part. Part four concludes the thesis by summarizing the main contributions of this thesis and discussing future work. We now briefly describe what each chapter contains.

- Chapter 2 discusses the related work in the area of studying power. We first summarize the social science theories about power before discussing different stands of computational analysis of power. We also discuss a brief history of computational studies performed in both the organizational email genre and political debates genre. We postpone the discussion of literature that relates to specific strands of our analysis (e.g., dialog act modeling, committed belief analysis etc.) to their respective chapters.

- Chapter 3 describes the different datasets we use in this thesis in great detail. This chapter sets the background for the later chapters where we present the actual contributions of the thesis. Also, in describing the datasets, we limit our detailed discussions to preexisting resources/annotations; we postpone the discussion of resources that are contributions of this thesis (e.g., Overt Display of Power annotations) to their respective chapters.

- Chapter 4 describes the different methods we use for our analysis as well as to build our systems. We summarize the underlying software platform, machine learning algorithms, and feature representation techniques that are used across many chapters.

- Chapter 5 presents our work on obtaining the dialog act tags for our analysis. We introduce a two new multi-class classification methods in this chapter, both of which improves the performance of minority classes. We also present an improved dialog act tagger using these methods, and discuss the different experiments and results.

- Chapter 6 introduces the notion of overt display of power in interactions. We describe in detail the process of obtaining manual annotations, as well as present an automatic overt display of power tagger along with experiments and results. This is a major contribution of this thesis, as the problem, data, and the tagger are all introduced in this thesis.

- Chapter 7 is the first major power analysis chapter. In this chapter, we introduce the problem of predicting direction of power between pairs of people from single threads of interactions. We describe the problem formulation in detail and lay out the analysis framework that becomes the basis of analysis for further chapters (Chapters 8, 9 and 10). We present a detailed statistical analysis of how power is manifested along different dialog structural aspects of interactions. We also present an automatic direction-of-power predictor using dialog structure features, along with experiments and results.

- Chapter 8 has three major contributions. First, it describes the Gender Identified Enron Corpus, which is an extension to the Enron email corpus with 87% of email senders' gender identified. Second, it presents the results of a detailed statistical analysis of the interplay of power and gender. It also introduces the notion of gender environment to capture the gender makeup of the discourse participants of a particular interaction and presents a study of how

gender environment affects the manifestations of power. Third, it shows the utility of gender information in the task of predicting direction of power between pairs of participants.

- Chapter 9 has three major contributions. First, it presents an automatic tagger that detects the level of belief expressed about stated propositions. We discuss the tagger in great detail as well as present experiments and results. Second, we present a statistical analysis of whether the proportion of belief tags correlate with power relations. Third, we show different ways of integrating the belief tags into the machine learning framework for direction-of-power prediction.

- Chapter 10 focuses on studying how different types of power manifest in interactions. We describe the procedure of obtaining manual annotations of different types of power. We then present the results of a statistical analysis of how different types of power differ in how they patter in dialog structure features. We also present automatic person-with-power prediction systems for each type of power.

- Chapter 11 introduces our analysis of power in the genre of political debates. We describe in detail how we model power of confidence in that genre and present a statistical analysis of how the power of confidence is manifested along different structural aspects of the interaction in the debates. We then present an automatic ranking system rank the participants of a debate in terms of their relative power.

- Chapter 12 focuses on how topic dynamics correlate with power. In this chapter, we first present the results of our analysis on how the distribution of different topics correlate with power. We also describe different ways to detect topic shifts in debates and present our analysis of whether the topic shifting behavior of candidates correlate with their power. We also show that topic shifts can improve the predictive performance of the automatic power ranker presented in Chapter 11.

- Chapter 13 concludes the thesis and summarizes the major findings from the thesis. We also discuss the limitations of our analysis and plans for future work.

# Chapter 2

# Literature Review

In this chapter, we provide a comprehensive literature survey in order to situate this thesis among the large body of work on the study of power and social interactions. We start by discussing work in the social sciences about power, how power should be defined, what kinds of power are there, and how power is manifested in the language and structure of social interactions. We then discuss the related work in the field of computational analysis of organizational interactions in Section 2.2, followed by work on political speech (Section 2.3), both of which are genres of interactions we analyze in this thesis. Finally, we describe the growing array of work in computational analysis of social power in interactions in Section 2.4.

Our focus in this chapter is on work that relates to the overall thesis, i.e., work on power and on the genres of interactions we analyze in this thesis. We postpone the discussion of literature that relates only to specific strands of our analysis in the respective chapters where we introduce them. We list below those sections that discuss related literature as forward pointers:

- Chapter 5, Section 5.1, page 55: related work on speech act theory and dialog act analysis as the basis for our dialog structure analysis
- Chapter 6, Section 6.2, page 75: related work on face, politeness, and impoliteness as the theoretical framework for our notion of overt display of power.
- Chapter 8, Section 8.1, page 122: sociolinguistics studies on gender differences in language use and workplace interactions, as well as computational approaches on studying gender.

- Chapter 9, Section 9.1, page 159: related work on modeling expressions of cognitive states in text as a basis for our discussion of the notion of committed belief.
- Chapter 12, Section 12.1, page 241: related work on computational approaches towards modeling topic dynamics in interactions.

## 2.1 Power: Definitions, Typologies, and Manifestations

Power is a difficult concept to define, but is easily recognizable when expressed. There is a large body of literature in social sciences that studies power as a social construct (e.g., (Bierstedt 1950, French and Raven 1959, Dahl 1957, Emerson 1962, Pfeffer 1981, Handy 1985, Wartenberg 1990)) and how it relates to the ways people use language in social situations (e.g., (Bales et al. 1951, Bales 1970, O'Barr 1982, Van Dijk 1989, Bourdieu and Thompson 1991, Ng and Bradac 1993, Sexton and Helmreich 1999, Fairclough 2001, Locher 2004)). Most of the classical definitions of power in the sociology literature include "an element indicating that power is the capability of one social actor to overcome resistance in achieving a desired objective or result" (Pfeffer 1981). For example, Dahl (1957) defines power as follows: "A has power over B to the extent that he can get B to do something that B would not otherwise do". Emerson (1962) considers the basis of power to be dependency — "A depends on B if A has goals and needs that B can fulfill".

As a way to better understand power, sociolinguists have also categorized power into different types. One of the most widely used typologies of power is the five bases of power proposed by French and Raven (1959) and its extensions. French and Raven propose that power should be analyzed along the following five bases:

- *Reward power*: based on a person's ability to reward another person
- *Coercive power*: based on a person's ability to coerce another person into some action
- *Legitimate/Positional power*: based on a person's position with respect to another person that legitimizes the right to expect compliance
- *Referent power*: based on a person's perceived attractiveness (e.g., charisma)
- *Expert power*: based on a person's expertise or knowledge

Wartenberg (1990)'s notion of power is based on the relational work; i.e., considering power in a way that incorporates its exercise or reception by an individual in the context of an interaction. Wartenberg makes the distinction between two notions of power:

- *Power-over*: refers to power relations between interactants sourced from external power structures which can result in control, dominance etc.
- *Power-to*: refers to the ability an interactant may possess and uses, even if temporarily, within a situation

In his treatment of power, Wartenberg (1990) identifies the restriction of action environment of an interactant as a basic tenet of exercise of power. He identifies three types of exercise of power: *force*, *coercion*, and *influence* based on three different ways that can restrict the action environment of an interactant.

We find these definitions and typologies helpful as a general background, but too abstract to be used as the framework of analysis for our data-oriented study on how power is expressed in social interactions. However, our work draws great inspiration from these theories. Our notion of overt displays of power in interactions (Chapter 6) draws from the theory on action-restriction as a means of exercise of power. Similarly, the power typology we introduce in Chapter 10 aligns with the various typologies we discussed above. We consider our notions of hierarchical power, situational power and power over communication to be French and Raven (1959)'s positional power; although the former two can also have bases in coercion and rewards. The bases of our notion of influence are mainly referent and expert power. Our power typology can also be mapped to Wartenberg's distinctions of power. Our notions of hierarchical power and influence are special cases of Wartenberg's notion of power-over. Hierarchical power is determined by organizational hierarchy, whereas influence is determined by knowledge, expertise etc. Our notions of situational power and power over communication are special cases of power-to. Situational power applies to the situation/task at hand, while power over communication applies to the interaction itself.

Studies in social psychology have looked into the correlation between dialog behavior of a discourse participant and how influential he or she is perceived to be by the other discourse participants (Bales et al. 1951, Bales 1970, Scherer 1979, Brooke and Ng 1986, Ng et al. 1993; 1995). Specifically, factors such as frequency of contribution, proportion of turns, and number of successful

interruptions have been identified as being important indicators of influence. Bales (1970) argued that "To take up time speaking in a small group is to exercise power over the other members for at least the duration of the time taken, *regardless of the content* ... [emphasis added]". Simply successfully claiming the conversational floor represents a feat of power, regardless of what was spoken. Later, Ng and Bradac (1993) found that turns gained through interruptions were more powerful predictors of influence than non-interruptive turns. Like Bales (1970), they also argue that conversation is a resource to gain influence and power. However, in contrast to Bales (1970), they argue that the content of the turns play an important role in predicting influence and power. In fact, in later work (Reid and Ng 2000), they found evidence to that argument; interruptions gained through prototypical utterances (i.e., utterances that provide information that defines speakers and listeners within a given social context) were more strongly correlated with influence than interruptions encoded in non-prototypical utterances.

Sociolinguists have also studied the interaction of power and language use in great detail (e.g., (Brown and Gilman 1960, O'Barr 1982, Sexton and Helmreich 1999, Locher 2004, Pennebaker 2011)). Brown and Gilman (1960) introduced the distinction between the *T-form* (informal) vs. *V-form* (formal) pronouns. They argued that the use of informal pronouns would lead to solidarity between interactants, whereas the use of formal pronouns would lead to distance between interactants. Along the same line, Brown and Ford (1961) extended this idea to addressing forms; they argue that different address forms can lead to intimacy or distance. O'Barr (1982) analyzed courtroom conversations and defined linguistic markers that denote "powerful" and "powerless" speech. They characterized the powerless speech with frequent use of intensifiers, hedges, hesitation forms, and questioning intonation, whereas powerful style had less frequent use of these markers. Sexton and Helmreich (1999) analyzed cockpit conversations and found that linguistic indicators identify the status differences between flight crew. They found that the use of first person plural (we, our, us) pronouns increases over the life of a flight crew, and captains speak more in the first person plural than others. Locher (2004) connects Wartenberg (1990)'s notion of action-restriction with the study of politeness in dialogs and identifies linguistic means to restrict interactants' action environments. Pennebaker (2011) analyzed interactions between undergraduates, graduate students, and faculty and found that their relative use of first person singular pronouns was a strong marker for identifying their relative status in an interaction.

Most of the previous work just discussed was conducted entirely on spoken dialog. In our work, we analyze both written and spoken interactions and we show that the core insight — conversation is a resource for power — carries over to written dialog as well, and that computational techniques can benefit from looking into the discourse structure. However, some of the characteristics of spoken dialog do not carry over straightforwardly to written dialog, most prominently among them the important issue of interruptions: there is no interruption in written dialog. Our work draws on findings for spoken dialog, looking at correlates for written dialog.

## 2.2 Computational Analysis of Organizational Interactions

In this section, we review computational studies in the field of analyzing organizational interactions, one of the two genres we study in this thesis. A majority of early computational work on analyzing manifestations of power in interactions was done on organizational email. This was largely due to the availability of Enron email corpus, which is a large collection of emails from the Enron Corporation, collected and released by the Federal Energy Regulatory Commission as part of its investigation after the company's collapse in 2001. This corpus was the first large publicly-available collection of real email messages, which spurred research interest from many different disciplines. We also use the Enron email corpus in our thesis. We start with a brief history of the Enron email corpus, before we describe different computational studies performed using the corpus.

### 2.2.1 A Brief History of Enron Email Corpus

The FERC released the Enron email corpus on the web in May 2002. The corpus contained around 600K emails that were from the mailboxes of 158 Enron employees at the top level. The email data included the information about the sender, the set of recipients, date, time, subject, and the body of the email. The attachments of the emails were not included in the initial release. After the initial release, various researchers noticed many integrity issues in the corpus. Subsequently, the corpus underwent many iterations of cleaning up and reformatting, which resulted in many different versions of the corpus. Klimt and Yang (2004) performed the first major iteration of cleaning up and fixing some data integrity issues. Shetty and Adibi (2004) performed further cleaning up and released a MySQL version of the corpus, to which Diesner and Carley (2005) added the position

and location information. In a separate line of work, using the original FERC release, Yeh and Harnly (2006) automatically assembled the thread information of the emails in the corpus, to which (Agarwal et al. 2012) added organizational hierarchy information. There is also a more recent version of the corpus released by EDRM that is further cleaned up to remove personal information from the email content,[1] and is used by the TREC (Tomlinson 2010) for evaluation in the legal track. We now discuss in detail the different cleaned up and/or extended versions of the corpus.

The first cleaned up version of Enron email corpus was released by Klimt and Yang (2004). The released raw corpus contained 619,446 messages belonging to 158 users organized as separate folders.[2] They continue to delete messages from the released corpus "as part of a redaction effort due to requests from affected employees". Along with the corpus release, (Klimt and Yang 2004) also presented some experiments on automatically categorizing emails to corresponding folders. For those experiments, they used a cleaned up version of the corpus, in which they removed certain folders that seemed to be computer generated. They also removed the folder named "all documents" from all user mailboxes since they contained duplicate email messages. Their cleaned up version of the corpus contained 200,399 messages belonging to 158 users, approximately one third of the original corpus in terms of size. They also presented some initial work on detecting the email threads automatically. Their approach for thread detection relied on two factors: "Emails were considered to be in the same thread if they contained the same words in their subjects and they were among the same users (addresses). Messages with empty subjects were not considered to be a thread." However, they did not perform any evaluation of this approach.

Shetty and Adibi (2004) released a cleaned up version of the corpus in the form of a MySQL database. They used the dataset by Klimt and Yang (2004) as their starting point, but cleaned the dataset by removing duplicate emails, junk data, blank emails, as well as returned emails failure reports. They removed auto-generated folders such as "discussion threads", "all documents" (that Klimt and Yang (2004) had already found to contain duplicate emails), and the duplicate emails in the "sent messages" folder. They also fixed some invalid email addresses. Their cleaned up Enron email dataset contained 252,759 messages from 151 employees distributed in around 3000 user defined folders.

[1] `http://www.edrm.net/resources/data-sets/edrm-enron-email-data-set`

[2] `http://www-2.cs.cmu.edu/~enron/`

Diesner and Carley (2005) enhanced and refined the version released by Shetty and Adibi (2004) by adding information about positions (identified by titles) held by a subset of employees and their geographic location. They used three existing resources — a file with the positions of former employees released by ISI, a file with with job information from FERC, and a list from FERC with information on people's location — to assemble this information. They identified 15 unique job titles that were associated with 212 employees, and 5 unique locations of 67 people. They also performed additional normalization on email addresses of the individuals they were able to identify either the position or location.

In a parallel effort, Yeh and Harnly (2006) built a version of the corpus in which the thread structure of email messages is automatically reassembled. They consider thread reassembly as "the task of relating messages by parent-child relationships, grouping messages together based on which messages are replies to which others." They use similarity based measures to group emails together taking into account various heuristics such as subject, time, and sender/recipient information of emails. They also recover emails that are missing from the corpus by extracting them from other emails in which they are quoted. They also resolved multiple email addresses belonging to the same person, and assigned unique identifiers and aggregated names from different sources to persons. Therefore, each person is associated with a set of email addresses and names (or name variants), but has only one unique identifier. This version of the Enron corpus was further enhanced by Agarwal et al. (2012). They added organizational hierarchy based dominance relations into corpus. They extract this hierarchy information manually by reviewing the original Enron organizational charts. We use the version of the Enron corpus released by Yeh and Harnly (2006) along with the enhancement by Agarwal et al. (2012) for the analysis presented in this thesis. We describe them both in more detail in Chapter 3, Section 3.1.

### 2.2.2 Computational Analysis of Enron Email Corpus

Many researchers have applied social network analysis (SNA) on the Enron email corpus to study how the crisis Enron was going through affected the internal communication patterns (e.g., (Diesner and Carley 2005, Chapanond et al. 2005, Murshed et al. 2007)). For example, Diesner and Carley (2005) used the position information they added to the corpus to study the organizational behavior patterns in Enron before and during the turmoil in Enron. They found that "during the crisis the

communication among Enron employees had been more diverse with respect to people's formal positions and that the top executives had formed a tight clique with mutual support and highly brokered interactions with the rest of organization." Chapanond et al. (2005) used graph theoretical analysis to study various graph metrics such as degree distribution, centrality measures, average distance ratio, clustering coefficient and so on. Murshed et al. (2007) also found evidence for high levels of clique activity in response to the enveloping crisis.

The work described above mostly looked at the communication patterns as captured by the meta data (e.g., who sent how many messages to whom?). Researchers have also looked at the language used in emails. One line of analysis focuses on the linguistic patterns in emails related to the fall of Enron. Louwerse et al. (2010) investigated whether fraudulent events can be related to linguistic cues of deception in the emails dataset using a model of interpersonal language use. They found that "during times of fraud, emails were composed with higher degrees of abstractness".

Researchers have also applied linguistic analysis to reveal insights about organizational interactions in general. Keila and Skillicorn (2005) studied the Enron emails using word frequency profiles and length of messages and found that word use of individuals correlated with their function within the organization and that the relative changes in individuals' word usage over time can be used to identify key players in the organization. McCallum et al. (2007) applied their Author-Recipient-Topic (ART) model, which learns topic distributions based on the direction-sensitive messages sent between entities, on the Enron email corpus. They found that their topic model could predict people's roles in addition to detecting relevant topics. (Peterson et al. 2011) studied email formality in workplace using the Enron corpus. They build a formality tagger that they then apply to the corpus to study how the level of formality aligns with social distance, relative power, and weight of imposition. They found, among other things, that the more emails exchanged between a pair of people, the more informal their conversations were. Mohammad and Yang (2011) analyzed the way gender affects expressions of emotions in Enron emails and found that women send and receive emails with relatively more words that denote joy and sadness, whereas men send and receive relatively more words that denote trust and fear.

There is also work on analyzing manifestations of power in Enron using both social network analysis (SNA) techniques as well as NLP techniques. Early work used SNA based approaches (Diesner and Carley 2005, Shetty and Adibi 2005, Creamer et al. 2009) or email traffic patterns

(Namata et al. 2007) for extracting power relations. These studies use only meta-data about messages: who sent a message to whom when. For example, Creamer et al. (2009) find that the response time is an indicator of hierarchical relations; however, they calculate the response time based only on the meta-data, and do not have access to information such as thread structure or message content, which would actually verify that the second email is in fact a response to the first.

In fact, using NLP to deduce social relations from online communication is a relatively new area which has only recently become an active area of research. Bramsen et al. (2011) and Gilbert (2012) are two prominent studies which applied NLP based techniques to predict power relations in Enron emails. Using knowledge of the actual organizational structure, Bramsen et al. (2011) create two sets of messages: messages sent from a superior to a subordinate, and *vice versa*. Their task is to determine the direction of power (since all their data, by design in the construction of the corpus, has a power relationship). They approach the task as a text classification problem and build a classifier to determine whether the set of all emails (regardless of thread) between two participants is an instance of up-speak or down-speak. Similarly, Gilbert (2012) considers a message to be *upward* only when every recipient of that message outranks the sender. Any message that is not an *upward* message is labeled *non-upward*. This formulation is slightly different from that of (Bramsen et al. 2011) which considers only those messages that have a power relationship, upward or downward. Gilbert (2012) extracts a list of phrases that signal *upward* messages using penalized logistic regression model.

While the objectives of both these studies and our work are the same, there are major differences. We also use lexical features in our study and find them very useful, however, our focus is on understanding how the dialog structure and other deeper linguistic patterns (such as overt display of power, expressions of beliefs etc.) correlate with power. Consequently, our data unit is a naturally occurring thread, not data units assembled by researchers. Using email threads as our data units also allows us to focus on the structure of interactions, which wouldn't have been the case with a single message or an arbitrary aggregation of single messages.

## 2.3 Computational Analysis of Political Speech

In this section, we summarize recent work on computationally analyzing political speech. There is a growing body of research applying linguistic analysis to political discourse (Thomas et al. 2006,

Cardie and Wilkerson 2008, Guerini et al. 2008, Rosenberg and Hirschberg 2009, Strapparava et al. 2010, Nguyen et al. 2013, Sim et al. 2013, Iyyer et al. 2014). Researchers have looked at a variety of applications such as identifying markers of persuasion (Guerini et al. 2008, Strapparava et al. 2010), predicting voting patterns (Thomas et al. 2006, Gerrish and Blei 2011), and detecting ideological positions (Sim et al. 2013, Iyyer et al. 2014), to state a few. We summarize some of this work below.

Strapparava et al. (2010) uses CORPS corpus of political speeches released in (Guerini et al. 2008) to predict persuasiveness in political discourse. They conducted experiments using lexical features to predict persuasive passages in the discourses that trigger a positive audience reactions. Studies have also analyzed how personal attributes of political personalities. Rosenberg and Hirschberg (2009) analyze speeches made in the context of 2004 Democratic presidential primary election and identify lexical and prosodic cues that signal charisma. More recently, Nguyen et al. (2013) analyze the 2008 presidential and vice presidential debates to study how speaker identification helps topic segmentation and how candidates exercise control over conversations by shifting topics. In this thesis, we also study topic shift behavior by candidates in presidential debates, however, our focus is to correlate the dialog behavior with external factors such a poll scores. Iyyer et al. (2014) apply recursive neural networks to political ideology detection and shows that their approach detects bias more accurately than existing methods which uses bag-of-words models and hand-designed lexical resources. We do not use neural network based methods in our work, but we have identified it as one of the directions to take our work further in future.

## 2.4 Computational Power Analysis on Other Genres

Within the dialog community, researchers have studied notions of control and initiative in dialogs. Walker and Whittaker (1990) define "control of communication" in terms of whether the discourse participants are providing new, unsolicited information. They use utterance level rules to determine which discourse participant (whether the speaker or the hearer) is in control, and extend it to segments of discourse. One of the types of power we study is also the power or control over communication. However, their notion of control differs from our notion of power over communication. They model control locally over discourse segments. What we study is the possession of controlling power by one (or more) participant(s) across the entire dialog, i.e. how a participant controls the

communication in a dialog thread in order to achieve its intended goals. Despite this difference in definition, we find in our study that our notion of power over communication correlates with Walker and Whittaker (1990)'s notion of control over discourse segments. Jordan and Di Eugenio (1997) suggest that "initiative" applies to the level of problem solving, just as "control" applies to the dialog. Our notion of situational power is closely related to this notion.

More recently, there has been substantial research in analyzing manifestations of power in online written interactions (Strzalkowski et al. 2010, Danescu-Niculescu-Mizil et al. 2012, Biran et al. 2012, Swayamdipta and Rambow 2012, Bracewell et al. 2012, Taylor et al. 2012). Wikipedia talk pages is an online genre that has seen the most interest in the study of power, since it has established power structures such as administrators/moderators. It is also the case that Wikipedia discussions are mostly task oriented, a contrast from most other online discussion genre.

Danescu-Niculescu-Mizil et al. (2012) study the notion of language coordination — a metric that measures the extent to which a discourse participant adopts another's language — in relation with various social attributes such as power, gender, etc. They perform their study on Wikipedia discussion forums and Supreme Court hearings. They also look into situational power; however they define situational power in terms of the dependence between interactants: "x may have power over y in a given situation because y needs something that x can choose to provide or not". They model this dependence "using the exchange-theoretic principle that the need to convince someone who disagrees with you creates a form of dependence." We adopt a broader definition of situational power in our work based on context and perception. They study how power affects language coordination — a metric that measures the extend to which y adopts x's language, while we focus primarily on the structure of the dialog.

Strzalkowski et al. (2010) and Taylor et al. (2012) are also interested in power in written dialog. However, their work concentrates on lower-level constructs called *Language Uses* which will be used to predict power in subsequent work. This said, one of their language uses is agenda control, which is very close to our notion of power over communication. They model power using notions of topic switching, exploiting mainly complex lexical features. Biran et al. (2012) use content-related dialog behavior such as attempts to persuade and agreement/disagreement, and discourse structure-related dialog behavior such as initiative and investment, in order to find influencers in Wikipedia discussion forums and LiveJournal blogs.

Bracewell et al. (2012) and Swayamdipta and Rambow (2012) try to identify participants pursuing power in discussion forums. Bracewell et al. (2012) devise a set of eight *social acts* which largely overlaps with the dialog constructs used by (Biran et al. 2012). Swayamdipta and Rambow (2012) on the other hand, adopt an unsupervised learning approach and obtained results at par with a supervised models. They also use many dialog structure features to build their model, and find lexical features to be not helpful. Our work also falls into this category of studies in the sense that we look beyond purely lexical features.

# Part I

# DATA AND METHODS

# Chapter 3

# Data

In this chapter, we describe the different datasets we use in the work presented in this thesis. We describe in detail the source of the data as well as preexisting annotations on it that we make use of. In addition, we also summarize the new annotations/extensions that are contributions of this thesis, and give pointers to the chapters that describe them in more detail. We start by describing the datasets and annotations we use for our study in the domain of organizational email in Section 3.1. We then discuss the data and resources we use for the domain of political debates in Section 3.2. In Section 3.3, we describe the other datasets we use in this thesis, before we summarize the chapter in Section 3.4.

## 3.1 Organizational Emails

In this section, we describe the dataset and different annotations we use for the analysis we perform in the domain of organizational email.

### 3.1.1 Data Source

We use the version of Enron email corpus built by Yeh and Harnly (2006) for our study. They start with the the original collection of email messages released by the FERC which contained emails from the 158 mailboxes, which they assess to be owned by 149 people. Like the other cleaned up versions of the corpus (Klimt and Yang 2004, Shetty and Adibi 2004), they also removed auto-generated folders such as “all documents”, “discussion threads”, “contacts” etc. from each mailbox.

In addition, they also used some heuristics to eliminate "Exchange-specific" files from the folders that were not email messages and grouped duplicate messages. They report that this process resulted in a corpus of 269,257 unique messages; an average of 1,704 messages per mailbox. They also found that a large number of emails belonged to a small group of users; around 35% of messages were from the 10 largest mailboxes.

#### 3.1.1.1 Approach 1: Using Microsoft's Exchange Header

They used the header field called "Thread-Index" defined in the Microsoft Exchange Protocol that associates multiple emails to an email thread. This is a high precision method to identify the parent-child relations between emails (i.e., it never makes a false positive). However the Thread-Index header is not always available and hence the coverage of this approach is very low.

#### 3.1.1.2 Approach 2: Similarity Matching and Heuristics

They used a similarity matching algorithm along with some heuristics in order to reconstruct the thread structure of emails interactions for cases where the Approach 1 fails. We describe the steps they took below.

**Preprocessing steps** They applied the following preprocessing steps to the set of emails to normalize the meta data and separate the email body into original content and quotations.

- **Duplicate message grouping**: They grouped together the same email messages existing in different mailboxes (e.g., an email from A to B will be in A's *Sent* folder and B's *Inbox* folder) by matching date and time, subject, message body, and From/To/Cc/Bcc headers.
- **Datetime normalization**: They convert the time-stamp of each message into a corresponding time-stamp in the same time zone in order to enable easy comparison
- **Subject normalization**: They remove common prefixes and suffixes, such as RE:, FW:, FWD:, etc. from the email subject line.
- **Sender/ recipient identification and normalization**: They identify email addresses that likely belong to the same individual. For this step, they used the following heuristics:

- if the same email contains an email address in the RFC 2822 'From' header, and the another one in the Microsoft-specific 'ExchangeFrom' header, then both addresses are assumed to be of the same individual
- if there are multiple email addresses that are in the 'From' header of emails in a sent-mail folder, they are all considered to be of the same individual.
- if two email addresses are labeled with the same name in emails that have all other participants the same, both addresses are considered to be of the same individual; i.e, two people may have the same name, but it is unlikely for them to be interacting with exactly the same set of people.

5) **Reply and quotation extraction**: They separated the reply and quotation parts of emails using a set of 25 splitter texts (e.g., "—-Original Message—-"). They report that their approach correctly separated 98% of 1000 randomly selected emails.

**Finding the Parents:** They applied a similarity matching algorithm to the preprocessed emails to find the parent of each email, if one exist. The algorithm takes as input a set of email messages and outputs a set of email threads. The algorithm is summarized below:

1. Sort all emails in chronological order
2. Consider each message $m$ as an initial thread $T$, and put all messages that fall within a time-window (they set this to 14 days) and have the same normalized subject line as $m$ into the set $M$, the candidate children.
3. Add each email $m_i \in M$ to $T$, if $T$ already contained a possible parent of $m_i$. Return to step 2 until all emails are processed. To find the parent of $m_i$ in $T$, they used a series of tests comparing their sender/recipient relationships as well as similarity between the top level quotations of $m_i$ and the reply part of the candidate parent emails.

**Finding missing messages:** They applied the similarity matching to the automatically detected quotation text fragments to obtain sequences of missing messages that were not originally present in the corpus.

#### 3.1.1.3 Corpus Statistics

Using Approach 1, they identified 3705 email threads. The email threads obtained from Approach 1 is a reliable set of email threads which they use to evaluate Approach 2. They report a high recall of 87.4% in identifying the parent/child relationships captured by Approach 1. Table 3.1 shows the number of email threads reconstructed using both approaches.

| | Number of threads |
|---|---|
| Approach 1 | 3,705 |
| Approach 2 | 32,910 |
| Total | 36,615 |

Table 3.1: Enron email corpus statistics: number of threads (Yeh and Harnly 2006).

Using Approach 2, they obtained 32,910 email threads, which consist of 95,259 unique messages. Table 3.2 shows the distribution of email threads with respect to thread sizes. The mean thread size is 3.14, with a mean depth of 1.71. The median thread size is 2. The total number of threads with 2 to 5 messages is 30,940; only 1,970 have more than five. Hence, the corpus contains a large number of small threads.

| Thread Size | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11-20 | 20+ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| # of threads | 19,941 | 6,753 | 2,868 | 1,378 | 770 | 406 | 241 | 170 | 121 | 221 | 41 |

Table 3.2: Enron email corpus statistics: distribution of email thread sizes (Yeh and Harnly 2006).

### 3.1.2 Existing Annotations

In this section, we describe the different annotations or extensions added to the corpus by other researchers, that we make use of in this thesis.

#### 3.1.2.1 Dialog Act Annotations (Hu et al. 2009)

A small subset of 122 email thread from the corpus by Yeh and Harnly (2006) described in Section 3.1.1 was annotated by Hu et al. (2009) for dialog acts. Their unit of annotation was Dialog Functional Units (DFU) which represent abstract units of interaction. They derive the notion of DFU from previous work in intention-based segmentation (Passonneau and Litman 1997) and on mixing formal schemas with natural language descriptions (Nenkova et al. 2007). They annotate each DFU with an extent, a dialogue act (DA) label along with a description, and possibly one or more forward and/or backward links. The extent of a DFU roughly corresponds to that portion of a turn (conversational turn; email message; etc.) that corresponds to a coherent communicative intention. They capture the communicative function of a DFU by assigning one of following seven dialog acts:

- Inform: This DFU conveys information. This covers many different types of information that can be conveyed, including answers to questions, elaborations, reporting completion of a requested action and so on.
- Commit: This DFU commits the speaker/writer to performing a task.
- Request-Information: This DFU obliges the hearer/reader, or opens an option to the hearer/reader, to provide information (either facts or opinion), either in the dialog or through another form of communication.
- Request-Action: This DFU obliges the hearer/reader, or opens an option to the hearer/reader, to perform some non-communicative action, i.e., an action that cannot be part of the dialog.
- Conventional: These are greeting, introductions, expression of thanks, etc.

In addition to the dialog act labels, the annotations also capture links between DFUs. They annotate three kinds of links:

- Forward link (Flink): a DFU is annotated with a forward link if it sets up an expectation in the dialog that the reader/hearer perform a certain action

- Backward link (Blink): a DFU is annotated with a backward link if it relates to a previous DFU, by performing an action which responds in some sense to the previous DFU. A DFU can have both a flink and a blink.

- Secondary forward link (SFlink): If a backward link connects back to a DFU that does not contain a forward link, then the it will be annotated with a secondary forward link.

We describe more statistics of these annotations in Chapter 5 in which we use this data to build an automatic dialog act tagger.

#### 3.1.2.2 Gold Standard for Enron Organizational Hierarchy (Agarwal et al. 2012)

We use the organizational hierarchy relations Agarwal et al. (2012) added to the corpus. They collected this information by studying the original Enron organizational charts. They discovered these charts by performing a manual, random survey of a few hundred emails, looking for explicit indications of hierarchy. They initially found that organizational charts are often present as Excel or Visio files. Hence they searched all remaining emails for attachments of the Excel or Visio files, and examined those with additional organizational charts. Then they manually transcribed the information contained in all organizational charts.

They define a dominance relation to be the relation between superior and subordinate in the hierarchy. Their gold standard for hierarchy relations contains a total of 1518 employees. They found 2155 immediate dominance relations spread over 65 levels of dominance (CEO, manager, trader etc.) among these 1518 employees. In the next step, they obtained the transitive closure of these relations. That is, they obtained the set of all valid organizational dominance relations. If an employee A immediately dominates another employee B and if B immediately dominates another employee C, then the set of valid organizational dominance relations are A dominates B, B dominates C and A dominates C. This step obtained 13,724 dominance relations, which forms the gold standard of organizational hierarchy they released. This data set is much larger than any other data set used in the literature for predicting organizational hierarchy.

### 3.1.3 New Annotations

In this section, we briefly describe the extensions or annotations added to the Enron email corpus as part of this thesis.

#### 3.1.3.1 Overt Display of Power Annotations

As part of this thesis, we built a corpus of annotated with instances of overt display of power at an utterance level. We obtained these annotations on the same corpus that has dialog act annotations (Section 3.1.2.1). The details of overt display of power annotations are described in Chapter 6.

#### 3.1.3.2 Power Types Annotations

We also built a corpus of email threads annotated with different types of power relations between participants. The annotations capture instances of situational power, influence, power over communication, as well as perceived hierarchical power. These annotations were also obtained on the same corpus that contained dialog act annotations. The details of these annotations are described in detail in Chapter 10.

#### 3.1.3.3 Gender Identified Enron Corpus

We released an extension to the entire corpus in which we assigned the gender of authors of 87% of the emails. The procedure followed to perform the gender assignment is described in detail in Chapter 8.

### 3.1.4 Corpus Subdivisions

Our starting point is the email corpus released by (Yeh and Harnly 2006) that we described in detail in Section 3.1.1. We use different subsets of this corpus for different analyses presented in this thesis, depending on the information required for each analysis. We formally define each such subset below.

- ENRON-SMALL: A subset of 122 email threads from the original corpus that contains the dialog act annotations (Hu et al. 2009), overt display of power annotations (Chapter 6) as

well as the power types annotations (Chapter 10). This is the corpus that is used in Chapter 5, Chapter 6, and Chapter 10.

- ENRON-EXCLUDE: Another subset of 297 email threads that were used for a re-annotation effort for dialog acts in order to capture the annotations in a slightly different granularity, using a modified annotation manual from (Hu et al. 2009). However we did not utilize these annotations in this thesis.

- ENRON-LARGE: This subset contains all the email threads that are not part of either ENRON-SMALL or ENRON-EXCLUDE. We use this corpus for our analysis presented in Chapter 7 and Chapter 9.

- ENRON-APGI: This is a subset of ENRON-LARGE that contains the set of email threads with all participants' gender identified as per the gender assignment procedure described in Chapter 8. This subset is used in the analysis presented in that chapter.

We list the number of threads in each thread in Table 3.3.

| | Number of threads |
|---|---|
| Total (Yeh and Harnly 2006) | 36,615 |
| ENRON-SMALL | 122 |
| ENRON-EXCLUDE | 297 |
| ENRON-LARGE | 36,196 |
| ENRON-APGI | 17,788 |

Table 3.3: Enron email corpus statistics: number of threads in corpus subdivisions.

## 3.2 Political Debates

In this section, we describe the data we use for our study on political debates.

| | |
|---|---|
| Number of debates | 20 |
| Interaction time | 30-40 hrs |
| Average number of Candidates per debate | 6.6 |
| Average number of Turns per debate | 245.2 |
| Average number of Words per debate | 20466.6 |

Table 3.4: GOP debates corpus statistics.

### 3.2.1 Data Source

We obtain the manual transcripts of presidential debates that are collected as part of the The American Presidency Project.[1] The American Presidency Project is a collaboration between John T. Woolley and Gerhard Peters at the University of California, Santa Barbara (Woolley and Peters 2011). Their archives contain a large collection of documents related to the study of the American Presidency. They have consolidated, coded, and organized this information into a single searchable online resource that contain documents such as party platforms, election debates, candidates' remarks and speeches, voter turnouts, acceptance speeches, inaugural addresses, President's approval ratings, State of the Union addresses and so on.

We downloaded the transcripts of the 2012 Republican Party primary debates from the The American Presidency Project website.[2] The transcripts are manually coded. Each debate's transcript lists the presidential candidates who participated and the moderator(s) of the debate. Transcripts demarcate speaker turns and also contain markups to denote applause, laughter, booing and crosstalk during the debates. We preprocessed the transcripts to avoid minor formatting errors and unified them into an XML format. The transcript of all debates follow similar formats, except for a few exceptions (e.g., the format of listing participants was different for few debates), which we manually corrected during the conversion to XML. Table 3.4 shows various statistics on the debates.

[1] `americanpresidency.org`

[2] http://www.presidency.ucsb.edu/debates.php

### 3.2.2 Candidate Poll Standings

We used Wikipedia as a source for obtaining candidate poll standings during the course of the 2012 presidential primary election campaign. Two Wikipedia pages kept track of opinion polls from a variety of sources during the campaign, one for the statewide polls,[3] and the other for the national polls.[4] We use both of these resources in our analysis. The sources of opinion polls include Gallup, Pew Research, Public Policy Polling as well as various national and regional news agencies such as CNN, Fox News, CBS etc.

The poll results are listed in the Wikipedia pages within a table environment under an html element of class: 'wikitable'. This made it easy to parse the html pages to obtain the poll scores, enabling us to easily obtain poll results from a wide variety of sources in one step. The parse of the Wikipedia page listing the state poll results returned the poll scores from 45 states from 153 different polling sources, ranging from polls conducted as far in the past as April 27th of 2009 till May 15th of 2012. We do not use all these poll scores in our analysis. We describe in Chapter 11 (Section 11.2.2, page 223) in detail which ones we use these scores.

## 3.3 Other Datasets

### 3.3.1 LU Corpus Annotations

We use the LU Corpus annotations (Diab et al. 2009) that capture whether a speaker/writer (SW) intends the reader to interpret a stated proposition as the writer's strongly held belief, as a proposition which the writer does not believe strongly (but could), or as a proposition towards which the writer does not express a belief, but rather a different cognitive attitude, such as desire or intention. We describe this corpus in more detail in Section 9.2.1, page 161.

[3] http://en.wikipedia.org/wiki/Statewide_opinion_polling_for_the_Republican_Party_presidential_primaries,_2012

[4] http://en.wikipedia.org/wiki/Nationwide_opinion_polling_for_the_Republican_Party_2012_presidential_primaries

| Corpus | Annotations | Chapters |
|---|---|---|
| ENRON-LARGE | Organizational Power | Chapters 7 and 9 |
| ENRON-SMALL | Organizational Power<br>Types of Power<br>Dialog Acts<br>Overt Displays of Power | Chapters 5, 6 and 10 |
| ENRON-APGI | Organizational Power<br>Gender | Chapters 8 |
| DEBATES-2012 | Poll Scores | Chapters 11 and 12 |

Table 3.5: Summary of different datasets used in this thesis.

### 3.3.2 DEFT Corpus Annotations

We use the DEFT Corpus that extends the 3-way belief distinction in LU Corpus to a 4-way scheme. The DEFT Corpus annotations capture whether a speaker/writer (SW) intends the reader to interpret a stated proposition as the writer's strongly held belief, as a proposition which the writer does not believe strongly (but could), as a proposition the writer is reporting someone else's belief about, or as a proposition towards which the writer does not express a belief, but rather a different cognitive attitude, such as desire or intention. We describe this corpus in more detail in Section 9.2.2, page 162

## 3.4 Summary

In Table 3.5, we summarize the different corpora, the annotations present in them and the chapters they are used in for analysis.

# Chapter 4

# Methods

In this section, we describe the methods used in this thesis. We start by describing the general software framework that we use to build the different systems of analysis. We then describe the natural language processing techniques that we apply as basic preprocessing steps for our analysis. After that we discuss the different feature representations we use across the different analyses presented. Finally, we explain the machine learning algorithms and some of the specific issues associated with them that the tackle in this thesis.

## 4.1 Software Framework

**UIMA:** We use the UIMA framework to build the complex natural language processing systems and perform the analysis and experiments described in this thesis. UIMA stands for Unstructured Information Management Architecture and was originally developed by IBM (Ferrucci and Lally 2004) and was later released under the Apache open source license.[1] UIMA provides a framework that facilitates analysis of large amounts of unstructured context such as text, audio and video.

UIMA enables applications to be decomposed into components, each of which implements interfaces defined by the framework and provides self-describing meta data via XML descriptor files. The framework gives the user an object called a common analysis structure (called CAS, for short) to which different components can add their analysis output to. UIMA provides a way for the user to specify how the analysis output of a component be represented in the CAS, and leaves

[1] https://uima.apache.org/

the flexibility of how to implement the analysis to the user. For example, a typical natural language processing system would start with a CAS that contain its input text, passing through a tokenizer component which adds token annotations to the CAS, followed by a part-of-speech tagger that adds part-of-speech tags to the token annotations, followed by a parser and so on. UIMA provides just the framework, but not the implementations of any of these components.

**ClearTK:** ClearTK (Ogren et al. 2008) is a a suite of tools and wrappers that provides easy access to many state-of-the-art machine learning and natural language processing components that are seamlessly integrated with the UIMA architecture.[2] It provides UIMA wrappers for common NLP tools; OpenNLP tools, MaltParser dependency parser and Stanford CoreNLP to name a few. It also provides a common interface and wrappers for popular machine learning libraries such as SVMlight, LIBSVM, OpenNLP MaxEnt, and Mallet. Another advantage of ClearTK is its rich feature extraction library that can be used with any of the machine learning classifiers. We use ClearTK to build our machine learning models as well as to perform the basic NLP preprocessing steps. We use some of the built-in feature extractors; however, for most of our features we wrote new feature extractors, which are easy to integrate with the ClearTK-UIMA analysis framework.

## 4.2 NLP Preprocessing

In this thesis, we utilize the analysis produced by basic NLP steps, namely, tokenization, part-of-speech tagging, lemmatization and parsing (in some cases) as the building blocks. We describe the tools we use, in this section.

- **Tokenization** is the task of splitting running text into pieces of text called *tokens*. A *token* is an instance of a sequence of characters that are grouped together as a useful orthographic unit for processing. Tokenization is a relatively easy task in English in which words are separated by white spaces and punctuations. But not all punctuations separate tokens (e.g., "U.S." should be considered as one single token). Machine learning models have been built that do the job of accurately splitting text into tokens. We use the OpenNLP tokenizer that

[2] `http://cleartk.github.io/cleartk/`

comes as the default tokenizer with the ClearTK suite of tools for our analysis, except in Chapter 9 where we use the Stanford CoreNLP tokenizer.

- **Sentence Segmentation** is the task of splitting sequences of tokens into sentences. We use the OpenNLP sentence splitter that comes as the default sentence splitter with ClearTK, except in Chapter 9 where we use the Stanford CoreNLP system.

- **Part-of-speech tagging** is the process of marking up the tokens identified in text as corresponding to a particular part of speech (e.g., noun, verb, adjective), based on the context in which it is used. The state of the art systems for part-of-speech tagging report accuracies above 97% on identifying the correct part-of-speech tags. We use the OpenNLP part-of-speech tagger that comes as the default ClearTK part-of-speech tagger, except in Chapter 9 where we use the Stanford CoreNLP part of speech tagger.

- **Dependency parsing** is the step that analyzes the grammatical structure of a sentence, establishing dependency relationships between word tokens in a sentence as a tree structure. In the dependency tree, each word token becomes a node and each edge between tokens is labeled with a dependency relation (e.g., subject, object). We use the Stanford CoreNLP system to obtain the dependency parses for ou analysis. We use the dependency parse information only in Chapter 9.

## 4.3 Machine Learning Algorithms

As part of this thesis, we build six machine learning systems — overt display of power tagger, dialog act tagger, committed belief tagger, direction of power predictor, person with power predictor, and power ranker. In all six cases, we adopt a supervised learning framework, where we supply the machine learning algorithm with a set of labeled *instances* and let the algorithm learn to classify the labels on unseen instances. The instances the algorithm learns from are called *training instances* and we call the unseen instances that we test the system during experiments *test instances*. The instances are usually represented as *feature vectors*, i.e., points in a high dimensional feature space. This representation is essentially a mapping of the real world problem instance (like classifying power relations between *Sara* and *Kim*) to a geometric space that encapsulates the features of the

real world instance that might help solve the problem (of classification).

### 4.3.1 Binary Support Vector Machines

Three of the systems we build solve binary classification problems — classifying a sentence to be an overt display of power or not, classifying whether the first person of a pair of participants is the superior of the other, and classifying whether a participant has a certain kind of power or not. In all three cases, we use the Support Vector Machine (SVM) algorithm to build our systems.

A Support Vector Machine is a discriminative classifier that finds a maximum-margin hyperplane that optimally separates the instances in the high-dimensional feature space so that it can classify unseen instances. The binary classification problem is the simplest formulation of a classification problem where the training instances will be labeled as positive and negative instances. SVMs learn a decision function $f$ from the set of positive and negative training instances such that an unlabeled instance $x$ is labeled as positive if $f(x) > 0$. This function $f$ represents the maximum-margin hyperplane that separates the positive and negative instances. We use the ClearTK wrapper for the SVMlight (Joachims 1999) package in our experiments.

### 4.3.2 Multi-class Support Vector Machines

Two of our systems solve multi-class classification problems — classifying the dialog act of a segment of text in an interaction to be one of the four dialog acts, classifying words in a sentence to be propositional heads of one of the types of belief expressions or not a propositional head. The basic SVM formulation deals with only binary classification tasks. A commonly used extension to apply SVMs to multi-class situations is by using a one-vs-all algorithm, where separate models are trained to recognize each of the classes separately using the binary SVM formulation and then at prediction time, assigning the label based on the predictions (and their confidences) made bye each individual model.

We use this one-vs-all method to build our model for belief tagging in Chapter 9. We use the ClearTK wrapper for the SVMlight (Joachims 1999) package which internally implements the one-vs-all approach. As part of our dialog act tagging experiments presented in Chapter 5, we introduced a new multi-class classification algorithm that outperforms the one-vs-all method.

### 4.3.3 Support Vector Ranking

The basic SVMlight implementation also perform ranking. There is also the SVMrank package which is a faster implementation of the ranking algorithm. The SVMrank algorithm solves the quadratic program through an extension to the ROC-are optimization algorithm (Joachims 2006). We use the ClearTK wrapper for the SVMrank for our experiments.

### 4.3.4 Handling Class Imbalance

Since SVMs optimize on training set accuracy to learn the decision function $f(x)$, it performs better on balanced training sets (i.e., equal number of instances for each class/label). As a result, in situations where the dataset is imbalanced, SVMs perform poorly. We use a threshold adjusting method to handle this issue across our experiments. We find a better threshold for $f(x)$ based on posterior probabilistic scores, $p = Pr(y = 1|x)$, calculated using the ClearTK implementation of Lin et al. (2007)'s algorithm. It uses Platt (1999)'s approximation of $p$ to a sigmoid function $P_{A,B}(f) = (1 + exp(Af + B))^{-1}$, where A and B are estimated from the training set. Then, we predict $x$ as positive if $p > 0.5$, which in effect shifts the threshold for $f(x)$ to a value based on its distribution on positive and negative training instances.

Another commonly used method is instance weighting, where training errors on majority class instances are outweighed by errors on minority class instances. This can be achieved using the *j* option in SVMlight to set the outweighing factor. This is in effect equivalent to oversampling by repeating minority class instances. In Chapter 6 we experiment with this approach, setting the outweighing factor to be the ratio of negative to positive instances in the training set and show that the threshold fitting approach described above works better than instance weighting approach.

## 4.4 Feature Representations

We use lexical features in all our experiments. We describe how we represent the lexical features here. Other features are described in detail in the respective chapters. This set of features include ngram features that can be extracted from the lemma and part-of-speech tags of the tokens in a span of text. An ngram is a contiguous sequence of $n$ items from the span of text, where an item could be words, word lemmas, or part-of-speech tags. In this thesis, when we say "ngrams with $n = 2$",

we mean that the feature set includes indicator features of all ngrams of length 1 to 3. We use three types of ngram features — word lemma ngrams, part-of-speech ngrams, and mixed ngrams. We describe each of type of ngrams below using the example sentence "I took the report"

**LemmaNgram:** contiguous sequences of word lemmas of length $n$ or smaller. For example, LemmaNgram ($n = 3$) of the above sentence will contain the indicator features for the following ngrams set to 1: *i*, *take*, *the*, *report*, *_BOS_ i*, *i take*, *take the*, *the report*, *report _EOS_*, *_BOS_ i take*, *i take the*, *take the report*, and *the report _EOS_*. Note that we use word lemmas ("take") instead of surface forms ("took") in our experiments. We did use ngrams of surface form words in our preliminary experiments, but found lemma ngrams to perform consistently better and hence adopted lemma ngrams for the majority of our experiments. We do use surface word ngrams as our baseline methods in some experiments, in which case we describe them. In the above example, "_BOS_" and "_EOS_" denote beginning-of-sentence and end-of-sentence respectively.

**PosNgram:** contiguous sequences of part-of-speech tags of length $n$ or smaller. For example, PosNgram ($n = 3$) of the above sentence will contain the indicator features for the following ngrams set to 1: *PRP*, *VBD*, *DT*, *NN*, *_BOS_ PRP*, *PRP VBD*, *VBD DT*, *DT NN*, *NN _EOS_*, *_BOS_ PRP VBD*, *PRP VBD DT*, *VBD DT NN*, and *DT NN _EOS_*. Here, PRP stands for personal pronoun, VBD stands for past tense verb, DT stands for determiner and NN stands for singular noun.

**MixedNgram:** a special formulation of word lemma ngrams where the lemmas of open-class words (nouns, verbs, adjectives and adverbs) are replaced with their corresponding POS tags. For example, MixedNgram ($n = 3$) of the above sentence will contain the indicator features for the following ngrams set to 1: *i*, *VBD*, *the*, *NN*, *_BOS_ i*, *i VBD*, *VBD the*, *the NN*, *NN _EOS_*, *_BOS_ i VBD*, *i VBD the*, *VBD the NN*, and *the NN _EOS_*.

# Part II

# MODELING DIALOG BEHAVIOR

# Chapter 5

# Dialog Act Tagging: Improving the Minority Class Identification

Dialog Act (DA) annotation and tagging, inspired by the speech act theory (Searle 1969), have long been used in the NLP community to understand and model the structure of dialog. A dialog act represents the communicative intent of an utterance, which is equivalent to the *illocutionary force* of (Austin 1975), *speech act* of (Searle 1969), and the *adjacency pair part* of (Sacks et al. 1974). From a computational perspective, assigning dialog act tags to utterances will provide a framework that "classifies utterances according to a combination of pragmatic, semantic, and syntactic criteria" (Stolcke et al. 2000). This serves as a way to model dialog structure that will help downstream tasks (e.g., a summarization system that needs to know who asked whom what). The dialog structure aspects captured in terms of dialog act tags can also be thought of as a way to model participants' dialog behavior, which may shed light to the social context of the interaction. In this thesis, we use dialog act analysis as one of the primary ways to model dialog behavior of the participants.

Early computational approaches towards dialog act modeling focused on spoken interactions (e.g., (Stolcke et al. 2000)). More recently, studies have explored dialog act tagging in written interactions such as emails (Cohen et al. 2004), online discussion forums (Kim et al. 2006; 2010b), instant messaging (Kim et al. 2010a) and Twitter (Zhang et al. 2012). Most early DA tagging systems for written interactions used a message/post level tagging scheme, and allowed multiple tags for each message/post (e.g., (Cohen et al. 2004)). Such a tagging scheme models dialog structure in

a rather coarse level — e.g., they detect that there was a request in a message, but do not identify the segment of text (e.g., a sentence) corresponding to the request. However, recent studies have found merit in segmenting each message into functional units and assigning a single DA to each segment (Hu et al. 2009), thereby capturing the dialog structure in a more fine-grained fashion. Our work falls in this paradigm (we choose a single DA for smaller textual units). In this thesis, we build on the work by (Hu et al. 2009) in organizational emails; we improve their dialog act prediction performance on minority classes using two new multi-class classification approaches we introduce in this thesis: the divide and conquer method that uses per-class feature optimization and the minority preference method that gives priority to minority class predictions. We obtain an overall accuracy error reduction of 10.6% and a minority class F-measure error reduction of 22.8% using the combination of these methods.

This chapter is structured as follows. We start by discussing related work on computational approaches towards dialog act modeling in Section 5.1, before describing the dialog act annotations we use in this work (Section 5.2). We then discuss the issue of identifying minority dialog act classes (Section 5.3) and present the novel multi-class classification techniques we introduce in this thesis in Section 5.4.3. In Section 5.4.4 we describe the experiments and results obtained in our automatic dialog act tagging experiments. Section 5.4.5 concludes the chapter.

## 5.1 Literature Review

The foundations of dialog act analysis stems from the speech act theory by (Searle 1969, Austin 1975). Austin (1975) proposed the analysis of speech acts at three levels: *locutionary act* which is the act of speaking an utterance, *illocutionary act* which is the act of using language to convey an intention (e.g., asking, answering, greeting etc.), and *perlocutionary act* which is the effect the utterance has on the hearer (e.g., being persuaded, scared, inspired etc.). Austin (1975) also proposed a classification of illocutionary acts along with a list of verbs that are examples of each class. Later, Searle (1976) argued that Austin's classification is of verbs rather than speech acts, and refined the classification of speech acts as Representatives, Expressives, Directives, Commissives, and Declarations, as described in Table 5.1.

One of the most significant early work in NLP community on dialog act analysis is the DAMSL

| Speech Act Type | Example Utterance | Speaker (S)'s communicative intention |
|---|---|---|
| Representatives | "[I state that] it is raining" | S commits (in varying degrees) to p |
| Expressives | "I thank you for leaving" | S expresses an attitude about p |
| Directives | "I order you to leave" | S wants H to do some action p |
| Commissives | "I will leave" | S commits self to some action p |
| Declarations | "You're fired" | S performs p by saying p |

Table 5.1: Speech acts classification proposed by Searle.

(Dialog Act Markup in Several Layers) annotation scheme developed by Core and Allen (1997). They argued that "an utterance might simultaneously perform actions such as responding to a question, confirming understanding, promising to perform an action, and informing". They also point out that Searle's speech acts do not capture how an utterance relates to the previous ones (e.g.: answering, accepting, rejecting). They proposed dialog act analysis of an utterance to be done in three layers: *Forward Communicative Functions* that captures what Searle's speech acts capture (e.g., statements, commissives etc.), *Backward Communicative Functions* to capture how the current utterance relate to previous parts of dialog (e.g., agreements, answering etc.), and *Utterance Features* to capture whether the utterance deal with the communication process or the content. In order to formulate the DAMSL tag-set, Core and Allen (1997) used the TRAINS corpus, a corpus of discussions on solving transportation problems involving trains. Later, Stolcke et al. (2000) adapted the DAMSL tag set to the switchboard corpus and proposed a SWBD-DAMSL tag set for a task-free environment. They used the SWBD-DAMSL annotations in the switchboard corpus to build a statistically trained Dialog Act tagger using hidden markov models achieving an accuracy of 71% using word transcripts.

Subsequently, many annotation schemes have been proposed within the NLP community to apply dialog act analysis on different genres of written interactions. Most of these schemes are specific to the genre of interactions and the granularity that is appropriate for that genre. In this direction, early work was done on dialog act tagging of email conversations. Cohen et al. (2004) proposed a 5-tag schema for email dialog acts: request, propose, amend, commit, and deliver. They also built a supervised learning system to automatically classify dialog acts, obtaining a best macro

average F-measure of around 56%. In further work (Carvalho and Cohen 2006), they applied more n-gram preprocessing and filtering techniques to obtain an error reduction of around 26.4%. Studies have also focused on identifying specific dialog acts such as action items in emails (Bennett and Carbonell 2005, Lampert et al. 2010). Bennett and Carbonell (2005) use emails from an educational institution to train an action-item detection system. They obtained a best F-measure of 77.9% using n-gram features and the kNN algorithm. In later work, Bennett and Carbonell (2007) used estimates of the sensitivity and variance of sentence-level action-item predictions to make more robust predictions at the macro-level. Lampert et al. (2010) used the Enron email corpus to train a system that can detect action items. Their features include the length of the utterance, usage of uppercase and Wh-words, in addition to n-grams. They also used n-gram preprocessing proposed by (Carvalho and Cohen 2006) as well as a method of segmenting email messages into different zones. They obtained a best F-measure of 84.3% using the zoning method.

There is also work in the genre of conversations happening in the web such as online discussion forums, instant messaging (IM) systems, and social media sites. (Kim et al. 2010b) proposed a 12-tag schema for forum dialog acts and presented a CRF-based automatic dialog act tagger obtaining a best F-measure of 75.3%. In later work, (Kim et al. 2010a) proposed a different 12-tag schema for instant messenger chat logs and presented a CRF-based tagger obtaining a best F-measure of 87.6%. More recently, (Zhang et al. 2012) proposed a 5-tag dialog act schema for twitter conversations. (Ferschke et al. 2012) proposed a dialog act annotation schema for Wikipedia discussion forums and presented a tagger that achieved an average F-measure of 82%.

Most DA tagging approaches on written interactions described above assign labels at a message/post level, allowing multiple tags for each post. For example, an email could be tagged as having both a request and proposal in (Cohen et al. 2004). However, recent studies have found merit in segmenting each message into functional units and assigning a single DA to each segment (Hu et al. 2009), thereby capturing the dialog structure in a more fine-grained fashion. Our work falls in this paradigm (we choose a single DA for smaller textual units). We use the annotations by (Hu et al. 2009), which we will describe in the next section.

## 5.2 Data and Annotations

In this thesis, we use the already existing dialog act annotations present in the ENRON-SMALL sub-corpus that was originally annotated by Hu et al. (2009). The details of these annotations and the annotation scheme they used is described in detail in Chapter 3, Section 3.1.2.1, page 40. Each message in an email thread is segmented into Dialog Functional Units (DFUs), which are contiguous spans within an email message which has a coherent communicative intention. In their annotations, each DFU is assigned a single DA label which is one of the following: INFORM, REQUEST-INFORMATION, REQUEST-ACTION, COMMIT, CONVENTIONAL, BACKCHANNEL, and OTHER. The annotation scheme they proposed was designed to work across different genres — written and spoken interactions. Some tags were more relevant to spoken interactions and less relevant to the written interactions, . For example, there were no instances of BACKCHANNEL, and only 3 instances of COMMIT (0.2%) and 2 instances of OTHER (0.1%) in the annotations they obtained on the Enron email corpus. We mapped the COMMIT and OTHER instances to the closely related class of INFORM, resulting in a 4-way distinction we use in this chapter, and the rest of the thesis. We briefly describe each of the 4-tags below (see Section 3.1.2.1, page 40 for more details).

- In a **REQUEST-ACTION**, the writer signals her desire that the reader perform some non-communicative act, i.e., an act that cannot in itself be part of the dialogue. For example, a writer can ask the reader to write a report or make coffee.

- In a **REQUEST-INFORMATION**, the writer signals her desire that the reader perform a specific communicative act, namely that he provide information (either facts or opinion).

- In an **INFORM**, the writer conveys information, or more precisely, the writer signals that her desire that the reader adopt a certain belief. It covers many different types of information that can be conveyed including answers to questions, beliefs (committed or not), attitudes, and elaborations on prior DAs.

- A **CONVENTIONAL** dialog act does not signal any specific communicative intention on the part of the writer, but rather it helps structure and thus facilitate the communication. Examples include greetings, introductions, expressions of gratitude, etc.

The relative proportion of each of the dialog act labels (under the 4-way distinction) in the ENRON-SMALL sub-corpus is given in Table 5.2

| Dialog Act Tag | Count | Percentage |
|---|---|---|
| REQUEST-ACTION | 35 | 2.5% |
| REQUEST-INFORMATION | 151 | 10.7% |
| CONVENTIONAL | 357 | 25.4% |
| INFORM | 853 | 60.7% |
| Total # of DFUs | 1406 | |

Table 5.2: Dialog act tag distribution in ENRON-SMALL corpus under the 4-tag distinction.

## 5.3 Automatic DA-Tagging: The Case of Minority Classes

In addition to the dialog act annotations, (Hu et al. 2009) also described the automatic dialog act tagger that they built using the annotations they obtained. They built two supervised learning systems — one using the Yamcha SVM framework, and another using the SVMstruct algorithm. Yamcha internally uses the regular one-vs-all SVM multi-class classification algorithm in which separate binary classifiers are built for each class and the final prediction is done by choosing the class based on the confidence of prediction by these individual classifiers. In contrast, the SVM-struct algorithm predicts the likelihood of a sequence of tags given an email thread, thereby obviating the need to have separate classifiers for each tag. They obtained an overall accuracy of 88.3% on 5-fold cross validation, using the Yamcha's regular multi-class SVM. The structured SVM did not report any significant improvement in the accuracy over using regular SVM, in the emails genre.

We further inspected their results at a per-class level. Table 5.3 presents the precision, recall, and F-measure for each class obtained using the regular one-vs-all multi-class classification approach. While the performance is pretty good as measured by accuracy, it performs poorly on the dialog act of REQUEST-ACTION. They reported a very low recall of 27.8% on REQUEST-ACTION with a precision of only 55.6%, resulting in a very low F-measure of 37.0%. The system reports a very high overall accuracy of 88.3% despite the low performance on REQUEST-ACTION, since it is a rare

class that accounts for only 2.5% of the data. However, for practical purposes such as systems that try to understand and model dialog behavior, these rare classes are of the most importance. For example, in the context of our study in which we analyze dialog behavior with respect to social power relations, being able to precisely identify requests for action issued by participants is potentially very important.

| | Precision | Recall | F-measure |
|---|---|---|---|
| REQUEST-ACTION | 55.6 | 27.8 | 37.0 |
| REQUEST-INFORMATION | 82.3 | 77.9 | 80.0 |
| CONVENTIONAL | 87.3 | 90.5 | 88.9 |
| INFORM | 90.6 | 92.5 | 91.5 |
| Accuracy | | 88.3 | |

Table 5.3: Dialog act tagging results reported by Hu et al. (2009) using Regular SVM

One of the primary reasons for the low performance in predicting REQUEST-ACTION is that it is a rare class. This leads the corresponding individual binary classifier to learn from heavily imbalanced training set. The class imbalance problem of SVM is a well studied one in the case of binary classification, and many different approaches have been proposed as solutions. But not much work has been done on this problem in a multi class setting. Apart from the underlying binary classifiers having to learn from skewed datasets, we suspect that there are more ways that the class imbalance will negatively impact the learned model. Specifically, we investigate two issues: 1) in the selection of the appropriate feature space, and 2) in using the same scale of confidence for all classes to make the final prediction. We describe them below.

### 5.3.1 Issue 1: Suboptimal Feature Spaces for Minority Classifiers

In order to better explain the issue of suboptimal feature spaces, we assume an empirical research methodology commonly followed within the applied machine learning community. Given the classification problem, a researcher trains different models using different feature combinations and selects the set of features that results in the model that gives the best overall performance, measured in terms of overall accuracy or micro/macro averaged F-measures of classes of interest. In other

words, the researcher is doing extrinsic feature optimization to find the "optimal" (within the set of experiments he/she conducts) feature space that represents the classification problem. In the case of multi-class classification problem, however, the minority classes will have minimal representation in the choice of this feature space since their performance will have very low impact on the overall accuracy. That is, features that are important only in distinguishing a minority class from other classes, will probably not make it to the final model, there by unfairly penalizing the minority class classifier.

For example, in our case, suppose a classifier built to distinguish REQUEST-ACTION needs a certain feature $f$ that will help make more accurate predictions. But for INFORM, $f$ is not helpful, and acts as noise thereby decreasing its performance. Hence, if we select the "optimal" feature set based purely on overall accuracy, we might end up excluding $f$, since the impact in improving the REQUEST-ACTION prediction on the overall accuracy is minimal. This will unfairly affect the REQUEST-ACTION class, since we are left with a suboptimal REQUEST-ACTION classifier, in the interest of INFORM classifier, by unnecessarily forcing them both to use the same feature set. In principle, each classifier is independent, and including $f$ only for REQUEST-ACTION does not affect the other classifiers' performance. So, we introduce a divide and conquer (DAC) method in selecting feature configurations. We do separate per-class feature optimizations in order to find the feature space that best captures the particular class in question. This will not affect the performance of any other classifier, but improves the performance of individual classifiers.

### 5.3.2 Issue 2: Unfair Ranking of Minority Classifier Confidences

Secondly, the classifier trained to detect the minority class has very small number of positive instances to learn from, and hence will result in relatively lower confidence in its positive predictions, compared to other classes. This may lead to true positive predictions by the minority class classifier being of relatively lower confidence, and hence being drowned by other classes, especially for the borderline cases. As part of this thesis, we propose two methods to handle this issue: Minority Preference (MP) and Cascaded Minority Preference (CMP). Both methods give preference to the minority classes; i.e., if a minority class predicts *true* despite having substantially fewer positive instances, then it is given preference over *true* predictions by classes with more positive instances in the data, regardless of the prediction confidence.

## 5.4 An Improved Dialog Act Tagger

In this section, we present an improved dialog act tagging system that uses specific techniques to handle the issues described in Section 5.3.1 and Section 5.3.2. We start by describing the machinery used to implement the system and the different sets of features we experimented with. We then describe the techniques to handle the issues with minority-class performance in the multi-class classification setting and present the experiments and results.

### 5.4.1 Implementation

We use the UIMA architecture and the ClearTK suite of UIMA tools (Ogren et al. 2008) to build the automatic dialog act tagger. We use the default ClearTk tokenizer, part-of-speech tagger and lemmatizer to obtain features for our experiments. We use a linear kernel Support Vector Machine (SVM) as the base machine learning algorithm, for which we use the ClearTK wrapper for SVM-Light (Joachims 1999). The ClearTK SVMlight wrapper internally shifts the prediction threshold based on posterior probabilistic scores calculated using the algorithm of Lin et al. (2007) which handles the class imbalance problem for the basic binary classification.

### 5.4.2 Features

We experimented using three categories of features — LEXICAL, VERB-BASED, and DIALOGIC. Table 5.4 lists the features in each category. We describe each feature below.

**LEXICAL:** This set of features include n-gram features and other token level features that can be extracted from the lemma and part-of-speech of the token in the DFU. It consists of three types of ngram features — word lemma ngrams, part-of-speech ngrams, and mixed ngrams. In addition, it contains a small set of specialized features. We describe all LEXICAL features below.

- *LemmaNgram*: word lemma ngrams.
- *PosNgram*: part-of-speech ngrams.
- *MixedNgram*: a special formulation of word lemma ngrams where the lemmas of open-class words (nouns, verbs, adjectives and adverbs) are replaced with their corresponding POS tags (see Chapter 4).

| Category | Feature Set | Description |
|---|---|---|
| LEXICAL | *LemmaNgram* | Lemma N-grams |
| | *PosNgram* | Part-of-speech N-grams |
| | *MixedNgram* | Mixed N-grams |
| | *StartLemma* | Lemma of the first word |
| | *StartPOS* | Part-of-speech tag of the first word |
| | *LastLemma* | Lemma of the last word |
| | *LastPOS* | Part-of-speech tag of the last word |
| | *MDCount* | Number of modal verbs in the DFU |
| | *QuestionMark* | Is there a question mark in the DFU? |
| VERB-BASED | *FirstVerbLemma* | Lemma of the first verb in the DFU |
| | *VerbBeforeNoun* | Did a verb occur before the first noun? |
| | *VerbBeforeNounLemma* | Lemma of the verb occurred before the first noun |
| DIALOGIC | *PosFromBegin* | Position of the DFU from the beginning of the message |
| | *PosFromEnd* | Position of the DFU from the end of the message |
| | *PosFromEitherEnd* | Position of the DFU from the either end of the message |
| | *Size* | Size of the DFU in terms of number of word tokens |

Table 5.4: Features used for dialog act tagging.

- *StartLemma* & *StartPOS*: word lemma and part-of-speech tag of the first word in the DFU.
- *LastLemma* & *LastPOS*: word lemma and part-of-speech tag of the last word in the DFU.
- *MDCount*: number of modal verbs in the DFU.
- *QuestionMark*: binary feature denoting whether there is a question mark in the DFU.

**VERB-BASED:** This set of features specifically looks at the first verb of the DFU. It includes the following three features.

- *FirstVerbLemma*: the lemma of the first verb (a word with a part of speech tag starting with 'VB') in the DFU.

- *VerbBeforeNoun*: a binary feature indicating that a verb occurred before the first noun (a word with a part of speech tag starting with 'NN') in the DFU.
- *VerbBeforeNounLemma*: the lemma of the verb that occurred before the first noun. feature will be assigned value only when VerbBeforeNoun is *true*.

**DIALOGIC:** This set of features capture extra-linguistic aspects of the DFU; i.e., the size of the DFU and its position with respect to the message. It includes the following four features

- *PosFromBegin*: the relative position of the DFU from the beginning of the message.
- *PosFromEnd*: the relative position of the DFU from the end of the message.
- *PosFromEitherEnd*: the minimum of *PosFromBegin* and *PosFromEnd*.
- *Size*: the number of tokens in the DFU.

### 5.4.3 Methods

In this section we describe three different methods we employ to handle the multi-class classification of dialog act labels — Divide And Conquer (DAC), Minority Preference (MP), and Cascaded Minority Preference (CMP). The first method attempts to handle the issue of suboptimal feature spaces, whereas the second and third methods address the issue of unfair ranking of minority class classifiers' confidences.

#### 5.4.3.1 Divide And Conquer (DAC)

We introduce the method of Divide And Conquer (DAC) to solve the issue with suboptimal feature spaces for the minority class classifiers described in Section 5.3.1. As in the regular multi-class SVM, the DAC system also builds a binary classifier for each dialog act separately, and the component classifier with highest probability score determines the overall prediction. The crucial difference in the DAC system is that the feature optimization is performed for each component classifier separately. This allows for us to find out the optimal (within the set of experiments conducted) feature space for each individual class, leading to more accurate predictions by each binary classifier. We optimize each component classifier for the F-measure of the class they are trying to predict.

#### 5.4.3.2 DAC with Minority Preference (DAC-MP)

This method builds upon the basic DAC system except for one crucial difference: overall classification is biased towards a specified minority class. If the minority class binary classifier predicts true, this system chooses the minority class as the predicted class. In cases where the minority class classifier predicts false, it backs off to the basic DAC system after removing the minority class classifier from the confidence tally.

#### 5.4.3.3 DAC with Cascaded Minority Preference (DAC-CMP)

This method is similar to the Minority Preference System; however, instead of a single supplied minority class, the system accepts an ordered list of classes. The classifier then works, in order, through this list; whenever any classifier in the list predicts true, for a given instance, it then assigns this class as the predicted class. The subsequent classifiers in the list are not run. If all classifiers predict false, we back off to the basic DAC system, i.e., the component classifier with highest probability score determines the overall prediction. For our experiments, we ordered the list of classes in the ascending order of their frequencies in the training data. This ordering is driven by the observation that the less frequent classes are also hard to predict correctly.

### 5.4.4 Experiments and Results

In this section, we describe the various experiments conducted using the methods introduced in Section 5.4.3 for the problem of multi-class classification of dialog act labels. We report precision, recall, and F-measure obtained on 5-fold cross validation performed on the entire corpus.

#### 5.4.4.1 Feature Optimization Experiments

The methods described in Section 5.4.3 crucially differ from the regular one-vs-all multi-class classification algorithm in terms of how the feature optimization is performed. We use the following same steps to find the optimal feature space for both the regular one-vs-all method and the DAC-based methods.

1. We first find the optimal width for each n-gram feature by varying the value of $n$ from 1 to 5.

2. We then find the optimal feature subset in each feature category through an exhaustive search in the space of all feature subsets of each category. This results in 511 ($2^9 - 1$) experiments for LEXICAL, 7 ($2^3 - 1$) experiments for VERB-BASED, and 15 ($2^4 - 1$) experiments for DIALOGIC.

3. We then perform all the seven combinations of these three best feature subsets to obtain the overall best feature subset.

In each step, if more than one feature subset give the same best performance, we choose the feature subset with the smallest cardinality; i.e., the minimal feature subset.

#### 5.4.4.2 Baseline: One-vs-All SVM (BAS)

This system uses the ClearTK built-in one-versus-all multiclass SVM in prediction. Internally, the multi-class SVM builds a set of binary classifiers, one for each dialog act. For a given test instance, the classifier that obtains the highest probability score determines the overall prediction. We performed feature optimization on the whole multiclass classifier, (as described in Section 5.4.4.1), i.e., the same set of features was available to all component classifiers. We optimized for system accuracy. Table 5.5 shows results using the baseline system. We give the performance of the system on the four dialog acts, using precision, recall, and F-measure. The dialog acts are listed in ascending order of frequency in the corpus (least frequent dialog act first). We also give an overall accuracy evaluation. As we can see, detecting REQUEST-ACTION is much harder than detecting the other dialog acts.

| | Precision | Recall | F-measure |
|---|---|---|---|
| REQUEST-ACTION | 57.9 | 31.4 | 40.7 |
| REQUEST-INFORMATION | 91.5 | 78.2 | 84.3 |
| CONVENTIONAL | 92.0 | 95.8 | 93.8 |
| INFORM | 91.6 | 95.1 | 93.3 |
| Accuracy | | 91.3 | |

Table 5.5: Results for baseline (BAS) system (standard one-vs.-all multi-class SVM)

#### 5.4.4.3 Divide And Conquer (DAC)

We first apply the Divide and Conquer method described in Section 5.4.3; i.e., we apply per class feature optimization. For each individual binary classifier, we perform the feature optimization steps as described in Section 5.4.4.1. The optimal set of features obtained for each classifier is summarized in Table 5.6. As can be seen from the table, the feature sets that worked best for each individual classifier differ considerably. For example, for REQUEST-ACTION, the MIXEDNGRAM was very useful, but it was not useful for any other classes. For both REQUEST-ACTION and REQUEST-INFORMATION, the part-of-speech ngrams were useful, but it was not useful for CONVENTIONAL and INFORM. Moreover, REQUEST-ACTION benefited from longer part-of-speech sequences (n=4), whereas REQUEST-INFORMATION required only the unigrams and bigrams of par-of-speech tags. The REQUEST-INFORMATION classifier performed best when using only the part-of-speech ngrams and the binary feature denoting whether there is a question mark in the sentence. All other lexical features other than ngrams were useful only for CONVENTIONAL and INFORM.

| Feature Set | REQ-ACTION | REQ-INFORM. | CONV. | INFORM |
|---|---|---|---|---|
| *LemmaNgram* | ✓(n=1) | | ✓(n=2) | ✓(n=2) |
| *PosNgram* | ✓(n=4) | ✓(n=2) | | |
| *MixedNgram* | ✓(n=2) | | | |
| *StartLemma* | | | ✓ | ✓ |
| *StartPOS* | | | ✓ | ✓ |
| *LastLemma* | | | | |
| *LastPOS* | | | | ✓ |
| *MDCount* | | | ✓ | |
| *QuestionMark* | | ✓ | ✓ | ✓ |

Table 5.6: Best features for individual classifiers obtained through the DAC method

The results obtained for the DAC system using the maximum confidence based choice with individual classifiers optimized separately is presented in Table 5.7. The biggest improvement in performance was obtained for REQUEST-ACTION, for which both precision and recall was considerably improved. It resulted in an F-measure error reduction of 15.6% for REQUEST-ACTION.

| | Precision | Recall | F-measure | Error Reduction (%) |
|---|---|---|---|---|
| REQUEST-ACTION | 66.7 | 40.0 | 50.0 | 15.6 |
| REQUEST-INFORMATION | 91.5 | 78.2 | 84.3 | 0.0 |
| CONVENTIONAL | 93.9 | 94.1 | 94.0 | 2.6 |
| INFORM | 91.4 | 96.1 | 93.7 | 5.7 |
| Accuracy | | 91.7 | | 4.9 |

Table 5.7: Results for the Divide And Conquer (DAC) system (per-class feature optimization followed by maximum confidence based choice). Error reduction is calculated with respect to the standard multi-class SVM

REQUEST-INFORMATION, on the other hand, posted no improvements. CONVENTIONAL improved the precision from 92.0 to 93.9, at the cost of recall going down from 95.8 to 94.1, resulting in an F-measure error reduction of 2.6%. INFORM improved recall from 95.1 to 96.1, at a marginal reduction in precision from 91.6 to 91.4, resulting in an F-measure error reduction of 5.7%. Overall, DAC method posted an accuracy error reduction of 4.9% just by doing per-class feature optimization of individual classifiers, the biggest improvement being for the minority class of REQUEST-ACTION.

**System 2: Divide And Conquer with Minority Preference (DAC-MP)** Now we apply the minority preference (MP) method presented in Section 5.4.3 to the DAC system to obtain the DAC-MP system. That is, if the minority class classifier (REQUEST-ACTION in our case) predicts true, then the system would choose the minority class as the predicted class. In cases where the minority class classifier predicts false, it backs off to the basic DAC system after removing the minority class classifier from the confidence tally.

Table 5.8 shows our results using this method. Since the DAC-MP approach is biased towards the minority class REQUEST-ACTION, the recall of REQUEST-ACTION improved considerably from 40.0 to 45.7, resulting in an F-measure of 54.2, a 22.8% F-measure error reduction from the original one-vs-all BAS classifier. The performance of REQUEST-INFORMATION and CONVENTIONAL did not change, whereas the precision of INFORM improved from 91.4 to 91.6 at a small decrease in recall from 96.1 to 96.0. This suggests that using the DAC-MP method, many cases of REQUEST-

| | Precision | Recall | F-measure | Error Reduction (%) |
|---|---|---|---|---|
| REQUEST-ACTION | 66.7 | 45.7 | 54.2 | 22.8 |
| REQUEST-INFORMATION | 91.5 | 78.2 | 84.3 | 0.0 |
| CONVENTIONAL | 93.9 | 94.1 | 94.0 | 2.6 |
| INFORM | 91.6 | 96.0 | 93.8 | 6.5 |
| Accuracy | | 91.8 | | 5.7 |

Table 5.8: Results for the DAC Minority Preference (DAC-MP) system (first consult REQUEST-ACTION tagger, then default to choice by maximum confidence). Error reduction is calculated with respect to the standard multi-class SVM

ACTION that were incorrectly classified as INFORM by the DAC classifier were corrected. Overall, the DAC-MP method reported an accuracy error reduction of 5.7% from the BAS classifier.

**System 3: Divide And Conquer with Cascaded Minority Preference (DAC-CMP)** Now we apply the cascaded minority preference (CMP) method presented in Section 5.4.3 to the DAC system to obtain the DAC-CMP system. This system is similar to the Minority Preference System; however, instead of a single supplied minority class, the system accepts an ordered list of classes. The classifier then works, in order, through this list; whenever any classifier in the list predicts true, for a given instance, it then assigns this class as the predicted class. The subsequent classifiers in the list are not run. If all classifiers predict false, we back off to the basic DAC system, i.e., the component classifier with highest probability score determines the overall prediction. We ordered the list of classes in the ascending order of their frequencies in the training data. This ordering is driven by the observation that the less frequent classes are also hard to predict correctly. It is also the case that the less frequent classes happen to be more useful for our subsequent processing and we want to increase their recall.

Table 5.9 shows our results using this method. The performance of REQUEST-ACTION did not change using DAC-CMP since the cases where the system predicts REQUEST-ACTION remain the same in both DAC-MP and DAC-CMP. However, the performance of all other classes go up considerably using DAC-CMP. The recall of REQUEST-INFORMATION and CONVENTIONAL both

| | Precision | Recall | F-measure | Error Reduction (%) |
|---|---|---|---|---|
| REQUEST-ACTION | 66.7 | 45.7 | 54.2 | 22.8 |
| REQUEST-INFORMATION | 91.0 | 80.8 | 85.6 | 8.4 |
| CONVENTIONAL | 93.7 | 95.3 | 94.5 | 10.1 |
| INFORM | 92.4 | 95.8 | 94.0 | 10.0 |
| Accuracy | | 92.2 | | 10.6 |

Table 5.9: Results for the DAC Cascading Minority Preference (DAC-CMP) system (consult classifiers in reverse order of frequency of class). Error reduction is calculated with respect to the standard multi-class SVM

improved by 2.6 and 1.2 percentage points respectively, while their precision dropped by a smaller margin (0.5 and 0.2 percentage points respectively). This resulted in an F-measure error reduction of 8.4% for REQUEST-INFORMATION and 10.1% for CONVENTIONAL. The improvement in performance for INFORM was on precision (91.6 to 92.4), at a smaller cost of recall (96.0 to 95.8), resulting in an F-measure error reduction of 10.0% over the BAS classifier. Overall, the DAC-CMP system posted an overal accuracy error reduction of 10.6% over the BAS classifier.

### 5.4.5 Post-hoc Analysis

Following (Guyon et al. 2002), we performed a post-hoc analysis by inspecting the feature weights of the best performing models created for each individual classifier in the DAC system. In a linear kernel SVM such as the one we built, the feature weights assigned in the model for each feature is an indicator of how that feature correlates with the class being predicted. Table 5.10 lists some interesting features chosen during feature optimization for the individual SVMs. We selected them from the top 25 features in terms of absolute value of feature weights.

Some features help distinguish different DA categories. For example, the feature *QuestionMark* is the feature with the highest negative weight for INFORM, but has the highest positive weight for REQUEST-INFORMATION. Features like *fyi* and *period (.)* have high positive weights for INFORM and high negative weights for CONVENTIONAL. Some other features are important only for certain classes. For example, *please* and *VB_NN* are important for REQUEST-ACTION, but not so for other

| REQUEST-ACTION | REQUEST-INFORMATION | CONVENTIONAL | INFORM |
|---|---|---|---|
| please (0.9) | QuestionMark (6.6) | StartPOS_NNP (2.7) | QuestionMark (-3.0) |
| VB_NN (0.7) | _BOS_PRP (-1.2) | thanks (2.3) | thanks (-2.2) |
| you_VB (0.3) | WRB (1.0) | . (-2.0) | . (2.2) |
| PRP (-0.3) | PRP_VBP (-0.9) | fyi (-2.0) | fyi (1.9) |
| MD_PRP_VB (0.3) | _BOS_MD (0.8) | , (0.9) | you (-1.0) |
| will (-0.2) | _BOS_DT (-0.7) | QuestionMark (-0.8) | can_you (-0.9) |

Table 5.10: Post-hoc analysis on models built by the DAC system for each class: some of the top features with corresponding feature weights in parentheses, for each individual tagger. (POS tags are capitalized; _BOS_ stands for Beginning Of Sentence)

classes. Overall, the most discriminating features for both INFORM and CONVENTIONAL are mostly word ngrams, while those for REQUEST-ACTION and REQUEST-INFORMATION are mostly POS ngrams. This shows why our approach of per-class feature optimization is important to boost the classification performance.

Another interesting observation is that the least frequent category, REQUEST-ACTION, has the least strong indicators (in terms of feature weights). Presumably this is because there are much fewer positive instances for this class in the training data. This explains why our cascading classifiers approach giving priority to the least frequent categories worked better than a simple confidence based approach, since the simple approach drowns out the less confident classifiers.

## 5.5 Conclusion

In this chapter, we described our work on automatically obtaining the dialog act tags that we use for our dialog structure analysis in the rest of this thesis. We introduced two new methods to improve the performance of multi-class SVM classification algorithms — the Divide and Conquer method and the Cascaded Minority Preference method. We also built an automatic dialog act tagger using these methods. The combination of our methods obtained an F-measure error reduction of 10.6% over using the regular one-vs-all multi-class classification algorithm. More importantly, we obtained

around 23% F-measure error reduction on the minority class (requests for action) prediction, which is a crucial improvement for the analysis we perform in the rest of this thesis.

Our methods to improve multi-class classification have already been applied to other problems by other researchers obtaining significant improvements (Hou et al. 2013). We performed a detailed analysis of how our methods are able to obtain the improvements. We showed that in a multi-class classification setting, different individual classes perform best using different feature spaces. For example, our REQUEST-ACTION classifier performs best when we include our formulation of mixed ngrams in the feature set, but classifiers for all other classes perform better without using them. This is the crucial point on which our DAC method is based on. Similarly, we showed that the least frequent classes often have the least strong indicators (in terms of feature weights), limiting their ability to make highly confident predictions. Hence, giving preference to their positive predictions, as we do in our cascaded minority preference method, improved their recall resulting in an overall error reduction.

# Chapter 6

# Overt Display of Power

Dialog is successful when all discourse participants exhibit cooperative dialog behavior. Certain types of dialog acts, notably requests for actions and requests for information (questions), "set constraints on what should be done in a next turn" (Sacks et al. 1974). For example, a request dialog act is the first part of an adjacency pair and thus requires a response from the addressee. From a dialog act perspective, in order to exhibit cooperative dialog behavior, a next turn must perform an act that is a response to the request issued. This response could also be the act of declining the request. The utterer may, however, formulate her request in a way that attempts to remove the option of declining it (e.g., *Come to my office now!*). In so doing, she restricts her addressee's options for responding more severely than a simple request for action would. Such a "restriction of an interactant's action-environment" is identified as one of the primary means of the exercise of power in interactions (Van Dijk 1989, Wartenberg 1990, Locher 2004, Bousfield and Locher 2008). In this chapter, we introduce the notion of "Overt Display of Power" (ODP) to denote such instances, and describe computational techniques to automatically detect them in interactions.

We start by formally defining the notion of overt display of power in Section 6.1. We then describe how this notion relates to the sociolinguistics literature on face, impoliteness, and exercise of power (Section 6.2). In Section 6.3, we describe the process we followed in obtaining manual annotations for ODP and present an in-depth analysis of them, discussing specific examples and annotation statistics. Section 6.4 describes the supervised machine learning system we built to automatically tag instances of overt displays of power in interactions.

## 6.1 What is Overt Display of Power?

We use the term "Overt Display of Power" (ODP) to capture utterances in dialog that display the exercise of power in an overt way. Although we characterize it in terms of linguistic form, we are more interested in how it affects the dialog, more specifically, how it affects its responses. We start by illustrating the notion of ODP using an example scenario. Suppose the utterance in Example 6.1 is from an email message exchanged within an organizational setting. Although phrased as a declarative statement, its communicative intention is to request an action; the utterer requesting the addressee to come to her office.

**Example 6.1.** *It would be great if you could come to my office as soon as you can.*

Using the dialog act tag-set we discussed in Chapter 5, this utterance would be labeled as a REQUEST-ACTION. As a request, this utterance sets up an expectation that there be a response. Acceptable responses from a dialog perspective include a commitment to performing the action, actually performing the action, or rejecting the request (with or without an explanation), while unacceptable responses include silence, or changing the topic. If the addressee responds by declining (*Would love to, but unfortunately I need to pick up my kids*), he has still met the dialog expectation, and hence has exhibited cooperative dialog behavior.

However, the high-level dialog act of an utterance provides only an initial description of what constraints apply to its response. Other sources of constraints include the social relations between the utterer and the addressee, and clues from the language used in the utterance. Suppose the utterance in our Example 6.1 had come, say, from the CEO to a lower-level employee. In this case, the addressee's response declining the request would not have met the constraints set by the utterance within the social context, even though it is still analyzed as the same dialog act (a request for action). Detecting such social relations and determining their effect on dialog is a hard problem, and is the ultimate goal of this thesis. In this chapter, we focus on another source of constraint — linguistic clues from the utterance.

Consider the the utterance in Example 6.2. It is a request for action, requesting the same action as that in Example 6.1, however the linguistic form used in making the request is different.

**Example 6.2.** *Please come to my office immediately.*

If the addressee now declines the request, he is clearly not adhering to the constraints the sender has signaled, though he is adhering to the general constraints of cooperative dialog by responding to the request for action. That is because the utterer has chosen a linguistic form in her utterance to signal that she is imposing further constraints on the addressee's choices of how to respond, constraints which go beyond those defined by the standard set of dialog acts. We consider such utterances with linguistic forms that create additional constraints on its addressee's response as instances of overt displays of power.

**Definition 6.1.** *We define an utterance to have* **Overt Display of Power** *(ODP) if it is interpreted as creating additional constraints on its response beyond those imposed by the general dialog act.*

Note that we define ODP as a pragmatic concept, i.e., in terms of the dialog constraints an utterance introduces to its response, and not in terms of specific linguistic markers. For example, the use of politeness markers (e.g., use of "please" in Example 6.2) does not, on its own, determine the presence or absence of an ODP. Also, as we will see in Section 6.3.3, presence of ODP cannot be determined solely based on syntactic patterns alone. Instead, we adopt a data-oriented approach where we learn the linguistic markers that are salient to ODP.

We also do not make any assumptions about the intention of the utterer. That is, we call an utterance to be an ODP based on whether it can be interpreted as creating additional constrains on its response, irrespective of whether or not the utterer intended to. In other words, we do not try to *guess* the utterer's intention. For example, in the utterance *please come to my office immediately* (Example 6.2), the utterer may not have intended to overtly display power, but it will still be analyzed as an utterance with ODP.

Furthermore, our focus in this chapter is on the clues introduced through the language used in the utterance, and not on the factors external to the dialog such as social relations between the utterer and the addressee. For example, the presence of an ODP does not presuppose that the utterer actually possess social power.

## 6.2 Theoretical Framework

In this section, we relate the notion of overt display of power to different sociolinguistics theories.

### 6.2.1 ODP as Restriction of Action Environment

Sociolinguistics studies have looked at how exercise of power can be recognized in language. Locher (2004) identifies restrictions of the "action-environment" of an interactant as one of the key elements by which the exercise of power in interactions can be identified. Locher (2004) derives the notion of action-environment from prior work by Van Dijk (1989), Wartenberg (1990) on defining power. Wartenberg (1990) defines power as follows: "an agent who exercised power over another agent does so by affecting the circumstances within which the other agent acts and makes choices". In our analysis, the action-environment is the set of actions an addressee can take in a next turn, and by using a specific linguistic form, the utterer is imposing constraints on that set of actions. In other words, ODP is an instance of what Locher would consider as exercise of power.

### 6.2.2 Relation to Face and Politeness

The concept of face proposed by Brown and Levinson (1987) is the seminal work on politeness theory. Face is a "public self-image" of an individual within an interaction. (Brown and Levinson 1987) define two types of faces: positive face and negative face. Positive face is the "want of every member that his wants be desirable to at least some others". Negative face is the "want of every 'competent adult member' that his actions be unimpeded by others". Positive Face refers to one's self-esteem, while negative face refers to one's freedom to act. The two aspects of face are the basic wants in any social interaction, and so during any social interaction, cooperation is needed amongst the participants to maintain each other's faces. (Brown and Levinson 1987) argue that "face respect is not an unequivocal right". In other words, face wants may not always be recognized. Although the interactants may cooperate to maintain each other's face for common interest, there are acts that intrinsically threaten face. These acts are called *face-threatening acts*. Overt display of power is a face threatening act in that it threatens the negative face of the addressee. There are other face threatening acts that are not considered overt display of power; e.g., disapprovals (threatening the positive face of addressee) and apologies (threatening the positive face of utterer).

The sociolinguistics construct of *Impoliteness* (Hickey 1991, Culpeper 1996, Rudanko 2006, Bousfield and Locher 2008, Locher and Bousfield 2008) is very related to our notion of overt display of power. Locher and Bousfield (2008) argue that the lowest common denominator on different schools of thoughts on what impoliteness is that "Impoliteness is behavior that is face-aggravating

in a particular context". Bousfield (2008) defines impoliteness as "constituting the issuing of intentionally gratuitous and conflictive face-threatening acts (FTAs) that are purposefully performed". Culpeper (2008) defines it as involving "communicative behaviour intending to cause the 'face loss' of a target or perceived by the target to be so". Both of these definitions stresses the speaker's intention and the hearer's understanding of that intention. Terkourafi (2008) argues that impoliteness occurs "when the expression used is not conventionalised relative to the context of occurrence", and, in contrast to the above definitions, that "no face-threatening intention is attributed to the speaker by the hearer". Our notion of overt display of power is closer to (Terkourafi 2008) in that we do not make any assumptions about the speaker's intention, but rather limits our focus to the linguistic form. Regardless, Locher and Bousfield (2008) argue that impoliteness is inextricably tied up with power because an interlocutor whose face is damaged by an utterance suddenly finds his or her response options to be sharply restricted, a notion central to the exercise of power (Wartenberg 1990). She also argues that "even interactants with a hierarchically lower status can and do exercise power through impoliteness", which is also the case with ODP.

## 6.3 Data and Annotations

In this section, we describe the process of obtaining manual annotations for instances of overt displays of power in interactions. We used the ENRON-SMALL corpus (Chapter 3; Section 3.1.4, page 42) for this purpose. The corpus contains 122 email threads with 360 messages. The corpus already contains dialog act annotations by Hu et al. (2009) in which each message is segmented into dialog functional units, and each dialog functional unit is further split into utterances. Here, an utterance is roughly equivalent to a sentence. Refer to Section 3.1.2.1, page 40 for a detailed discussion of dialog act annotations. Our annotations for overt displays of power were obtained at the level of utterances. There were 1734 utterances in the corpus.

### 6.3.1 Annotator Training

We hired and trained a manual annotator, an undergraduate student who is a native speaker of English. The annotation task was to analyze each utterance and label whether it is an instance of overt display of power or not. The original instruction given to the annotator was this:

> Label a DFU with a Y if it requires the recipient to react according to one of a finite set of options. Do not consider the use of polite language alone as giving an option.

The instructions also included Examples 6.3 to 6.5; [Y] denote an instance of overt display of power and [N] denote an utterance that do not contain an overt display of power.

**Example 6.3.** *If you have not, can you do so immediately.* : [Y]

**Example 6.4.** *Will you be able to sit in on this and decide if we should participate further for our group?* : [N]

**Example 6.5.** *If there is any movement of these people between groups can you plese keep me in the loop.* : [Y]

The annotator was given full email threads, in which the messages were already manually segmented into separate utterances. After the annotator went through a small set of threads, we held different follow up discussions with her to clarify her questions. These discussions with the annotator lead to refining the definition of ODP to Definition 6.1. The annotation manual was appended with the following main clarification questions brought up by the annotator:

Q: Does an example of this [Overt Display of Power] have to be a question where the correct response is yes or no?

A: Examples for Overt Display of Power need not necessarily be yes/no questions. A DFU is an overt show of power if the sender gives the receiver a limited set of options. For example “Please do job A or B” and “Please do job A” would both be labeled as overt display of power. However “Please do job A if you have the time” is not an overt display of power.

Q: Reference utterance: “If you have any additional questions or comments, please call me.”, . . . Is this really a request to do something, or just an open invitation that does not require any response?

A: [It] should be labeled [N] because a response is optional.

Q: Should the [N] label be explicit?

A: The label [N] need not be explicit - the example in the manual is just an illustration - you only need to label the [Y]s.

### 6.3.2 Annotation Statistics

Out of the 1734 utterances in the corpus, our annotator identified 86 utterances (about 5%) to have an overt display of power.

| | |
|---|---|
| Number of threads | 122 |
| Number of utterances | 1734 |
| Number of utterances with overt display of power | 86 |
| Percentage of utterances with overt display of power | 4.96% |

Table 6.2: Overt display of power annotation statistics.

In order to validate the annotations, we trained another annotator, also an undergraduate student who is a native English speaker, using the same definitions and examples and had him annotate 46 randomly selected threads from the corpus, which contained a total of 595 utterances (34.3% of whole corpus). We obtained a reasonable inter annotator agreement, $\kappa$ value of 0.669, which validates the annotations while confirming that the task is not a trivial one.

| | |
|---|---|
| Number of threads | 46 |
| Number of utterances | 595 |
| Observed agreement, P(a) | 96.5% |
| Expected agreement, P(e) | 89.3% |
| Cohen's Kappa, $\kappa$ | 0.669 |

Table 6.3: Inter annotator agreement of overt display of power annotations.

### 6.3.3 Syntactic Configurations and ODP

Identifying instances of overt displays of power is not a purely syntactic task. An utterance with ODP can be an imperative sentence (Example 6.6), an interrogative sentence (Example 6.7) or even a declarative (Example 6.8) sentence.

**Example 6.6.** *Please give me your views ASAP.* (Reference thread: A)

**Example 6.7.** *Would you work on that?* (Reference thread: A)

**Example 6.8.** *I need the answer ASAP, as we are going to discuss the additional summer intern positions this afternoon.* (Reference thread: A)

However, not all imperatives (Example 6.9) or interrogatives (Examples 6.10, 6.11) are overt displays of power. Examples 6.7, 6.10 and 6.11 are all syntactically questions. However, Example 6.7's discourse function within the email thread (Appendix A) is to request/order to work on "that" which makes it an instance of ODP, while Example 6.10 is merely an inquiry and Example 6.11 is a rhetorical question.

**Example 6.9.** *Keep up the good work!* (Reference thread: A)

**Example 6.10.** *... would you agree that the same law firm advise on that issue as well?* (Reference thread: A)

**Example 6.11.** *can you BELIEVE this bloody election?* (Reference thread: A)

### 6.3.4 Dialog Acts and ODP

Identifying an instance of overt display of power is closely related to identifying the dialog act of the utterance. An utterance with ODP can be a request that is an explicit order or command (Example 6.6) or an implicit one (Examples 6.7, 6.8). However, not all requests are ODPs. Even requests for actions could be expressed in a non-ODP way. In Example 6.12, the communicative intent of the author is to request the addressee to "leave the call-in number", but there is no overt display of power.

**Example 6.12.** *Sorry to bother you with this, but I'm travelling, and if you could leave the call-in number for tomorrow's meeting on my voice mail, I'll be forever indebted.* (Reference thread: A)

Similarly, 6.13 is a request for action, but the presence of conditional "if interested . . ." in this case acts as a way to not limit the addressee's options, and hence the utterance is analyzed as not an overt display of power.

**Example 6.13.** *if interested in putting these contracts in place, forward to me all required information on the first page of the GISB contract form.* (Reference thread: A)

| Dialog Act (DA) Tag | # of utterances | % of utterances with ODP |
|---|---|---|
| REQUEST-ACTION | 39 | 79.5% |
| REQUEST-INFORMATION | 155 | 15.5% |
| INFORM | 1125 | 2.8% |
| CONVENTIONAL | 415 | 0.0% |

Table 6.4: Distribution of overt displays of power across different dialog act tags

However, the presence of a conditional does not always open up addressee's action environment. In Example 6.14, the request to "keep me in the loop" is an overt display of power in spite of the presence of the conditional. Here, the difference is that the conditional is about a world event, where as the conditional in the above example is about the addressee's volition.

**Example 6.14.** *If there is any movement of these people between groups can you please keep me in the loop.* (Reference thread: A)

Table 6.4 shows the percentage of each dialog act tag that was judged to be an ODP. Around 79.5% of REQUEST-ACTION utterances were annotated as instances of ODP. That means that there is a substantial percentage of REQUEST-ACTION utterances (over 20%) that are not ODP. When it comes to REQUEST-INFORMATION, only around 16% was annotated as ODP. To our surprise, around 2.7% of the utterances that were labeled as INFORM in the underlying DA annotations were tagged as ODP. On further analysis, we found that these were incorrectly tagged as INFORM in most cases. For example, Example 6.15 was tagged as INFORM in the underlying DA annotations, instead of REQUEST-ACTION. Our annotator, however, correctly identified this utterance as an instance of ODP.

**Example 6.15.** *Per Daren's repsonse below, can you correct price on this deal for 03/01.* (Reference thread: A)

## 6.4 Automatic ODP Tagging

In this section, we describe the ODP Tagger, a system that can automatically detect instances of overt display of power in interactions. Within this thesis, we use this ODP tagger to obtain ODP

labels in large amounts of data in which manually obtaining those labels is not feasible. This approach enables us to perform large-scale data-oriented investigations on how the usage of ODP relates to other social factors such as power relations and gender. Automatically identifying ODP in interactions has many practical applications, that go beyond this thesis. For example, it could help model organizational behavior, aid in email summarization, and even work as a stand-alone email-analytics tool for an end user (as we show in the gSPIN browser extension in Section 7.6).

We use an SVM-based supervised learning approach to build the ODP Tagger. For this task, we want to label each utterance as as either positive (*ODP*) or negative (*NotODP*). We use the ODP annotations described in Section 6.3 to obtain gold labels (*ODP* vs. *NotODP*) for each utterance and built a binary SVM classifier using linguistic, syntactic and dialog structure features.

### 6.4.1 Features

Since linguistic form is an important factor of ODP, we expect lexical and syntactic features to be useful for this prediction task. Also, since ODP is defined also in terms of the underlying communicative intention, we use dialog act features to capture the dialog context of the utterance. For each utterance, we extract five sets of features. The first four sets of features contain lexico-syntactic features extracted using information entirely from the utterance, whereas the last set of features (i.e., DIALOGACT) takes into account the dialog context.

We use three types of n-gram features to capture linguistic and syntactic patterns — lemma n-grams (LEMMANGRAM), part-of-speech n-grams (POSNGRAM), and mixed n-grams (MIXEDNGRAM). Mixed n-gram is a restricted formulation of lemma n-gram where open-class lemmas (nouns, verbs, adjectives and adverbs) are replaced by part-of-speech tags. These n-gram formulations are described in more detail in Chapter 4. The fourth set (FIRSTVERB) contains features referring to the first verb in the utterance. We look at the POS tags assigned to each token in the sentence, and select the first token with a POS tag starting with 'VB' and choose the lemma of that token as a feature. The fifth feature set captures the dialog act of the utterance, which could be one of the following: INFORM, REQUEST-ACTION, REQUEST-INFORMATION, CONVENTIONAL.

Table 6.5 describes these features using the utterance from Example 6.8 — "I need the answer ASAP . . . " as the reference. LEMMANGRAM captures patterns such as {i, need, i need, . . . }, while POSNGRAM captures {PRP, VBP, PRP VBP, . . . } and MIXEDNGRAM captures {i VBP the NN,

. . . }. FIRSTVERB would be 'need' and DIALOGACT would be 'INFORM'.

| Feature Set | Description | Example |
|---|---|---|
| LEMMANGRAM | Lemma N-grams | {i, need, i need, . . . } |
| POSNGRAM | Part-of-speech N-grams | {PRP, VBP, PRP VBP, . . . } |
| MIXEDNGRAM | Mixed N-grams | {i VBP the NN, . . . } |
| FIRSTVERB | First Verb Lemma | need |
| DIALOGACT | Dialog Act | INFORM |

Table 6.5: Features used for ODP prediction.

Feature values are illustrated with respect to the utterance in Example 6.8: "I need the answer ..."

### 6.4.2 Handling Class Imbalance

In its basic formulation, an SVM learns a decision function $f$ from a set of positive and negative training instances such that an unlabeled instance $x$ is labeled as positive if $f(x) > 0$. Since SVM optimize on training set accuracy to learn $f$, it performs better on balanced training sets. However, our dataset is highly imbalanced (around 5% positive instances). Handling the class imbalance problem for SVM is an active area of research, which we discuss in more detail in Section 4.3.4, page 51. We explore two ways of handling this class imbalance issue in this problem:

- *InstWeight*: an instance weighting method where training errors on negative instances are outweighed by errors on positive instances.
- *SigThresh*: a threshold adjusting method to find a better threshold for $f(x)$ to make a positive prediction.

For *InstWeight*, we used the *j* option in SVMlight to set the outweighing factor to be the ratio of negative to positive instances in the training set for each cross validation fold. For *SigThresh*, we used a threshold based on a posterior probabilistic score, $p = Pr(y = 1|x)$, calculated using the ClearTK implementation of Lin et al. (2007)'s algorithm. It uses Platt (1999)'s approximation of $p$ to a sigmoid function $P_{A,B}(f) = (1 + exp(Af + B))^{-1}$, where A and B are estimated from the

training set. Then, we predict $x$ as positive if $p > 0.5$ which in effect shifts the threshold for $f(x)$ to a value based on its distribution on positive and negative training instances.

### 6.4.3 Experiments and Results

We used the ClearTK (Ogren et al. 2008) framework for extracting features and developing the classifier under the Apache UIMA framework. We used ClearTK's built-in tokenizer, part-of-speech tagger, and lemmatizer. We also used the ClearTK wrapper for the SVMLight (Joachims 1999) package to build our models and perform experiments. We use a linear SVM kernel with default values for all other parameters in our experiments, unless specified otherwise.

#### 6.4.3.1 Baseline Systems

We present three baseline approaches. The first two are simple baselines employing no analysis of the training data — ALL-TRUE, where an utterance is always predicted to have an ODP, and RANDOM, where an utterance is predicted at random, with 50% chance to have an ODP. The third one is a strong baseline WORD-UNG, which uses a linear kernel SVM model trained using surface-form words (unigrams) as bag-of-words features. In other words, the WORD-UNG represents a sophisticated machine learning based model, but uses no NLP processing.

#### 6.4.3.2 Results

In this section, we describe results obtained in different experiments. Since our dataset is relatively small, we use 5-fold cross validation to evaluate different experiments. Our folds do not cross thread boundaries. We report precision, recall and F-measure. We calculate F-measure as the harmonic mean of precision and recall. All the results described in this section are obtained using gold dialog act labels from the underlying corpus for both training and testing times. This enables us to study how useful dialog act labels are for the task of predicting overt displays of power. Table 6.6 lists results comparing different feature settings as well as the two class imbalance handling techniques.

**Baseline results:** The first three rows of Table 6.6 show the results obtained by the baseline systems. As expected, the ALL-TRUE and RANDOM baselines obtained very low F scores of 9.5 and 10.4 respectively. The WORD-UNG baseline obtained substantially higher F score of 34.7 under

| | | INSTWEIGHT | | | SIGTHRESH | | |
|---|---|---|---|---|---|---|---|
| | | P | R | F | P | R | F |
| Baselines | ALL-TRUE | 5.0 | 100.0 | 9.5 | 5.0 | 100.0 | 9.5 |
| | RANDOM | 5.7 | 58.1 | 10.4 | 5.7 | 58.1 | 10.4 |
| | WORD-UNG | 43.1 | 29.1 | 34.7 | 63.0 | 39.5 | 48.6 |
| Individual feature sets | LN | 36.6 | 34.9 | 35.7 | 64.3 | 41.9 | 50.7 |
| | PN | 24.4 | 60.5 | 34.8 | 46.4 | 30.2 | 36.6 |
| | MN | 61.0 | 29.1 | 39.4 | 58.7 | 43.0 | 49.7 |
| | FV | 30.5 | 61.6 | 40.8 | 57.6 | 39.5 | 46.9 |
| | DA | 28.4 | 64.0 | 39.3 | 79.5 | 36.1 | 49.6 |
| All feature sets | LN, PN, MN, FV, DA | 66.7 | 46.5 | 54.8 | 69.8 | 51.2 | 59.1 |
| | PN, MN, FV, DA | **66.7** | **48.8** | **56.4** | 72.3 | 54.7 | 62.3 |
| | LN, MN, FV, DA | 72.0 | 41.9 | 52.9 | 77.4 | 47.7 | 59.0 |
| | LN, PN, FV, DA | 47.7 | 60.5 | 53.3 | 72.6 | 52.3 | 60.8 |
| | LN, PN, MN, DA | 66.7 | 46.5 | 54.8 | 73.0 | 53.5 | 61.7 |
| | LN, PN, MN, FV | 64.4 | 44.2 | 52.4 | 65.2 | 50.0 | 56.6 |
| Best feature subset | PN, MN, DA | 64.5 | 46.5 | 54.1 | **75.8** | **58.1** | **65.8** |
| | MN, DA | 74.0 | 43.0 | 54.4 | 71.9 | 53.5 | 61.3 |
| | PN, DA | 29.0 | 62.8 | 39.7 | 80.4 | 43.0 | 56.1 |
| | PN, MN | 56.7 | 39.5 | 46.6 | 55.7 | 45.4 | 50.0 |

Table 6.6: ODP Tagging Results.

INSTWEIGHT: Instance weighting, SIGTHRESH: Sigmoid thresholding

WORD-UNG: Word unigrams, LN: Lemma n-grams, PN: POS n-grams, MN: Mixed n-grams,

FV: First verb, DA: Dialog acts

INSTWEIGHT, and improved it to 48.6 under SIGTHRESH. This shows that the words themselves, without any processing, help greatly in identifying instances of ODP.

**Individual feature sets:** The next section of rows present the results obtained by individual feature sets. Note that the number of features in each of these feature sets vary greatly. For example, LEMMANGRAM include thousands of features, whereas LEMMANGRAM contains only 4 features (one of the four dialog acts). For LEMMANGRAM, POSNGRAM and MIXEDNGRAM, we first found the best value for $n$ to be 1, 2 and 4, respectively, by separate tuning experiments. The results indicate interesting patterns in the utility of these features.

First, let us look at the results for the INSTWEIGHT method. The F-measure for LEMMANGRAM and POSNGRAM did not improve by much compared to the WORD-UNG baseline, but when you look at the precision and recall, they behave differently. Both LEMMANGRAM and POSNGRAM improved recall at the cost of precision, but POSNGRAM improved the recall to almost two times, at a higher cost of precision. MIXEDNGRAM on the other hand, did not improve the recall, but made significant improvement in precision (61.0%, which is the highest precision by any individual feature set). This suggests that MIXEDNGRAM captures patterns that helps identify false positives better. FIRSTVERB and DIALOGACT both resulted in high-recall low-precision models, with DIALOGACT obtaining the best overall recall (64.0%) of all settings (both individual and combinations of feature sets). This suggests that DIALOGACT is valuable in finding false negatives. The best F-measure by an individual feature set was obtained by FIRSTVERB.

The corresponding results for the SIGTHRESH however did not improve by much compared to the WORD-UNG baseline. The LEMMANGRAM improved on both precision and recall marginally and obtained the best F-measure of individual feature sets (50.7%). MIXEDNGRAM improved the recall further, but at the cost of precision. POSNGRAM performed the worst with lowest precision and recall. DIALOGACT obtained the highest precision (79.5%) of all settings (both individual and combinations of feature sets), but at the cost of recall. One thing to note here is that SIGTHRESH is less susceptible to whether the individual feature set is better at preventing false negative or false positives (high-recall or high-precision), since it internally shifts the prediction threshold which will in-turn normalize these differences.

**Using all feature sets:** Using all five feature sets, the F-measure improves significantly for both INSTWEIGHT and SIGTHRESH settings over using any single feature set. INSTWEIGHT obtained the highest precision of 66.7% using all feature sets, with a recall of 46.5% which falls mid-way in the range of recall measures obtained using individual feature sets. The SIGTHRESH method improved on both precision and recall substantially and obtained an F-measure of 59.1%. The next set of rows represents ablation experiments, removing each feature set separately from the model. Removing LEMMANGRAM improved the performance for both INSTWEIGHT and SIGTHRESH. Removing DIALOGACT on the other hand, decreased the performance by a large margin in both cases. Removing FIRSTVERB improved the performance of SIGTHRESH while it did not have any impact on INSTWEIGHT performance.

**Best feature subset:** We then performed experiments using all remaining combinations of LEMMANGRAM, POSNGRAM, MIXEDNGRAM, FIRSTVERB and DIALOGACT under both INSTWEIGHT and SIGTHRESH (total of 31 experiments under each setting). The overall best performing feature set for each setting is highlighted in **boldface**. The combination of POSNGRAM, MIXEDNGRAM, FIRSTVERB and DIALOGACT obtained the best F-measure of 56.4% for INSTWEIGHT. For SIGTHRESH, the best F score of 65.8 was obtained using POSNGRAM, MIXEDNGRAM and DIALOGACT, which was also the overall best performance across all settings and feature combinations. We performed feature set ablation experiments on the best feature set as well, to measure how important each feature set is. POSNGRAM was the least contributing feature; removal of POSNGRAM decreased the F-measure only by 4.5 percentage points. However, MIXEDNGRAM and DIALOGACT were much more important. Removal of MIXEDNGRAM brought down the F-measure by almost 10 percentage points, while removal of DIALOGACT brought down the F-measure by 16 percentage points.

**SIGTHRESH vs. INSTWEIGHT:** In all our experiments, SIGTHRESH obtained better F-measures than INSTWEIGHT. We did perform experiments combining these two techniques for dealing with class imbalance. However they gave worse performance than using either one alone. We conclude that SIGTHRESH is the right approach for our task.

### 6.4.4 Post-hoc Analysis

One of the criticisms often raised about SVM-based approaches is that the models are not interpretable. However, since we use a linear kernel for training, the weights assigned for each feature will denote the strength and direction of their relation with the class that is being predicted (in our case, ODP). This approach of inspecting feature weights was initially proposed by (Guyon et al. 2002). We used the model created for the last fold of the cross validation experiment of our best performing feature set for this analysis. Table 6.7 shows the top 10 positive and negative weighted features in the trained model.

| Positive Weighted Features | | Negative Weighted Features | |
|---|---|---|---|
| DialogAct_Request-Action | 2.5 | DialogAct_Inform | -1.4 |
| MixedNGram:you_VB | 1.0 | DialogAct_Conventional | -1.0 |
| PosNGram:_BOS__VB | 0.9 | PosNGram:MD_VB | -0.6 |
| PosNGram:MD_PRP | 0.9 | MixedNGram:VB_you | -0.5 |
| PosNGram:VB_VB | 0.8 | MixedNGram:what | -0.5 |
| PosNGram:_BOS__MD | 0.7 | MixedNGram:VB_VB_me_VB | -0.5 |
| MixedNGram:can_you | 0.6 | MixedNGram:we_VB | -0.4 |
| MixedNGram:. | 0.6 | PosNGram:,__EOS_ | -0.4 |
| MixedNGram:NN_for | 0.6 | MixedNGram:you_VBP | -0.4 |
| MixedNGram:for | 0.5 | PosNGram:WP | -0.4 |

Table 6.7: Post-hoc analysis of ODP trained models.

The feature *DA:RequestAction* got the highest positive weight of 2.5. The other interesting positive weighted features include patterns like *you_VB* (*you* followed by a verb in its base form), *_BOS__VB* (utterances beginning with a verb), *MD_PRP* (a modal verb followed by a personal pronoun), *VB_VB* (a verb following another verb) and *_BOS__MD* (utterances beginning with a modal verb), where _BOS_ denotes the beginning of sentence. *DA:Inform* got the most negative weight of -1.4, followed by *DA:Conventional* with -1.0. The other interesting top ten negative weighted features include patterns like *MD_VB* (a verb following a modal verb), *VB_you* (a verb followed by *you*), *what* (the word *what*), *VB_VB_me_VB* and *WP* (WH-pronouns). In both cases, the top DA fea-

tures got almost 2.5 times higher weight than the highest weighted n-gram pattern, which reaffirms their importance in this task, as was seen in the results. Also, mixed n-grams helped to capture long patterns like "please let me know" by *VB_VB_me_VB* (the POS tagger tagged "please" to be a verb) without increasing dimensionality as much as lemma n-grams. They also distinguish one of the top positive weighted feature *you_VB* (1.0) from one of the top negative weighted feature *we_VB* (-0.4), which pure part-of-speech n-grams couldn't have been able to.

### 6.4.5 Experiments using Automatically Obtained Dialog Act Tags

Since the ODP annotations are obtained on top of the existing dialog act annotations (Hu et al. 2009), we were able to use the gold dialog act labels as features for our system. Using the gold dialog act labels enabled us to study how useful they are for the task of predicting overt displays of power, as we did in the previous section. However, a completely automatic ODP tagger will not have access to gold dialog act tags in unseen sentences, and hence the performance obtained by a system that uses gold dialog act labels in order make a prediction is not representative of the ODP tagger's end-to-end performance. Hence we also perform experiments using automatically obtained dialog act tags.

Although at prediction time, we have to restrict ourselves to automatically obtained dialog act labels, we have the option of training the models with either gold or automatically obtained dialog act labels. Previous studies have shown that in such scenarios, it is better to use automatically obtained labels to train the models (Marton et al. 2013), so that the resulting models are trained on features that are more closer to the features available at prediction time. Since our experiments are done in a cross validation set up, and because the same data contains the gold annotations for dialog acts and ODP, we cannot use a dialog act tagger trained on the entire corpus to obtain the dialog act labels. We have to ensure that the test folds for ODP were excluded from training the taggers to obtain DA tags. Hence, in each ODP cross validation step, we retrained a DA tagger using DA annotations present in the training folds for that step and then used the tags produced by that tagger for both training and testing the ODP tagger for that step.

We use the best performing configuration from Table 6.6 as the reference for this set of experiments. Table 6.8 summarizes the results obtained in our experiments. The first row repeats the best performing setting from Table 6.6 using gold dialog act labels for comparison. If we drop the gold

| | Precision | Recall | F-measure |
|---|---|---|---|
| BEST (using gold DIALOGACT) | 75.8 | 58.1 | 65.8 |
| BEST minus DA | 55.7 | 45.4 | 50.0 |
| BEST using BAS DA | 60.6 | 46.5 | 52.6 |
| BEST using DAC-CMP DA | 67.2 | 45.4 | 54.2 |

Table 6.8: Results for ODP tagger using different sources of DA tags.

DIALOGACT features, the F-measure drops to 50.0. We compared two different approaches of automatic dialog act tagging — the baseline one-vs.-all method (BAS) and the DAC-CMP methods described in Chapter 5. Using BAS tagged DIALOGACT, the F-measure of ODP system reduced by 13.2 points to 52.6 from using gold dialog acts (F=65.8). Using DAC-CMP, the F-measure improved over BAS by 1.6 points to 54.2. This constitutes an error reduction of 12.1% over using the standard one-vs.-all SVM to generate dialog act tags.

## 6.5 Conclusion

In this chapter, we introduced the notion of overt display of power (ODP) in interactions. We defined an overt display of power as an utterance that adds constraints on the possible responses from its addressees. We presented a corpus of 122 email threads annotated with instances of overt displays of power and described the annotation procedure in detail. Our annotations obtained a reasonable inter-annotator agreement of $\kappa = 0.67$. We also presented an overt display of power tagger that obtains an F-measure of 66% despite the fact that ODPs are very rare in the corpus.

We found that the dialog act features are the most useful features for detecting overt displays of power, followed by the mixed ngrams. We obtained an overall best F-measure using a combination of part-of-speech ngrams, mixed-ngrams and dialog act features. We also experimented with two different methods to handle the class imbalance problem (our positive class is only around 5% of the data). We found that using a post-processing step of shifting the prediction threshold by using sigmoid fitting works better to handle this issue than the instance weighting approach where training errors on positive instances are penalized more.

# Part III

# MANIFESTATIONS OF POWER IN DIALOG

# Chapter 7

# Hierarchical Power in Organizational Email

In this chapter, we will present our study on manifestations of hierarchical power relations in the domain of organizational interactions. In the field of studying manifestations of power in interactions, the domain of organizational interactions has generated considerable research interest. For instance, a significant number of computational studies of power in interactions are performed in the domain of organizational email (Shetty and Adibi 2005, Diesner and Carley 2005, Rowe et al. 2007, Creamer et al. 2009, Bramsen et al. 2011, Gilbert 2012). This is in part due to the fact that this domain has a well-defined notion of power, i.e., organizational hierarchy. Another reason that triggered substantial research in this domain is the availability of the Enron email corpus which contains a large collection of real-world organizational interactions that occurred over a span of a few years.

Early computational approaches relied on social network analysis based purely on email metadata (i.e., without looking at the email content) (Shetty and Adibi 2005, Diesner and Carley 2005, Rowe et al. 2007, Creamer et al. 2009), whereas more recent approaches such as (Bramsen et al. 2011, Gilbert 2012) have shown that the content of the email messages also holds important clues about power relations. Both Bramsen et al. (2011) and Gilbert (2012) predict hierarchical power relations between people in the Enron email corpus using lexical features extracted from the set of all messages exchanged between them. However, their approaches primarily apply to situations

where large collections of such messages exchanged between pairs of people are available. Also, they do not use the context in which each email was exchanged. By context, we mean the interaction as part of which an email was sent. We hypothesize that the dialog context of an interaction will reveal important clues about power relations that exist between its participants. In this chapter, we look beyond the content of the emails, for patterns that also capture their dialog context. We show that power is manifested in the structure of the interactions as well as in the language used, and that these structural and linguistic patterns can help infer power relations between participants.

We start with a motivating example in Section 7.1, before formally describing the data and terminology in Section 7.2. Section 7.3 discusses the different features we use in this study and Section 7.4 presents the results of a statistical analysis we performed on how power is manifested among those features. Section 7.5 presents an automatic system to predict the direction of power between pairs of people and Section 7.6 presents an end-to-end demonstration of the power predictor system: a Google Chrome browser plugin called gSPIN. In Section 7.7, we summarize the work and discuss limitations and future work.

## 7.1 A Motivating Example

Let us start by looking at an email thread from our corpus of Enron email threads in order to motivate the rest of this chapter. Table 7.1 presents an abridged version of an email thread discussing the status of an ISDA agreement with the City of Glendale.[1] The only "abridging" done to the thread was by omitting parts of the first email message, which was too long (more than 200 words) to include entirely. The parts that are omitted are indicated by '[...]' in the table. The email thread contains nine email messages, and four participants: Kim S Ward, Sara Shackleton, Marie Heard, and Steve Lins; the first three actively participate in the thread (i.e., each of them sends at least one email message) while the last one is a silent participant (i.e., does not send any messages, but receives at least one one). Each email message is given a unique identifier M1-M9. The email identifiers follow the chronological ordering of the messages. The date, time, sender, and recipient information of each message is also shown in the header line. In the table, the email thread is arranged in the table

[1]ISDA (International Swaps and Derivatives Association) agreement is used between a derivatives dealer (Enron, in this case) when discussions begin concerning a derivatives trade.

**M1 — 05 Oct 2001 2:59 PM; From: Kim S Ward; To: Sara Shackleton;**

Sara,

Believe it or not, we are very close getting our signed ISDA from the City of Glendale. Steve Lins, the City attorney had a couple of questions which I will attempt to relay without having a copy of the documents.

1) I am assuming that he obtained a for legal opinion letter or document of some sort. [...] What is your opinion regarding this?

2) We sent him a couple of form documents to facilitate the documents required under the ISDA. [...] Will this suffice?

When you return, I may try to do one last conference call [...]

Thanks for your help,

**M2 — 08 Oct 2001 9:02 AM; From: Sara Shackleton; To: Kim S Ward; CC: Marie Heard;**

Kim: Can you obtain the name of Glendale's bond counsel (lawyer's name, phone number, email, etc.)?

Thanks. SS

**M3 — 08 Oct 2001 9:26 AM; From: Kim S Ward; To: Sara Shackleton;**

Glendale's City Attorney is Steve Lins. His phone number is 818-548-2080 and his email is slins@ci.glendale.ca.us. Please let me know if you need anything else. I will be in their offices on Wednesday.

**M4 — 08 Oct 2001 9:27 AM; From: Sara Shackleton; To: Kim S Ward;**

I need the city's bond counsel (outside counsel).

**M5 — 08 Oct 2001 10:03 AM; From: Kim S Ward; To: Sara Shackleton;**

Is this to obtain outside opinion? I thought we were going to do that at our own expense.

**M7 — 08 Oct 2001 10:38 AM; From: Sara Shackleton; To: Kim S Ward;**

We are going to do this at our own expense. But we would like to hire Glendale's bond counsel. I don't know the name of Glendale's bond counsel or how to get in touch with them.

**M8 — 08 Oct 2001 11:43 AM; From: Kim S Ward; To: Sara Shackleton;**

I will work on this for you - and will be in touch. Thanks!

**M6 — 08 Oct 2001 10:15 AM; From: Marie Heard; To: Sara Shackleton;**

Sara: I do not see a copy of an opinion in the file nor have we received one since I sent the execution copies of the ISDA to Steve Lins.

**M9 — 08 Oct 2001 4:18 PM; From: Kim S Ward; To: Steve Lins;**

Steve, could you provide the name, phone number, etc. of your bond council for our attorney, Sara Shackleton? Thanks,

Table 7.1: Example email thread from the Enron email corpus. Subject line: "City of Glendale".

in such a way that the thread structure is easier to follow. The indentation denotes the depth of the message within the message tree. Each message is in response to the message with the next smallest indentation above it (either as a reply to the latter message or forwarding it); e.g., M3, M6 and M9 are in response to M2, and M4 is in response to M3. It is important to note that the messages are not arranged chronologically; e.g., M6 was sent after M5 and before M7, chronologically, but is listed after M7 and M8 in the table in order to display the thread structure.

Kim initiates the email thread by sending an email (M1) to Sara detailing the updates on the ISDA agreement signing with the City of Glendale. She also lists two questions the city attorney Steve Lins had about the agreement, and asks Sara's feedback on both. Sara responds to Kim's email adding Marie to the conversation (M2). She ignores Kim's questions, and asks her to provide information about Glendale's bond counsel. Kim and Sara goes on to exchange few emails back and forth (M3, M4, M5, M7, and M8) until Kim understands what Sara is asking for and why. Finally, Kim forwards Sara's request to Steve Lins to obtain the required information (M9). Meanwhile, Marie responds to Sara's email (M6) giving information about the status of ISDA from her end.

On careful observation of the interaction in this email thread, one can infer that it is likely the case that Sara has power over Kim. There are many indicators that lead to this inference. In M2, Sara ignored all the questions raised by Kim in M1, and asks Kim to obtain some specific information. Sara does not provide any explanation for why she needs that information; it becomes clear to Kim only after she asks clarification questions. In M4, Sara uses an overt display of power (see Chapter 6) in the sentence *I need the city's bond counsel* when the information Kim provided in M3 was not what she requested for. Sara's response in M4 (*I need the city's bond counsel*) is an instance of overt display of power (Chapter 6). Kim, on the other hand, appears to be following orders from Sara, although she does raise clarification questions when necessary, and finally commits to *will work on this for you*, and in fact relays Sara's question to Steve in order to obtain the information Sara needs. On the other hand, the relationship between Sara and Marie is hard to judge, since there aren't any clear indicators that suggest that one has power over the other.

The question we ask in this thesis is whether a computational system can be trained to automatically detect the power relations between pairs of participants based solely on single threads of interaction. Such a computational system will help detect such linguistic and dialogic patterns that are indicative of power relations. We start by discussing the data we use to build this system.

## 7.2 Data and Terminology

### 7.2.1 Enron Email Threads

For the work presented in this chapter, we use the ENRON-LARGE corpus described in Chapter 3, page 36. The ENRON-LARGE is the version of Enron email corpus put together by Yeh and Harnly (2006). They reconstructed the thread structure of email messages in the original Enron corpus, by restoring the missing messages from other emails in which they were quoted. Most email threads are concerned with exchanging information, scheduling meetings, and solving problems, but there are also purely social emails. There were around three messages per thread on average. More details about this corpus is given in Section 3.1, page 36. The corpus contains 36,196 email threads. We divide the threads in the corpus to *train* (50%), *dev* (25%) and *test* (25%) sets by random sampling. The first row of Table 7.2 shows the number of threads in each set.

Let $t$ denote an email thread and $M_t$ denote the set of all messages in $t$. Each message $m \in M_t$ has one sender and one or more recipients. Recipients of a message include those to whom the message is addressed (*To* list) as well as those to whom the message is carbon-copied (*CC* list). Let $P_t$ be the set of all participants in $t$, i.e., the union of senders and recipients of all messages in $M_t$. We are interested in analyzing the power relations between pairs of participants who interact within the email thread $t$. Not every pair of participants $(p_1, p_2) \in P_t \times P_t$ interact with one another within $t$. If a pair of participants have no common email message that they are part of (as a sender or as one of the recipients), then they are considered to be not interacting within the thread; e.g Sara and Steven do not interact in the example email thread given in Table 7.1, nor do Marie and Steven. Being co-recipients of the same email also is not sufficient for a pair to be considered as interacting; e.g., Kim and Marie are not interacting within the thread. Let $IM_t(p_1, p_2)$ denote the set of *Interaction Messages* of the pair $(p_1, p_2)$ — 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 $t$ ($IPP_t$). i.e.,

$$IPP_t = \{ (p_1, p_2) \mid |IM_t(p_1, p_2)| > 0 \}$$

The second row of Table 7.2 shows the total number of interacting participant pairs in $IPP_t$ in all the threads in our corpus and across *train*, *dev* and *test* sets.

### 7.2.2 Enron Organizational Hierarchy

We use the Enron gold organizational hierarchy released by Agarwal et al. (2012) to model hierarchical power relations between participants of email threads. Their corpus was manually built using information from the Enron organizational charts. It contains relations of 1,518 employees and captures 13,724 dominance pairs among them. They define the dominance pairs as "pairs of employees such that the first dominates the second in the hierarchy, not necessarily immediately". Theirs is the largest such data set available to the best of our knowledge.

Let $DomPairs$ denote the set of dominance pairs that contains pairs of Enron employees $(p, q)$ such that $p$ dominates $q$ as per the gold organizational hierarchy by **?**. Also, let $PeopleInHierarchy$ denote the set of Enron employees who are part of the gold hierarchy; i.e., union of all $p$, $q$ for all $(p, q) \in DomPairs$. For each thread $t$, for each $(p_1, p_2) \in IPP_t$, we assign their hierarchical power relation $HP(p_1, p_2)$ as follows:

$$HP(p_1, p_2) = \begin{cases} superior & \text{if } (p_1, p_2) \in DomPairs \\ subordinate & \text{if } (p_2, p_1) \in DomPairs \\ neither & \text{if } (p_1, p_2) \notin DomPairs \text{ and } (p_2, p_1) \notin DomPairs, \\ & \text{but } p_1, p_2 \in PeopleInHierarchy \\ unknown & \text{if } p_1 \notin PeopleInHierarchy \text{ or } p_2 \notin PeopleInHierarchy \end{cases}$$

The four cases are described below.

- *superior*: $p_1$ is above $p_2$ in the hierarchy
- *superior*: $p_1$ is below $p_2$ in the hierarchy
- *neither*: $p_1$ and $p_2$ are not related by a superior/subordinate relation as per the hierarchy
- *unknown*: $p_1$ or $p_2$ is not captured in the hierarchy, hence we do not know their relation

The first two cases—*superior* and *subordinate*—are the most contrasting relations. The interacting participant pairs that belong to one of these cases are the ones which we reliably know to be hierarchically related. We call the set of such pairs in thread $t$ the *related interacting participant pairs* of $t$ (denoted $RIPP_t$). The *neither* pairs are the ones who we reliably know are not guided

| Description | Total | Train | Dev | Test |
|---|---|---|---|---|
| # of threads | 36,196 | 18,079 | 8,973 | 9,144 |
| $\sum_t \lvert IPP_t \rvert$ | 355,797 | 174,892 | 91,898 | 89,007 |
| $\sum_t \lvert RIPP_t \rvert$ | 15,048 | 7,510 | 3,578 | 3,960 |

Table 7.2: Enron email corpus: 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 them.

by a superior-subordinate relation. Both participants of these pairs take part in some of the dominance relations captured by $DomPairs$ and hence they are covered in the organizational hierarchy that $DomPairs$ stands to represent, but the pairs themselves are not present in $DomPairs$. This means that these pairs may either be employees at the same level (peers) or in completely different sub-divisions/branches of the organization. We call the set of pairs in thread $t$ that are covered in the $DomPairs$ hierarchy as the *covered interacting participant pairs* of $t$ (denoted $CIPP_t$). The rest of the pairs, i.e., the $unknown$ pairs include cases where the entities may or may not be hierarchically related, but we do not have a way to reliably determine this. Remember that $DomPairs$ captures relations between 1518 of the Enron employees, whereas our corpus contains around 25K Enron employees. Hence, we exclude the $unknown$ pairs from our study. In this chapter, we focus mainly on the pairs in $RIPP_t$ since we are more interested in the differences between superiors and subordinates, than situations where such a power relation does not exist. The third row of Table 7.2 shows the total number of related interacting participant pairs in $RIPP_t$ in all the threads in our corpus and across *train*, *dev* and *test* sets.

### 7.2.3 Preprocessing

We use the UIMA architecture and the ClearTK suite of UIMA tools (Ogren et al. 2008) to build the computational framework required for this study. We applied various basic NLP preprocessing steps to the content of the email messages to enable downstream processing. We used the default ClearTk tokenizer, part-of-speech tagger and lemmatizer for this purpose. We then apply the dialog act tagger

described in Chapter 5 to each sentence in the email messages. The dialog act tagger assigns one of the four dialog act labels — REQUEST-ACTION, REQUEST-INFORMATION, CONVENTIONAL and INFORM— to each sentence. Following that, we apply the automatic tagger to detect overt displays of power (ODP Tagger) described in Chapter 6, also at the sentence level.

## 7.3 Features

In this section we describe various features we use to model the aspects of dialog behavior of the participants in an email thread. We focus on features in six different aspects of interactions — POSITIONAL, VERBOSITY, THREAD STRUCTURE, DIALOG ACTS, OVERT DISPLAY OF POWER, and LEXICAL. The first three aspects (POSITIONAL, VERBOSITY, and THREAD STRUCTURE) capture the structure of message exchanges without doing any NLP processing on the content of the emails (e.g., how many emails did a person send), whereas DIALOG ACTS and OVERT DISPLAY OF POWER capture the pragmatics of the dialog and requires an analysis of the content of the emails (e.g., did they issue any requests). LEXICAL features also analyze the content, but at a shallow level, looking solely at word lemma and part-of-speech ngrams.

Each feature $f$ is extracted with respect to a person $p$ over a reference set of messages $M$ (denoted $f^p_M$). For example, $MsgRatio^{Kim}_{M_t}$ denotes the ratio of messages sent by ***Kim*** to the total number of messages in the thread $t$, whereas $MsgRatio^{Sara}_{IM_t(Kim,Sara)}$ denotes the ratio of messages sent by ***Sara*** to the total number of interaction messages between ***Kim*** and ***Sara*** in the thread $t$. For each pair $(p_1, p_2)$, we extract 4 versions of each feature $f$.

$f^{p_1}_{IM_t(p_1,p_2)}$: features with respect to $p1$ and interaction messages between $p1$ and $p2$

$f^{p_2}_{IM_t(p_1,p_2)}$: features with respect to $p2$ and interaction messages between $p1$ and $p2$

$f^{p_1}_{M_t}$: features with respect to $p1$ and all messages in thread $t$

$f^{p_2}_{M_t}$: features with respect to $p2$ and all messages in thread $t$

The first two versions capture behavior of the pair among themselves, while the third and fourth capture their overall behavior in the entire thread. In Table 7.3, we list each feature $f$ we use in this chapter.

| Aspects | Features | Description |
|---|---|---|
| PST | *Initiator* | did $p$ sent the first message? |
| | *FirstMsgPos* | relative position of $p$'s first message in $M$ |
| | *LastMsgPos* | relative position of $p$'s last message in $M$ |
| VRB | *MsgCount* | Count of messages sent by $p$ in $M$ |
| | *MsgRatio* | Ratio of messages sent in $M$ |
| | *TokenCount* | Count of tokens in messages sent by $p$ in $M$ |
| | *TokenRatio* | Ratio of tokens across all messages in $M$ |
| | *TokenPerMsg* | Number of tokens per message in messages sent by $p$ in $M$ |
| THR | *AvgRecipients* | Avge. number of recipients in messages |
| | *AvgToRecipients* | Avge. number of To recipients in messages |
| | *InToList%* | % of emails p received in which he/she was in the To list |
| | *AddPerson* | did $p$ add people to the thread? |
| | *RemovePerson* | did $p$ remove people to the thread? |
| | *ReplyRate* | average number of replies received per message by $p$ |
| | *ReplyRateWithinPair* | *ReplyRate* from the other person of the pair |
| DA | *ReqAction%* | % of Request Action dialog acts in $p$'s messages |
| | *ReqInform%* | % of Request Information dialog acts in $p$'s messages |
| | *Inform%* | % of Inform dialog acts in $p$'s messages |
| | *Conventional%* | % of Conventional dialog acts in $p$'s messages |
| | *DanglingReq%* | % of messages with requests sent by $p$ that did not have a reply |
| ODP | *ODPCount* | Number of instances of overt displays of power |
| LEX | *LemmaNGram* | Word lemma ngrams |
| | *POSNGram* | Part of speech (POS) ngrams |
| | *MixedNGram* | POSNGrams, with closed classes replaced with lemmas |

Table 7.3: Aspects of interactions analyzed in organizational emails.

### 7.3.1 Positional Features

There are three features in this category — *Initiator*, *FirstMsgPos*, and *LastMsgPos*. *Initiator* is a boolean feature which gets the value of 1 (*true*) if the $p$ sent the first message in the thread, and 0 otherwise (*false*). *FirstMsgPos*, and *LastMsgPos* are real-valued features taking values from 0 to 1, capturing relative positions of $p$'s first and last messages. The lower the value, the earlier the participant sent his/her first (or last) message. The first two features relate to the participant's initiative. *LastMsgPos* capture whether the participant stays till the end of the email thread.

### 7.3.2 Verbosity Features

This set of features captures how verbose were the participants in the thread. There are five features in this set — *MsgCount*, *MsgRatio*, *TokenCount*, *TokenRatio*, and *TokenPerMsg*. The first two features measure verbosity in terms of $p$'s messages (raw counts and percentages), whereas the third and fourth features measure verbosity in terms of word tokens in $p$'s messages (raw counts and percentage). The last feature measure how terse or verbose on average were $p$'s messages.

### 7.3.3 Thread Structure Features

This set of features captures the structure of the email in terms of meta-data that is part of the email headers. It includes seven features — *AvgRecipients*, *AvgToRecipients*, *InToList%*, *AddPerson*, *RemovePerson*, *ReplyRate*, and *ReplyRateWithinPair*. The first two features capture the 'reach' of the person in terms of the average number of total recipients as well as recipients in the To list in emails sent by $p$. *InToList%* capture the the percentage of emails $p$ received in which he/she was in the *To* list (as opposed to the *CC* list); The next two features —*AddPerson* and *RemovePerson*— are boolean features denoting whether $p$ added or removed people when responding to a message. Next, we look at the responsiveness towards $p$ as the 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 the reference set of messages $IM_t(p_1, p_2)$.

### 7.3.4 Dialog Act Features

This feature set contains features that capture the dialog acts used by participants in the thread. The DA tagger labels each sentence to be one of the 4 dialog acts: REQUEST-ACTION, REQUEST-INFORMATION, INFORM, and CONVENTIONAL. Correspondingly, we use 4 features: *ReqAction%*, *ReqInform%*, *Inform%*, and *Conventional%* to capture the percentage of sentences in messages sent by $p$ that has each of these labels, respectively. We also use a feature to capture the percentage of $p$'s messages that had a request (either REQUEST-ACTION or REQUEST-INFORMATION), which did not get a reply, i.e., dangling requests (*DanglingReq%*).

### 7.3.5 Overt Displays of Power

This feature set is a singleton set that captures instances of overt displays of power in $p$'s messages. We apply the ODP Tagger described in Chapter 6 to the email threads in our corpus. The ODP tagger identifies sentences (mostly requests) that express additional constraints on its response, beyond those introduced by the dialog act. We use a feature *ODP%* to capture the percentage of sentences in messages sent by $p$ that was assigned an overt display of power label.

### 7.3.6 Lexical Features

Lexical features have already been shown to be valuable in predicting power relations (Bramsen et al. 2011, Gilbert 2012). We use the feature set LEXICAL to capture word lemma ngrams, POS (part of speech) ngrams and mixed ngrams. A mixed ngram is a special case of word ngram where words belonging to open classes are replaced with their POS tags, thereby being able to capture longer sequences without increasing the dimensionality as much as word ngrams do. We found the best setting to be using both unigrams and bigrams for all three types of ngrams, by tuning in our *dev* set.

## 7.4 Superiors vs. Subordinates: A Statistical Analysis

As a first step towards understanding how the dialog behaviors of superiors and subordinates differ, we perform an unpaired two-sample two-tailed Student's t-Test comparing the mean values of each

| Aspects | Features | **Mean**($f^{sub}_{IM_t}$) | **Mean**($f^{sup}_{IM_t}$) |
|---|---|---|---|
| POSITIONAL | Initiator*** | 0.45 | 0.56 |
| | FirstMsgPos | 0.04 | 0.03 |
| | LastMsgPos*** | 0.15 | 0.11 |
| VERBOSITY | MsgCount*** | 0.64 | 0.70 |
| | MsgRatio*** | 0.44 | 0.56 |
| | TokenCount | 91.22 | 83.26 |
| | TokenRatio*** | 0.45 | 0.55 |
| | TokenPerMsg* | 140.60 | 120.87 |
| THREAD STRUCTURE | AvgRecipients*** | 21.14 | 43.10 |
| | AvgToRecipients*** | 18.19 | 38.94 |
| | InToList% | 0.82 | 0.80 |
| | ReplyRate*** | 0.86 | 1.23 |
| | ReplyRateWithinPair*** | 0.16 | 0.10 |
| | AddPerson | 0.48 | 0.47 |
| | RemovePerson*** | 0.41 | 0.37 |
| DIALOG ACTS | ReqAction%*** | 0.02 | 0.04 |
| | ReqInform%*** | 0.05 | 0.06 |
| | Inform%*** | 0.78 | 0.72 |
| | Conventional%*** | 0.15 | 0.17 |
| | DanglingReq%*** | 0.12 | 0.15 |
| OVERT DISPLAY OF POWER | ODP%*** | 0.03 | 0.06 |

Table 7.4: Student's t-Test results comparing mean values of $f^{p}_{IM_t}$ of *Superiors* vs. *Subordinates*

* ($p < .05$); ** ($p < .01$); *** ($p < .001$)

feature for *subordinates* and *superiors*. We perform this analysis for all features other than LEXICAL. We assess significance at three different levels — significant* ($p < .05$), highly significant** ($p < .01$), and very highly significant*** ($p < .001$). Since we perform a large number of statistical tests in this analysis, to control for false discovery rates, we performed multiple-test correction on the p-values (Benjamini and Hochberg 1995). We use the related interacting participant pairs from our training set for this analysis.

### 7.4.1 Findings

Table 7.4 presents the results of the Student's t-Test. Columns 3 and 4 denote the mean values of features for *subordinates* and *superiors*, respectively, at the interaction level. The statistical significance level of the difference between mean values is denoted by asterisks(*) next to the feature name. Thread level versions of these features also obtained similar results overall in terms of direction of difference and significance.

In order to assess the magnitude of the difference between each groups, we calculated the relative difference between the mean values as follows:

$$RelativeDifference = \frac{Mean(f^{sup}_{IM_t}) - Mean(f^{sub}_{IM_t})}{Mean(f^{sub}_{IM_t})}$$

Figure 7.1 shows the relative difference for each feature. Dark bars indicate statistically significant differences, whereas light bars indicate features for which the difference was not significant. A dark bar to the right of the y-axis means superiors have significantly higher value for the corresponding feature, whereas a dark bar to the left means subordinates have significantly higher value for the corresponding feature. The length of the bar indicates the magnitude of the difference.

### 7.4.2 Discussion

As can be seen from Table 7.4 and Figure 7.1, many of the features (17 out of the 21) we analyzed were significantly different for *Superiors* and *Subordinates*. We discuss these findings in detail below for each set of features separately.

**POSITIONAL Features** Superiors initiated the threads significantly ($p < 0.001$) more often than subordinates (*Initiator*). However, in terms of the relative position of the first message (*FirstMsgPos*), the difference was not significant between superiors and subordinates. In other words, what

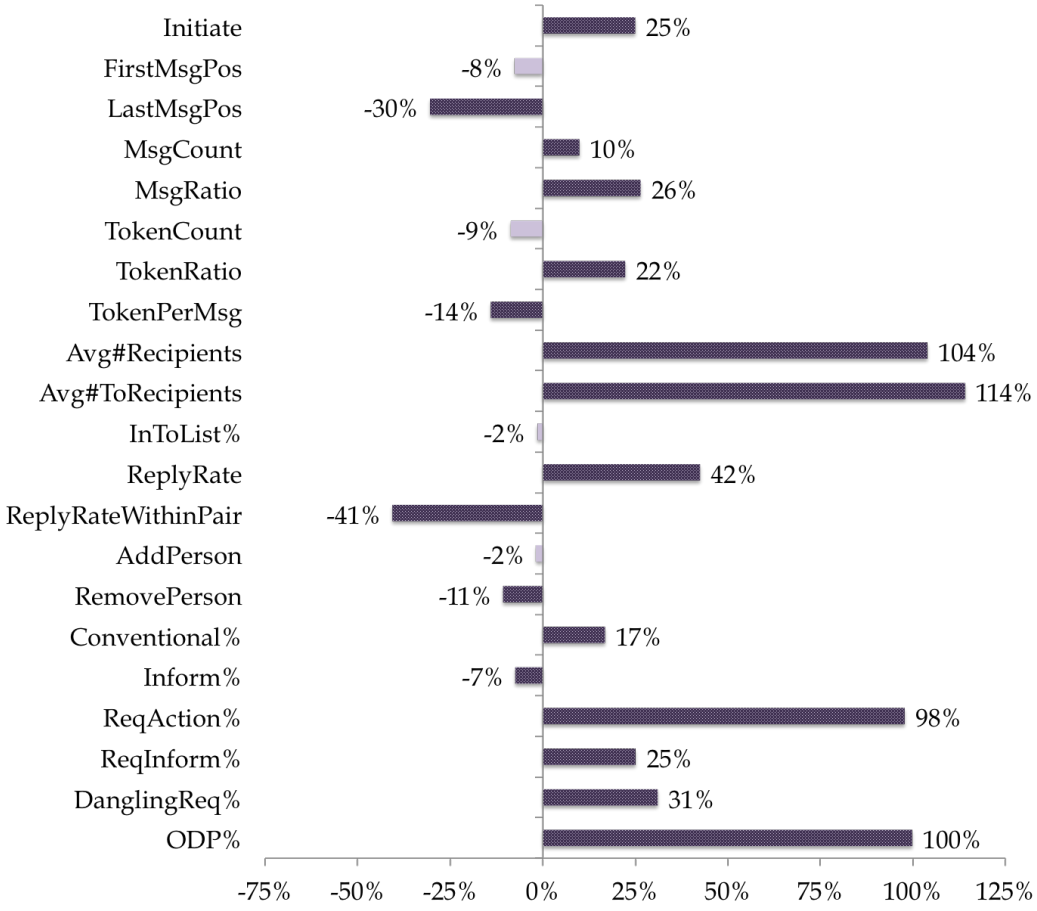



Figure 7.1: Relative difference of feature values between *Superiors* and *Subordinates*.

matters is whether a participant initiates the thread, and not whether the participant started participating earlier in the thread. On the other hand, superiors tend to leave the thread significantly ($p < 0.001$) earlier than subordinates (*LastMsgPos*).

**VERBOSITY Features** Superiors send significantly more ($p < 0.001$) messages (*MsgCount* & *MsgRatio*) in the thread. However, their messages were significantly shorter ($p < 0.05$) than those of subordinates (*TokenPerMsg*). In other words, superiors contribute more in the threads when interacting with subordinates, but with shorter messages. In terms of the absolute count of word tokens in their messages, there was no significant difference between superiors and subordinates. As Figure 7.1 shows, mean value of *MsgRatio* for superiors is almost 25% more than that of subordinates,

whereas the relative difference (in the opposite direction) in *TokenPerMsg* was only half as much. Hence, in terms of the ratio of word tokens to the total number of tokens exchanged in the thread, superiors still made a larger contribution.

**THREAD STRUCTURE Features** Out of the seven thread structure features we study, we found five of the features to be significantly different between superiors and subordinates. Superiors send messages addressed to significantly more people (*AvgRecipients* and *AvgToRecipients*). In fact, the mean values of both of these features for superiors are more than twice (relative difference more than 100%) than that of subordinates. This finding goes in line with findings from studies analyzing social networks that superiors have higher connectivity in the networks that they are part of (Rowe et al. 2007, Agarwal et al. 2014). Intuitively, those who have higher connectivity also send emails to larger number of people. Superiors also get significantly more replies to their emails than subordinates (*ReplyRate*). However, considering messages where the other person of the pair is addressed in the *To* list (*ReplyRateWithinPair*), subordinates get significantly more replies. On further analysis, we found that this is because many of superiors' messages are broadcast messages that do not require all recipients to respond. Since these messages have a large number of recipients, the superior gets replies to these messages from some of the recipients, resulting in a larger value for (*ReplyRate*) on average for superiors. However, when subordinates send an email addressing the superior, it is mostly for a purpose that aligns with a task being performed, and hence require responses.

**DIALOG ACTS Features** We found all the dialog act features we used to be significantly different for superiors and subordinates. All of these differences were very highly significant. As expected, superiors issue significantly more request (*ReqAction%* and *ReqInform%*) than subordinates. It is interesting to note that superiors issue about twice as many requests for actions than subordinates, whereas the relative difference was less than 25% for requests for information. This shows the importance of distinguishing these two dialog act labels. Subordinates, on the other hand, use significantly more of INFORM sentences. Although the magnitude of difference is not huge, this difference was very highly significant ($p < 0.001$). A counter-intuitive result here is in terms of the *DanglingReq%* feature. Superiors had a higher percentage of their request-containing messages that went without any replies. This might be because superiors issue more requests for actions that

need to be performed outside the conversation, and hence may not receive a reply. For example, if a superior sends an email to her subordinate saying *Please come to my office*, the subordinate may just show up at her office without replying to the email saying *Yes, I am coming.*

**OVERT DISPLAY OF POWER Features** We also found that superiors use almost twice as many overt displays of power in their messages (*ODP%*) than subordinates. This result is also very highly significant. This finding is not surprising, since it aligns with the general intuition about how superiors and subordinates behave within interactions. However, it is important to note that the notion of ODP is defined independent of power relations between participants. The ODP labels were obtained using a tagger trained on annotations that looked at the linguistic form and the dialog context alone to judge whether a sentence has an overt display or not. The annotators did not have access to the power relations between the participants. In other words, what we find here is not a cyclical effect of how ODP is defined.

## 7.5 Predicting Power Relations

In this section, we describe an automatic system that can distinguish superiors and subordinates based on their dialog behavior. Like prior work in this area (Bramsen et al. 2011, Gilbert 2012), we also focus on the problem of detecting the direction of power (*superior* vs. *subordinate*) of related interacting participant pairs.

The problem of distinguishing superiors and subordinates in the related interacting participant pairs of an email thread is a binary classification task. We use a supervised learning approach to solve this. For a given email thread $t$ and $(p_1, p_2) \in RIPP_t$, we would like the system to automatically predict $HP(p_1, p_2)$ to be either *superior* or *subordinate*. We use the Support Vector Machine (SVM) algorithm to build the classifier. We use the ClearTK (Ogren et al. 2008) wrapper for SVMLight (Joachims 1999) to perform our experiments and build the final model. We performed experiments using all the features we described in Section 7.3 to build the system.

### 7.5.1 Fixing the Order of Participants

In the terminology we described in Section 7.2, the order of the participants in a pair is arbitrary. However, for the prediction task and for the machine learning algorithm, it helps to make the prob-

lem uniform. Hence we remove the arbitrariness of the ordering of the pair by deterministically fixing the order of participants in $(p_1, p_2)$ such that $p_1$ is the sender of the first message in $IM_t(p_1, p_2)$. That is, the first person in the pair is always the one who sent the first message to the other in the thread. By fixing the order, we also ensure that features with respect to $p_1$ will always fire. If the interaction was a one-way communication, features with respect to $p_2$ will not fire. Note that we consider each person of the pair separately for the analysis performed in Section 7.4 and so this step does not have any effect on the findings presented in that section.

### 7.5.2 Handling the Issue of Missing Features

In our formulation, values of many features are *undefined* for some instances. For example, the feature *Inform%* is undefined when *MsgCount* = 0, which happens for $p_2$ in a one-way communication. The issue of such missing values has been well-studied in statistics community and most statistical analysis software (such as R, the one we use) automatically handle these cases. So the results presented in Section 7.4 already accounts for this issue by treating these cases as missing values when the t-Test statistics are calculated.

However, it is not straightforward to handle the undefined values for features in the SVM algorithm (or other machine learning frameworks). This problem—some features being meaningless for some instances—has been actively researched within the machine learning community (Pelckmans et al. 2005, Chechik et al. 2008, García-Laencina et al. 2010). The most common approach used to tackle such cases is to substitute a zero for the value. However this approach conflates the cases where *Inform%* is *undefined* with those where *Inform%* is truly 0. For example, in our problem, we know that superiors have a smaller value for *Inform%* (Section 7.4, page 102), but subordinate send fewer messages and hence have higher chance of getting the value of *Inform%* to be assigned 0 under this approach, which will end up confusing the machine learning algorithm. We use another simple approach that has been shown to perform better in prior research (Chechik et al. 2008): add a new flag feature for every such feature that can be *undefined*. That is, we introduce an indicator feature for each structural feature to denote whether or not it is valid. Since we use a quadratic kernel, we expect the SVM to pick up the interaction between each feature and its indicator feature. Our results show that this approach improves over using the default option of substituting zero.

### 7.5.3 Masking of Names

The content of an email often contains references to the participants, especially in greetings and signature lines. Leaving them as is will lead to a model that learns from the names of the interactants. For example, if *Smith* has power over everyone else in the organization and he signs off his emails with his name, then a model trained using LEXICAL features would incorrectly puts a larger positive weight on the word feature *Smith*. Such a model will not work for a different organization where *Jones* has power over everyone else. In other words, such a system will over-fit to the corpus we are performing experiments on.

In order to avoid this, one option is to remove all greetings and signatures from the email body. There is some work within NLP to automatically perform this (e.g., (Carvalho and Cohen 2004)). However, signature text and greetings count as CONVENTIONAL dialog acts which we use as features in our study. Hence we want to preserve the signature and greeting lines. Another option is to remove all names from the email text. However, this will disrupt the sentence structure and will cause our dialog act and overt display of power taggers to perform worse. We use an alternate approach; we mask all the names in the email content. The procedure followed for this is as follows: first find all the first names and full names of all the participants in an email thread, then mask the occurrences of those names (using simple case-insensitive surface text matching) by replacing them with "Bob". This way, we preserve the syntactic structure as well as the dialog structure, but prevents the system from incorrectly biasing towards individual names.

In earlier results published in (Prabhakaran and Rambow 2014), we had not used this masking technique. In all the results presented in this section, we use the masking method, and as expected, our results are slightly lower than previously reported in (Prabhakaran and Rambow 2014). However, as we will see later in this section, the bias introduced by leaving the names as is was only 0.7 percentage points, and the overall conclusions made in (Prabhakaran and Rambow 2014) still hold.

### 7.5.4 Evaluation

We train our models using the related interacting participant pairs in threads in the *train* set and optimize its performance on the pairs from the *dev* set. We use accuracy, i.e., the percentage of pairs for which the direction of power relation was correctly predicted by the system, as the metric to measure the performance. We report accuracy obtained on pairs from both *dev* and *test* sets.

### 7.5.5 Experiments and Results

In this section, we present the various machine learning experiments we performed and their results. We use all six sets of features described in Section 7.3 — POSITIONAL, VERBOSITY, THREAD STRUCTURE, DIALOG ACTS, OVERT DISPLAY OF POWER, and LEXICAL— in our experiments. We start by describing the sets of experiments conducted.

**Feature Optimization Experiments** First, we conducted experiments to find the best setting to be used for the n-gram features in LEXICAL. We varied the value of $n$ for each type of n-gram — word lemma n-grams, part-of-speech n-grams and mixed n-grams — from 1 to 5 and found the best configuration to be $n = 2$ for all three of them, which is the configuration we use for the rest of the experiments. That is, we use unigrams and bigrams of word lemma n-grams, part-of-speech n-grams and mixed n-grams, for all results presented in this section. Once we found the best configuration for LEXICAL, we then performed experiments using all subsets of each of the feature categories (total $2^6 - 1 = 63$ experiments). We do not perform within-category feature optimization for other feature categories. Table 7.6 presents the results obtained using various feature subsets.

**Baseline results:** We use two different baseline systems — MAJOIRTYPREDICTION and WORDNGRAM. MAJOIRTYPREDICTION predicts the majority class always; in our case, the majority class is always *superior*. WORDNGRAM is a stronger baseline which uses the same machine learning framework as the rest of the experiments and uses word unigrams and bigrams as features. Note that this baseline do not use any NLP preprocessing (i.e., we use word ngrams, not word lemma ngrams) in making the predictions. The first two rows of Table 7.6 show results obtained using the baseline systems. The MAJOIRTYPREDICTION obtains a 52.5% accuracy, whereas the stronger baseline of WORDNGRAM obtains much higher accuracy of 68.6%. This result shows that lexical features, even without any NLP preprocessing, are useful for this task.

**Individual feature sets:** The next set of results in Table 7.6 lists accuracies obtained using each feature set individually. All feature sets except OVERT DISPLAY OF POWER obtained better results than the MAJOIRTYPREDICTION. LEXICAL obtained the highest improvement, posting an accuracy of 70.9%, a 2.4 percentage point improvement over the WORDNGRAM baseline. This improvement

| | System Description | Accuracy |
|---|---|---|
| Baselines | MAJOIRTYPREDICTION | 52.54 |
| | WORDNGRAM | 68.56 |
| Individual Feature Sets | PST | 53.66 |
| | VRB | 54.30 |
| | THR | 55.90 |
| | DA | 54.30 |
| | ODP | 49.97 |
| | LEX | 70.91 |
| All Feature sets | LEX, THR, PST, VRB, DA, ODP | 68.59 |
| | THR, PST, VRB, DA, ODP | 62.44 |
| | LEX, PST, VRB, DA, ODP | 67.44 |
| | LEX, THR, VRB, DA, ODP | 68.56 |
| | LEX, THR, PST, DA, ODP | 72.28 |
| | LEX, THR, PST, VRB, ODP | 68.53 |
| | LEX, THR, PST, VRB, DA | 68.53 |
| Best Feature Sets | LEX, THR, DA, ODP | 72.30 |
| | LEX, THR, PST | 72.30 |
| | LEX, THR, ODP | 72.22 |
| | LEX, THR, DA | 72.28 |
| | LEX, THR | 71.97 |
| | LEX, DA, ODP | 68.70 |
| | LEX, PST | 70.91 |
| | THR, DA, ODP | 58.50 |
| | THR, PST | 55.90 |
| Best without LEXICAL | THR, PST, VRB, DA | 62.47 |
| Best with no content | THR, VRB | 61.57 |

Table 7.5: Classifying *Superiors* vs. *Subordinates*: results on *dev* set. PST: POSITIONAL, VRB: VERBOSITY, THR: THREAD STRUCTURE, DA: DIALOG ACTS, ODP: OVERT DISPLAY OF POWER, LEX: LEXICAL

is obtained solely using lemmas of words instead of surface forms, as well as including part-of-speech ngrams and mixed ngrams into the feature set. Other than LEXICAL, the performance of all other systems were only marginally better than the MAJOIRTYPREDICTION. The best non-LEXICAL feature set is THREAD STRUCTURE, obtaining 55.9% accuracy that is a 3.4 percentage point improvement over the MAJOIRTYPREDICTION.

**Using all feature sets:** The next set of results in Table 7.6 presents results obtained using all feature sets, as well as the effect of removing each feature set on the performance. Using all six feature sets, the accuracy decreases to 68.6% a significant decrease from using LEXICAL alone (70.9%). This is counterintuitive since most of the structural features were significantly correlated with power in our statistical analysis (Section 7.4). Further analysis of our results reveal that the VERBOSITY features cause confusion to the machine learning system and is hurting the performance. Removing every other feature set decreases performance of the system, in varying degrees, but removing VERBOSITY improves the accuracy to 72.3%, a 2.4 percentage point improvement. As expected, removing LEXICAL features hurts the performance the most, decreasing the accuracy to 62.4%. Removing THREAD STRUCTURE features reduced the accuracy to 67.4%, a 1.2 percentage point decrease. Removing other feature sets affected the accuracy only marginally.

**Best feature subsets:** The next set of results in Table 7.6 lists the best performing feature subsets in the 63 experiments we conducted. There are two winners: the combination of LEXICAL, THREAD STRUCTURE, DIALOG ACTS and OVERT DISPLAY OF POWER, and the combination of LEXICAL, THREAD STRUCTURE and POSITIONAL. Both feature subsets obtained an accuracy of 72.3%, a statistically significant improvement over using LEXICAL features alone. As we saw in the previous two sets of results, LEXICAL and THREAD STRUCTURE are very useful for this task and are part of the final best result. We also list the results obtained on removing each feature set from the best feature sets. Removing DIALOG ACTS or OVERT DISPLAY OF POWER from the best feature set affects the performance only marginally. However, removing both of them (i.e., using only LEXICAL and THREAD STRUCTURE) reduces the accuracy to 72.0, suggesting that these features do add value to the classifier. Removing the THREAD STRUCTURE features decreased the performance for both of the winning feature sets (3.6 percentage points decrease in accuracy for the first one and 2.4 percentage points accuracy reduction for the second one). Removing LEXICAL

features reduced the accuracy significantly to 58.5% and 55.9% respectively.

**Without using LEXICAL features (for training):** It has been well understood in the NLP community that systems built using bag-of-words features such as our LEXICAL feature set poses the risk of over-fitting and building a model that is very domain-dependent. In addition, since the cardinality of LEXICAL features that are fed to the machine learning algorithm is orders of magnitude larger than the other feature sets, the training time for models that use LEXICAL features is significantly longer (four to five times, in our experiments). So we also report the best results obtained using no LEXICAL features, which was 62.5% obtained by the combination of THREAD STRUCTURE, POSITIONAL, VERBOSITY and DIALOG ACTS.

**Without using any email content:** In some situations, it will be desirable to build systems that do not look at the email content (for example, due to privacy reasons). Our DIALOG ACTS features do use lexical features to make the predictions, and hence do look at the email content. Our feature sets that do not look at the content of emails are THREAD STRUCTURE, POSITIONAL and VERBOSITY (note that VERBOSITY features do look at the size of the messages in terms of number of words, but do not look at the identity of the words). The best accuracy reported among these three feature sets was 61.6% using the combination of THREAD STRUCTURE and VERBOSITY features.

**Results on the blind test set** We now discuss evaluation on our blind test set. Table **??** presents the results obtained using our different best feature sets on the blind test set. The model trained using LEXICAL alone obtained an accuracy of 68.2%. The best performing feature sets we obtained from our experiments in the *dev* set both obtained significantly better results (72.8% and 72.9%) than the LEXICAL-only model. The model trained using no LEXICAL obtained an accuracy of 63.2%, while the model trained using features that do not look at the content at all obtained an accuracy of 62.7%. We obtained better results in the *test* set; and the patterns of improvements we saw in *dev* set carried over to the *test* set as well. This means that our model has not over-fitted to the *dev* set. Overall, we obtain a 14.8% accuracy error reduction in blind test by adding structural features to the LEXICAL-only model.

| | System Description | Accuracy |
|---|---|---|
| Best Feature Sets | LEX, THR, DA, ODP | 72.75 |
| | LEX, THR, PST | 72.93 |
| | LEX | 68.23 |
| Best without LEXICAL | THR, PST, VRB, DA | 63.21 |
| Best with no content | THR, VRB | 62.65 |

Table 7.6: Classifying *Superiors* vs. *Subordinates*: results on *test* set.
PST: POSITIONAL, VRB: VERBOSITY, THR: THREAD STRUCTURE,
DA: DIALOG ACTS, ODP: OVERT DISPLAY OF POWER, LEX: LEXICAL

**Significance of Improvements** We use the Approximate Randomization technique (Yeh 2000) to measure the significance of our improvements. We found the improvements we obtained in our results by adding structural features to the model that uses only lexical features to be statistically significant ($p < 0.05$) in both *dev* and *test* sets.

## 7.6 gSPIN: A Browser Extension for Email Power Analytics

In this section, we describe a system that demonstrate a practical application of the work presented in this chapter. We use the term SPIN (Social Power in INteractions) system to refer to the end-to-end power prediction system described in this chapter. We present a browser extension called gSPIN that allows its users to perform power analytics on their personal (or official) emails using the SPIN system. The extension is currently made available for the Google Chrome web browser and it works seamlessly with Gmail email threads once the user installs the extension and grants it the necessary access permissions.

### 7.6.1 Functionality

The gSPIN extension can be installed directly into the Google Chrome browser. Once installed, a gSPIN button will be displayed in the right end of the Chrome address bar, when the user navigates to a Gmail email thread. If the user clicks on the gSPIN button, the extension gives the user different options to perform the SPIN analysis on the email thread that is currently being displayed in the

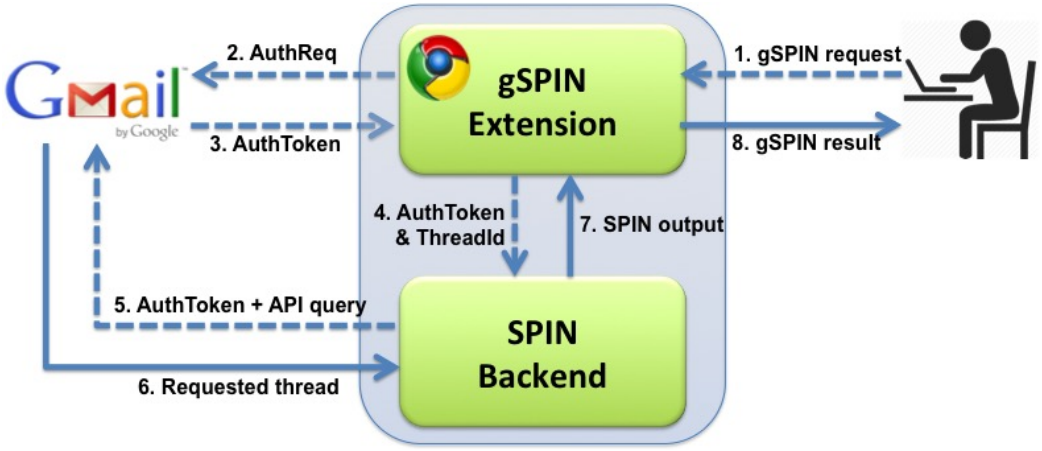



Figure 7.2: gSPIN plugin: process flow and system architecture.
Dotted arrows indicate communications that do not involve email content.
Solid arrows indicate email content being transferred

browser. The user can request to perform only the power prediction task, or both dialog analysis and power prediction. Upon clicking the submit button, the gSPIN extension will obtain authentication using the Google Chrome Identity API,[2] and securely obtain the email thread for processing directly from the Gmail server using the Gmail API.[3] The Gmail API gives a secure, RESTful access to the user's email threads. It will then process the email thread using the SPIN system and upon completion, display the results in a pop-up browser window.

### 7.6.2 System Architecture and Process Flow

In this section, we describe the system architecture and the process flow followed in the gSPIN system as shown in Figure 7.2. Dotted arrows indicate communications that do not involve email content; solid arrows indicate email content being transferred We describe below each step of the process, starting from the user initiating the gSPIN request until the gSPIN processing results are displayed to the user.

1. User requests gSPIN processing of the Gmail thread that is displayed in the Chrome browser.

[2]https://developer.chrome.com/apps/app_identity

[3]https://developers.google.com/gmail/api/

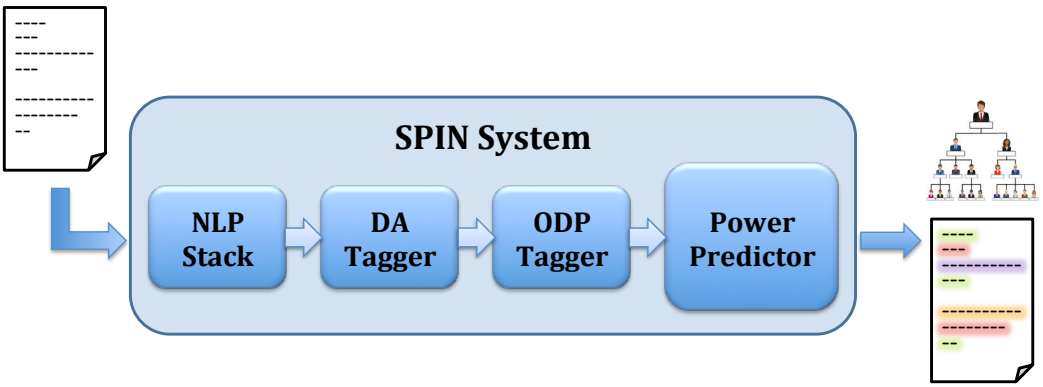



Figure 7.3: SPIN system: processing pipeline.

2. gSPIN sends an authentication request to the Google Chrome Identity API to obtain an authentication token
3. Google Chrome Identity API returns the authentication token after verifying user credentials
4. gSPIN sends the authentication token along with the thread identifier obtained from the email thread url to the backend SPIN server.
5. The backend SPIN server communicates directly with the Gmail API using the authentication token to request the content of the email thread.
6. The Gmail API returns the content of the requested email thread in the JSON format.
7. The SPIN system processes the email thread as per the processing pipeline shown in Figure 7.3 and returns the output in the SPINOut XML format to gSPIN
8. gSPIN unpacks the SPINOut results and displays it to the user in a pop-up window.

Steps 2 and 3 (authentication steps) are performed only for the initial request and when an already obtained authentication token has expired.

### 7.6.3 SPIN Processing Pipeline

As described in detail in Section 7.5, our system uses deep NLP analysis including dialog act tagging of email threads and detecting overt displays of power to make the predictions. Figure 7.3

shows the processing pipeline of SPIN system. The first step of SPIN processing applies the basic NLP steps such as tokenization, sentence segmentation, lemmatization, and part of speech tagging. This analysis is then used by the Dialog Act Tagger (Chapter 5) to assign dialog act tags to sentences, which are then used along with other lexical features by the Overt Display of Power Tagger (Chapter 6) to detect instances of overt displays of power. Features from all these stages contribute to the final stage that predicts the power relation between pairs of participants. In addition to the power relations between participants, the SPIN system also makes the output of the lower-level dialogic analysis available to the user, as they themselves give useful insights about the conversation. That is, the SPIN system takes as input an interaction, and outputs analysis results of three types:

- classification of dialog acts for sentences.
- instances of overt displays of power.
- superior/subordinate power relations between pairs of participants.

### 7.6.4 gSPIN at Work

The screen shot shown in Figure 7.4- 7.6 shows the output produced by the SPIN analysis on a sample email conversation. The first section of the output shows the power relations detected between pairs of interacting participants. The second section displays the original email thread with annotations of dialog acts and overt displays of power.

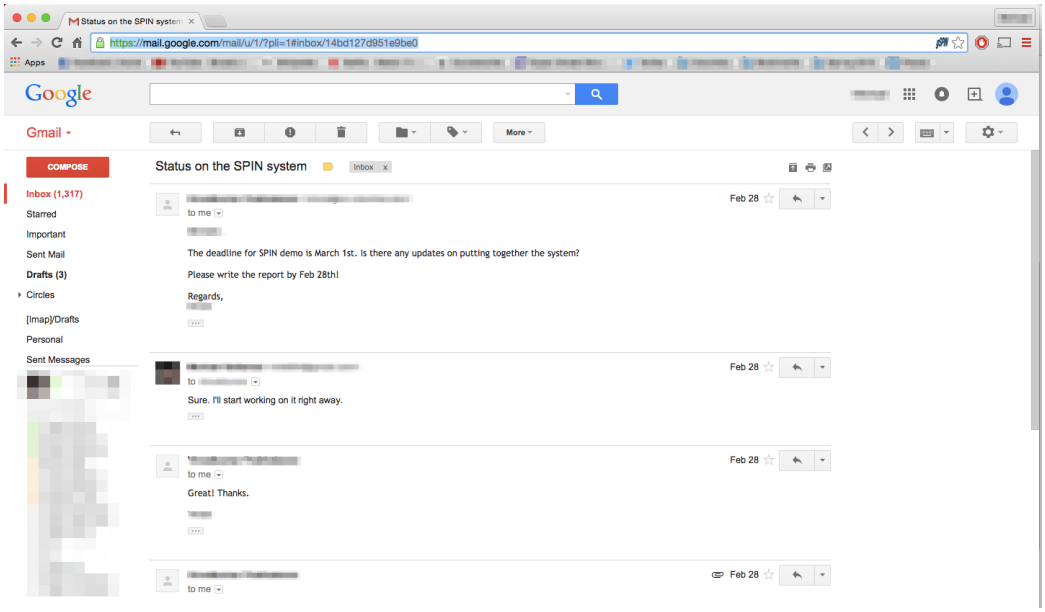



Figure 7.4: gSPIN at Work: screen shot 1. Once a Gmail thread has been loaded, the gSPIN icon will appear in the address bar.

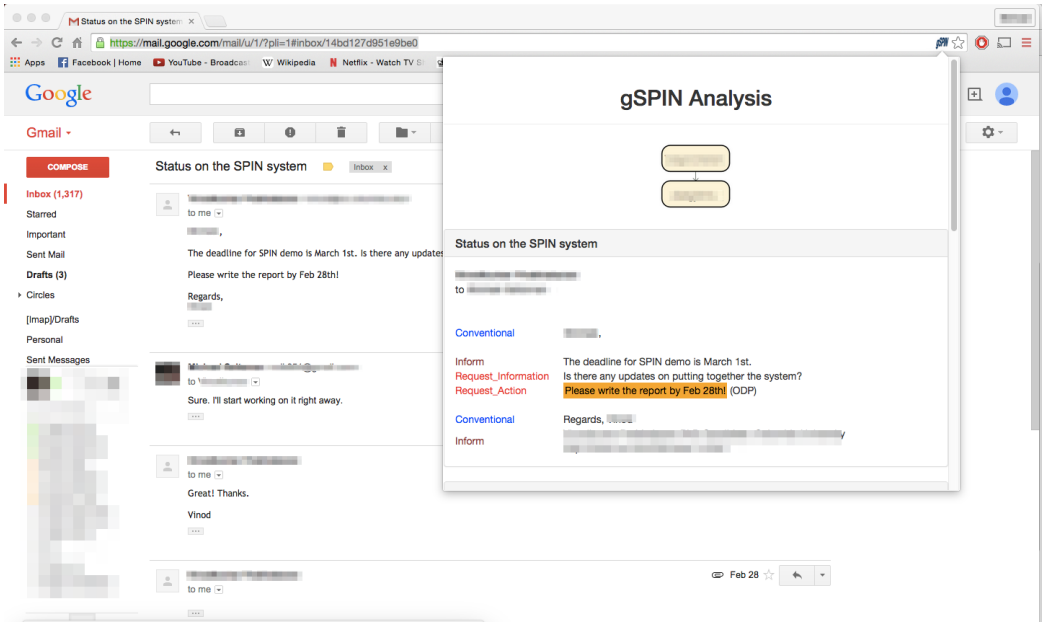



Figure 7.5: gSPIN at Work: screen shot 2. After processing, the gSPIN plugin displays the SPIN power prediction results. The first section shows the power prediction results.

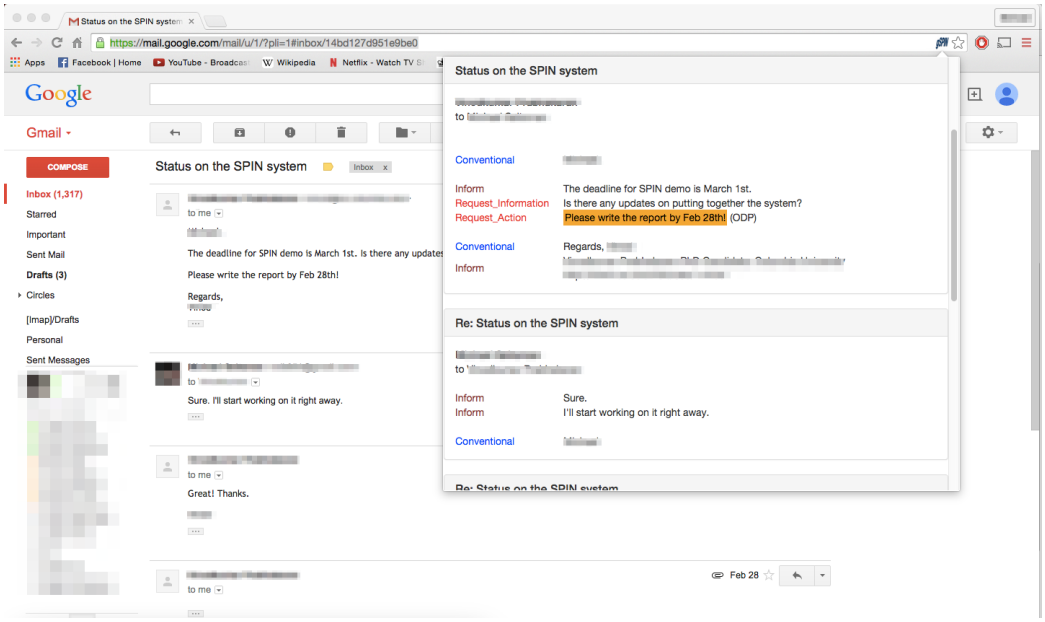



Figure 7.6: gSPIN at Work: screen shot 3. After processing, the gSPIN plugin displays the SPIN power prediction results. The second section shows the dialog analysis results, highlighting dialog act labels and instances of overt displays of power.

## 7.7 Conclusion

In this chapter, we introduced the problem of predicting direction of power between pairs of people from single threads of interactions. We performed this study in the organizational emails genre. We described the problem formulation in detail and laid out the analysis framework that becomes the basis of analysis for Chapters 8, 9 and 10. The contributions of this chapter are three fold. First, we presented a detailed statistical analysis of how power is manifested along different dialog structural aspects of interactions. Second, we presented an automatic direction-of-power predictor using dialog structure features, along with experiments and results. Third, we described a Google Chrome browser plugin that enables its user to apply the power prediction system described in this chapter, along with the dialog act tagger (Chapter 5) and overt display of power tagger (Chapter 6) to his/her email threads.

In our analysis, we found that power is manifested in the language as well as dialog structure of interactions. We showed that superiors and subordinates have significantly different values for their thread structure features as well as dialog act based features. Superiors send significantly more

messages than subordinates, but their messages are significantly shorter. Superiors also have significantly more recipients in their emails. In terms of dialog act features, we found that superiors issues almost twice as many requests for action as subordinates, whereas subordinates' contribute significantly more information in the conversation. Superiors also use significantly more overt displays of power than subordinates.

In our machine learning experiments, we found that lexical features have great predictive power for distinguishing between superiors and subordinates. A system that is trained using lexical features alone obtained an accuracy of 70.9%, compared to 61.6% obtained by a model trained purely on structural features without looking at the content of emails at all. However, adding structural features significantly improve the performance of the model that use lexical features alone. In fact, our best performing models include thread structure features, positional features, dialog act features, and the feature denoting overt displays of power. We obtain a best accuracy of 72.3% compared to the 70.9% obtained by a system that uses only lexical features, a 5.1% error reduction.

# Chapter 8

# Gender, Gender Environment, and Manifestations of Power

It has long been observed that men and women communicate differently in different contexts. There has been an array of studies in sociolinguistics that analyze the interplay between gender and power. These sociolinguistics studies often rely on case studies or surveys. The availability of large corpora of naturally occurring interactions, and the advanced computational techniques to process the language and dialog structure of these interactions, has given us the opportunity to study the interplay between gender, power, and language use at a larger scale than before. In this chapter, we study how gender correlates with manifestations of power in an organizational setting using the Enron email corpus. 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 power relations he or she has to the discourse participants. We concentrate on modeling the writer's choices related to the aspects of dialog behavior that we studied in Chapter 7. 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.

We make three major contributions in this chapter. First, we introduce an extension to the Enron corpus of emails: we semi-automatically identify the sender's gender of 87% of email messages in the corpus. This extension has been made publicly available. Second, we use this enriched version of the corpus to investigate the interaction of hierarchical power and gender. We formalize the no-

tion 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. This contribution shows how social science can benefit from advanced natural language processing techniques in analyzing corpora, allowing social scientists to tackle corpora that cannot be examined in their entirety manually. Third, we show that the gender information in the enriched corpus can be useful for computational tasks, specifically for improving the performance of the power prediction system presented in Chapter 7 that is trained to predict 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.

We start by discussing related work in sociolinguistics on the interplay between gender and power followed by work within the NLP community on gender and use of language (Section 8.1). In Section 8.2, we present the first contribution of this chapter — the Gender Identified Enron Corpus, and describe the procedure followed to build this resource and present various corpus statistics. In Section 8.4, we present the results from a statistical analysis of the interplay between gender, power and dialog behavior. Section 8.5 introduces the notion of gender environment and Section 8.6 presents the statistical analysis of how gender environment affects the manifestations of power. In Section 8.7, we demonstrate the utility of gender-based features in the problem of automatically predicting the direction of power between participants of an interaction, before we summarize the conclusions from this chapter in Section 8.8

## 8.1 Literature Review

There is much work in sociolinguistics on how gender and language use are interrelated (Tannen 1991; 1993, Holmes 1995, Kendall and Tannen 1997, Coates 1998, Eckert and McConnell-Ginet 2003, Holmes and Stubbe 2003, Mills 2003, Kendall 2003, Herring 2008). Some of this work look specifically at language use in the work environment and/or with respect to power relations, whereas some others study the gender differences in language use in general. In this section, we summarize the different strands of this research, focusing more on the studies that have influenced the work presented in this thesis.

### 8.1.1 Gendered Differences in Language Use

Many sociolinguistics studies have found evidence that men and women differ considerably in the way they communicate. Some researchers attribute this to psychological differences (Gilligan 1982, Boe 1987), whereas some others suggest socialization and gendered power structures within the society as its reasons (Zimmerman and West 1975, West and Zimmerman 1987, Tannen 1991). Tannen (1991) argues that "for most women, the language of conversation is primarily a language of rapport: a way of establishing connections and negotiating relationships", which she calls *rapport-talk*, whereas "for most men, talk is primarily a means to preserve independence and negotiate and maintain status in a hierarchical social order", which she calls *report-talk*. Along the same lines, Holmes (1995) argues that "women are much more likely than men to express positive politeness or friendliness in the way they use language". In addition to politeness, many other linguistic variables have been analyzed in this context. Lakoff (1973) describes women's speaking style as tentative and unassertive, and argues that women use question tags and hedges more frequently than men do. However, in a later study, Holmes (1992) found that the differential use of question tags in-fact depends on the function of the question tag in the interaction. She categorized the instances of question tags in terms of their functionality in the contexts in which they were used, and found that question tags used as a way to express uncertainty was done more by men, whereas question tags used as a way to facilitate communication was done more by women. Researchers have also looked into interruption patterns in interactions in relation to gender. For example, Zimmerman and West (1975) found that men interrupted conversations more often in cross-sex interactions, whereas there were no significant differences in interruptions in same-sex interactions.

However, recent studies have suggested the need for a more nuanced view on the interplay between gender and language use. They argue that the differences observed by above studies are due to more complex processes at play than gender alone, and that one needs to take into account the context in which the interactions happened to understand the gender differences better. Mills (2003) challenges the above line of analysis, especially Holmes (1995)'s theory regarding women being more polite. She argues that politeness cannot be codified in terms of linguistic form alone and calls for "a more contextualized form of analysis, reflecting the complexity of both gender and politeness, and also the complex relation between them". Along those lines, Coates (2013) also challenge Lakoff (1973)'s theory on women's language being unassertive. She points out that hedges are

multi-functional constructs and the greater usage of hedges by women "can be explained in part by topic choice, in part by women's tendency to self-disclose and in part by women's preference for open discussion and a collaborative floor". In other words, she argues that women using more hedges than men does not entail that women are unassertive, but instead is an artifact of what topics women often take part in. Kunsmann (2013) connects the gender differences in language specifically to status, dominance and power. He argues that "gender and status rather than gender or status will be the determinant categories" of language use. In our work, we follow a similar approach. We do not study gender in isolation, but in the context of the social power relations as well as the gender environment of the interaction.

### 8.1.2 Gender and Power in Work Place

Within the area of studying gender and language use, there is substantial amount of work that is specifically related to the language use in the work environment (West 1990, Tannen 1994, Kendall and Tannen 1997, Kendall 2003). These studies found that women use more polite language and are "less likely to use linguistic strategies that would make their authority more visible" (Kendall 2003). In this thesis, we study this aspect using our formulation of overt displays of power, which are face-threatening acts that reinforce the status differences. Our results align with these studies, however, we draw from a much larger-scale study than them, in which we analyze thousands of email interactions rather than a handful of case studies.

We summarize the findings of some of the above mentioned studies below. West (1990) found that male physicians and female physicians differed in how they gave directives to their patients. Male physicians aggravated their directives, whereas female physicians used forms that mitigated them. Similarly, in the study of gender, power and language in large corporate work environments, Tannen (1994) found that female managers use more face saving strategies (e.g., phrasing directives as suggestions: *You might put in parentheses*) when talking to subordinates, whereas male managers used language that reinforced status differences (e.g., *Oh, that's too dry. You have to make it snappier!*). Kendall (2003) shows that this behavior is specific to women operating in work environments. She studied the demeanor of a woman exercising her authority at work and at home, and found that while the woman used mitigating strategies to exercise her authority at work (as found by other studies before), she created a demeanor of explicit authority when exercising her authority

over her daughter at home. Our findings in this thesis on the Enron emails are also in line with above findings; we observe that male managers use significantly more overt displays of power when interacting with subordinates, whereas female managers use significantly fewer of them.

Another line of work that has influenced the work presented in this thesis, is by Holmes and Stubbe (2003) studying the effects of gendered work environments in the manifestations of power. They provide two case studies that analyze, not the differences between male and female managers' communication, but the differences 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 formulating the notion of "Gender Environment" in our study. We adapt this notion to the level of an interaction, and define the gender environment of an email thread in terms of the ratios of males to females on a thread, allowing us to look at whether the manifestations of power change within a more heavily male or female thread.

### 8.1.3 Computational Approaches towards Gender

Within the NLP community, there is a considerable amount of work on analyzing language use in relation to gender. Early work attempted to use NLP techniques to automatically predict the gender of authors using lexical features. Researchers have attempted gender prediction on a variety of genres of interactions such as emails, blogs, and online social networking websites such as Twitter (Corney et al. 2002, Peersman et al. 2011, Cheng et al. 2011, Deitrick et al. 2012, Alowibdi et al. 2013, Nguyen et al. 2014). In more recent work, Hovy (2015) argues for research in the other direction, showing the importance of using gender information for better performance on NLP tasks such as topic identification, sentiment analysis and author attribute identification.

While automatically detecting gender is an interesting problem, our focus in this thesis is not gender detection, but understanding the variations in linguistic patterns with respect to both gender and power. For this, we require a more reliable source of gender assignments. Hence, we use publicly available name databases to reliably determine the gender of participants as we have access to the email authors' names in our corpus. We believe that the gender-identified email corpus we are making available as part of this thesis will aid further research in the area of gender detection.

Existing work on gender prediction rely on relatively smaller datasets. For example, Corney et al. (2002) use around 4K emails from 325 gender identified authors in their study. Cheng et al. (2011) use around 9K emails from 108 gender identified authors. Deitrick et al. (2012) use around 18K emails from 144 gender identified authors. In contrast, we build a gender-assigned email dataset that is orders of magnitude larger than these resources. Our corpus contains around 97K emails whose authors are gender-identified, and these emails are from around 23K unique authors.

There has also been work on using NLP techniques to analyze gender differences in language use by men versus women (Mohammad and Yang 2011, Bamman et al. 2012; 2014, Agarwal et al. 2015). Mohammad and Yang (2011) analyze the way gender affects the expression of emotions in the Enron corpus. They found that women send and receive emails with relatively more words that denote joy and sadness, whereas men send and receive relatively more words that denote trust and fear. For their study, they assigned gender for the core employees in the 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. 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.

Bamman et al. (2012; 2014) study gender differences in the microblog site Twitter. One of the many insights from their work is that gendered linguistic behavior is determined 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. More specifically, people with many same-gender friends tend to use language that is strongly associated with their gender, whereas people with more balanced social networks tend not to. Our notion of gender environment captures the gender makeup of an interaction, and our findings reaffirms the need to also look into the audience's gender makeup in studying gender.

To our knowledge, ours is the first computational study of this scale that focus on the interplay between gender and power. We study the effects of gender in workplace interactions, not by considering the email senders' gender in isolation, but together with their power relations with the rest of the participants, as well as the gender makeup of the interaction.

## 8.2 Gender Identified Enron Corpus

In this chapter, our starting point is the same corpus (ENRON-LARGE) and terminology introduced in Chapter 7 (Section 7.2, page 96). The corpus contains emails from the mailboxes of 145 core Enron employees, and captures the thread structure of email messages. The presence of the email thread structures in the corpus allows us to go beyond isolated messages and study gender in relation to the dialog structure as well as the language use. However, there are 34,156 unique discourse participants (senders and recipients together) across all the email threads in the corpus, and manually determining the gender of all of them is not feasible. Hence, we adopt a two-step approach through which we reliably identify the gender of a large majority of discourse participants in the corpus.

Step 1: Manually determine the gender of the 145 core employees who have a bigger representation in the corpus

Step 2: 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.

### 8.2.1 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 sentence "Kay Mann is a strong addition to Noble's senior leadership team, and we're delighted to welcome *her* aboard" (gender-revealing pronoun emphasized) in the page we found for Kay Mann clearly identifies her gender.[1] We were able to correctly determine the gender of each of the 145 core

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

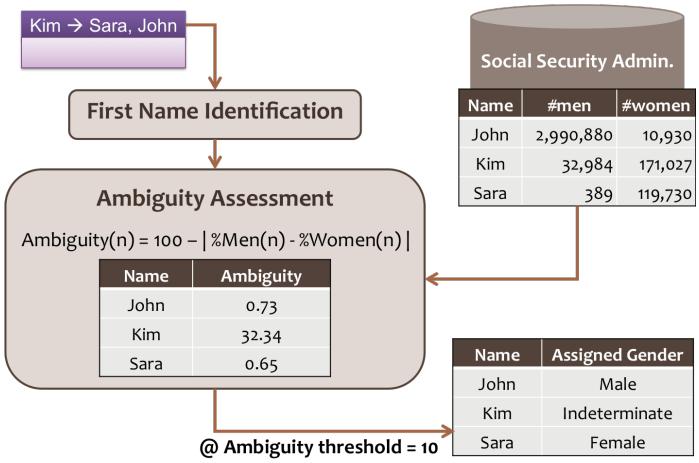



Figure 8.1: Automatic gender assignment process.

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.

### 8.2.2 Automatic Gender Assignment

Our corpus contains a large number of discourse participants in addition to the 145 core employees for which we manually identified the gender. The steps we follow to assign gender for these other discourse participants is pictorially represented in Figure 8.1. 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.

#### 8.2.2.1 SSA Names and Gender Dataset

The US Social Security Administration maintains a dataset of baby names, gender, and name count for each year starting from the 1880s, for names with at least five counts.[2] 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,[3] so we

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

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

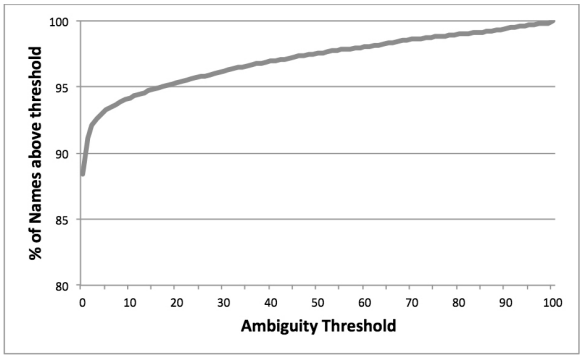



Figure 8.2: Plot of percentage of first names covered against ambiguity threshold.

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 $n$. The difference between these percentages of a name gives us a measure of how ambiguous it is; the smaller the difference, the more ambiguous the name. We define the ambiguity score of a name $n$, denoted by $AS(n)$, as follows:

$$AS(n) = 100 - |mp(n) - fp(n)|$$

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 $AS_T$.

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

Figure 8.2 shows the plot of the percentage of names that will be gender assigned in the SSA dataset against the ambiguity threshold. As the plot shows, around 88% of the names in the SSA dataset have $AS(n) = 0$, i.e., are unambiguous. 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. An ambiguity threshold of 10 means that we assign a gender only if at least 95% of people with that name were of that gender. In the gender assigned corpus that we released, we retain the $AS(n)$ of each name, so that the users of this resource can decide the threshold that suits their needs.

#### 8.2.2.2 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. The names in the email headers are populated from individual email clients the senders were using and hence do not follow a standard format. To make things worse, 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 set of candidate names 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; pick the first word if the name field does not contain a comma; pick the word following the comma if it does contain one.
- 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.

In addition, we cleaned up some irregularities that were present in the name field. One common issue was that many email fields started with the text '?S' possibly a manifestation of some data preprocessing step. We strip this portion of the string in order to obtain the part that denote the actual email address.

The above heuristics create a list of candidate names for each discourse participant. For each candidate name, we compute the ambiguity score (Section 8.2.2.1) 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.

#### 8.2.2.3 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. Table 8.1 presents the results of this evaluation. 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. For the cases where we correctly identify the first name, we obtain a 100% accuracy in assigning the gender. The errors in finding first name 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.

| | Percentage |
|---|---|
| Coverage | 90.3 |
| Accuracy | 89.3 |

Table 8.1: Performance of automatic gender assignment.

### 8.2.3 Corpus Statistics and Divisions

**Gender assignment coverage:** We apply the gender assignment system described above to all discourse participants of all email threads in the ENRON-LARGE corpus. Table 8.2 shows the coverage of gender assignment in our corpus at different levels: unique discourse participants, messages

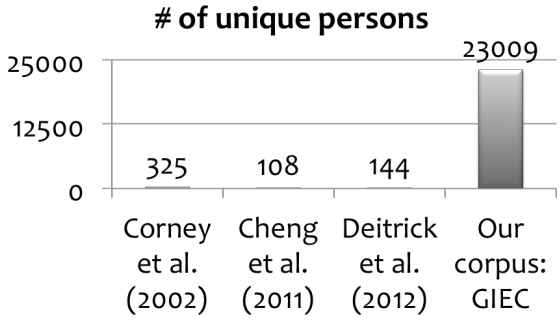
| | Count (%) |
|---|---|
| Total unique discourse participants | 34,156 |
| - gender identified | 23,009 (67.3%) |
| Total messages | 111,933 |
| - senders gender identified | 97,255 (86.9%) |
| Total threads | 36,615 |
| - All Senders Gender Identified (ASGI) | 26,015 (71.1%) |
| - All Participants Gender Identified (APGI) | 18,030 (49.2%) |

Table 8.2: Coverage of gender identification at various levels: unique discourse participants, messages and threads.

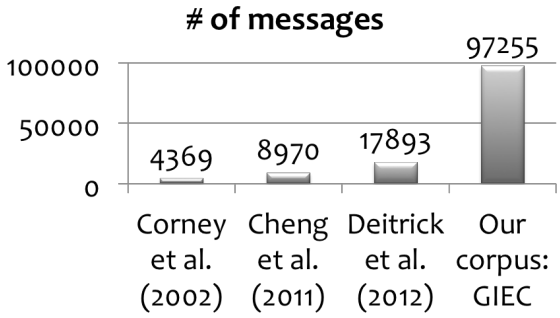
and threads. We were able to identify the gender of 67% of unique discourse participants in the corpus. This amounted to the senders of 87% of the messages in our corpus. We call the subset of threads for which we were able to identify the gender of all email senders, the *All Senders Gender Identified (ASGI)* sub-corpus, and those for which we were able to identify the gender of all participants including senders and all recipients, the *All Participants Gender Identified (APGI)* sub-corpus. ASGI covers around 71% of threads in the corpus, whereas APGI covers only about 49%. The users of this resource can limit their study to either subset, depending on their requirements.

In Figure 8.3, we show how the size of our gender identified Enron corpus compares to existing gender assigned resources (Corney et al. 2002, Cheng et al. 2011, Deitrick et al. 2012). Our corpus is orders of magnitude larger than existing resources. We have representation of over 23K authors in our corpus, as opposed to a few hundred in other existing resources. In terms of number of messages also, our corpus is more than 5 times the size of next biggest corpus.

**Gender assignment male/female split:** In Figure 8.4, 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). We have more male participants than female participants in the corpus (58% vs. 42%). When counted in terms of number of messages, around two thirds of the messages in our corpus were sent by men.




(a) Comparison in terms of number of unique discourse participants




(b) Comparison in terms of number of messages

Figure 8.3: Gender Identified Enron Corpus vs. existing gender assigned resources.

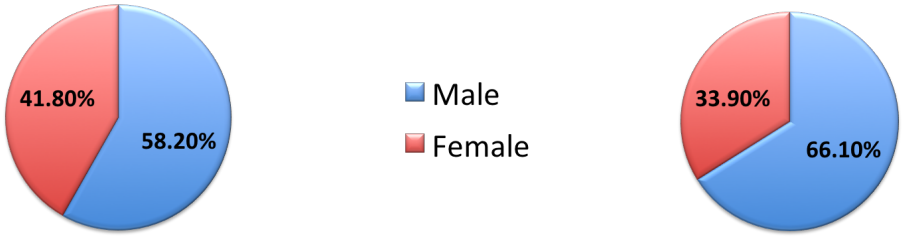



Figure 8.4: Male/Female split in gender assignments across a) all unique participants who were gender identified (left), b) all messages whose senders were gender identified (right)

## 8.3 Data Setup and Features

In this chapter, we use the gender assignments descried in Section 8.2 to study the interplay of gender and power. We use the same analysis framework — problem formulation, data splits, and features — introduced in Chapter 7. In this section, we briefly summarize the analysis framework and features we used, in order to give the necessary background required for the rest of this chapter. For a detailed account of the problem and features, refer Chapter 7, Section 7.2, page 96 & Section 7.3, page 99.

| 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 8.3: Data statistics in Gender Identified Enron Corpus.
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.

### 8.3.1 Problem

We study the pairs of participants $(p_1, p_2) \in RIPP_t$, the set of related interacting participant pairs in an email thread $t$. We are interested in the differences in the dialog behavior exhibited by superiors and subordinates. In this chapter, we study how their gender and the gender of other participants in the email thread affects these dialog behavior differences.

### 8.3.2 Data

We follow the same *train*, *dev*, *test* division of ENRON-LARGE described in Section 7.2. We limit our study in this chapter to the threads in which were able to identify the gender of all participants (i.e., threads that are part of the APGI subset of the corpus). Table 8.3 presents the total number of pairs in $IPP_t$ and $RIPP_t$ from all the threads in the APGI subset of our corpus and across the *train*, *dev* and *test* sets.

We choose APGI instead of ASGI (All Senders Gender Identified) because APGI allows us to also study the notion of Gender Environment (to be introduced in Section8.5) for which we need to know the gender of all participants. As an artifact of choosing the APGI, we also have a corpus with relatively smaller number of participants per thread than the full corpus. In other words, email threads with large number of participants, such as broadcast emails will have been excluded from the AGPI, since there is a higher chance that the automatic gender assignment step fails to assign the gender for at least one of the recipient. As a result, the findings from the analysis we perform in this chapter might sometimes differ from what we found in Chapter 7. However, knowing how the

two corpora differ in terms of the number of participants, it is interesting to note on which aspects of interactions the findings in both studies differ.

### 8.3.3 Features

We use the same aspects introduced in Chapter 7 (Section 7.3, page 99). We list the features again here in Table 8.4. For more details on each feature, refer to Section 7.3. We also describe each feature in more detail in Section 8.4, where we discuss the findings about them from our statistical analysis on how they differ with respect to gender and power.

## 8.4 Gender and Power: A Statistical Analysis

As a first step, we would like to understand whether male superiors, female superiors, male subordinates, and female subordinates differ in their dialog behavior. For this analysis, the ANOVA (Analysis of Variance) test is the appropriate statistical test as it provides a way to test whether or not the means of several groups are equal. In other words, ANOVA generalizes the Student's t-Test to situations with more than two groups. It also eliminates the possibility of making a type I error (false positives) if multiple two-sample t-Tests were applied to such a problem.

We perform ANOVA tests on all features keeping both Hierarchical Power and Gender as independent variables; i.e., there are four groups — male superiors, female superiors, male subordinates, and female subordinates. 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 use the Tukey's HSD (Honest Significant Difference) Test.

Altogether, there are twenty features each at the thread level and interaction level which are the dependent variables, and two independent variables — Power and Gender. That is a total of one hundred and twenty different statistical tests; in addition, for each ANOVA test, we also perform the Tukey's HSD test. This leads to a large number of results that we cannot discuss entirely in this section. We list the results obtained in all the statistical tests in Appendix B and discuss the main findings in each set of features below.

| Aspects | Features | Description |
|---|---|---|
| PST | *Initiator* | did $p$ sent the first message? |
| | *FirstMsgPos* | relative position of $p$'s first message in $M$ |
| | *LastMsgPos* | relative position of $p$'s last message in $M$ |
| VRB | *MsgCount* | Count of messages sent by $p$ in $M$ |
| | *MsgRatio* | Ratio of messages sent in $M$ |
| | *TokenCount* | Count of tokens in messages sent by $p$ in $M$ |
| | *TokenRatio* | Ratio of tokens across all messages in $M$ |
| | *TokenPerMsg* | Number of tokens per message in messages sent by $p$ in $M$ |
| THR | *AvgRecipients* | Avge. number of recipients in messages |
| | *AvgToRecipients* | Avge. number of To recipients in messages |
| | *InToList%* | % of emails p received in which he/she was in the To list |
| | *AddPerson* | did $p$ add people to the thread? |
| | *RemovePerson* | did $p$ remove people to the thread? |
| | *ReplyRate* | average number of replies received per message by $p$ |
| DA | *ReqActionCount* | # of Request Action dialog acts in $p$'s messages |
| | *ReqInformCount* | # of Request Information dialog acts in $p$'s messages |
| | *InformCount* | # of Inform dialog acts in $p$'s messages |
| | *ConventionalCount* | # of Conventional dialog acts in $p$'s messages |
| | *DanglingReq%* | % of messages with requests sent by $p$ that did not have a reply |
| ODP | *ODPCount* | Number of instances of overt displays of power |

Table 8.4: Aspects of interactions analyzed in organizational emails.

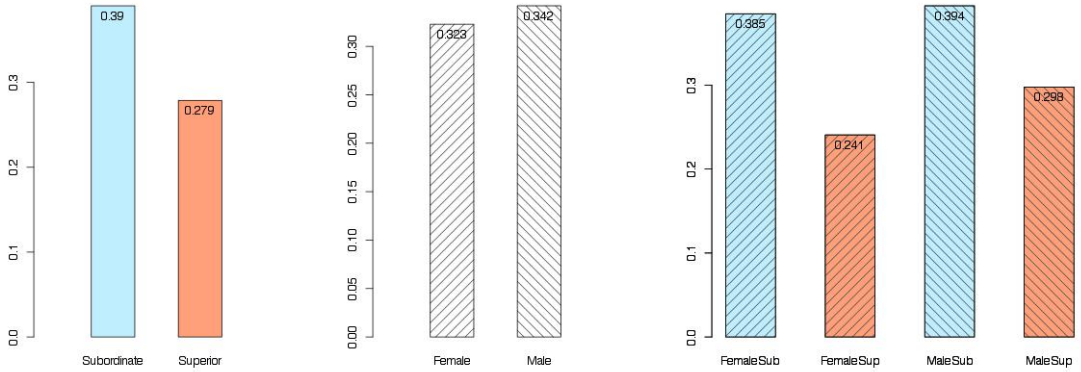



Figure 8.5: Mean value differences along Gender and Power: Initiator

### 8.4.1 Positional Features

There are three features in this category — *Initiator*, *FirstMsgPos*, and *LastMsgPos*. *Initiator* is a binary feature which gets the value of 1 (*true*) if the participant sent the first message in the thread, and 0 otherwise (*false*). *FirstMsgPos* and *LastMsgPos* are real-valued features taking values from 0 to 1. The lower the value, the earlier the participant sent the first (or last) message. The first two features relate to the participant's initiative. A higher average value for *Initiator* in a group indicates that participants in that group initiates threads more often; so does a lower average value for *FirstMsgPos*. *LastMsgPos* captures whether participant stayed on towards the end of the thread.

Figure 8.5 shows the mean values of each groups for the feature *Initiator*. *Initiator* and *FirstMsgPos* behave more or less similarly; hence we show the chart only for *Initiator*. Subordinates tend to initiate the threads significantly more often than superiors (average value of 0.39 against 0.28 for *Initiator*). This pattern was also seen in *FirstMsgPos* (0.18 over 0.23; lower value means earlier participation). Both differences are highly statistically significant $p < 0.001$. This finding appears to be in contrast with our finding in Chapter 7 (Section 7.4.2, page 104) that superiors initiate more conversations. As we discussed earlier, this is an artifact of the fact that broadcast messages with large number of participants get eliminated from our corpus because it is more likely to fail to assign gender to at least one of the participants. Putting together both findings, what we infer is that superiors tend to initiate email threads with large number of people; but in smaller conversations, it is the subordinates who initiate the conversations.

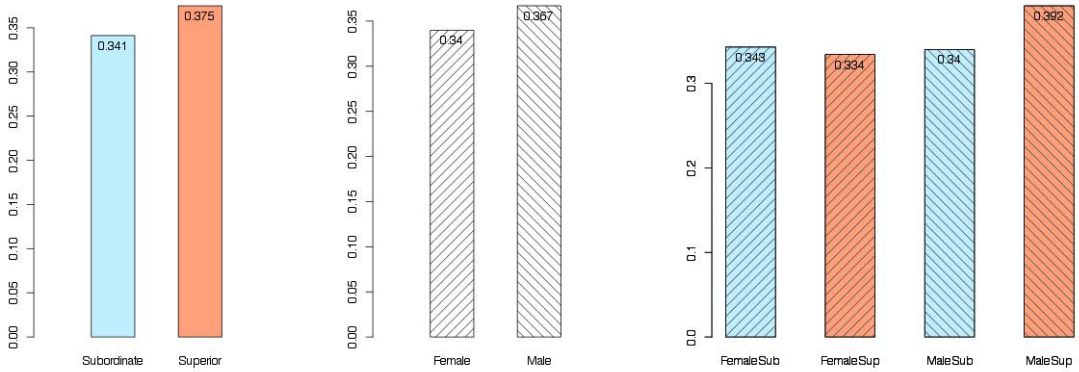



Figure 8.6: Mean value differences along Gender and Power: LastMsgPos

Gender was not a deciding factor. For *Initiator*, the t-Test result was significant ($p = 0.03$), however the magnitude of difference was very small (0.32 for females over 0.34 for males; Figure 8.5). The t-Test result was not significant for *FirstMsgPos*. For the ANOVA test for the combination of gender and power, the result was not significant for *Initiator*. The ANOVA test for *FirstMsgPos* was significant, however the Tukey's HSD test shows that the groups that were different were all cases where the superior vs. subordinate. In other words, male and female superiors behaved more or less the same way; similarly, male and female subordinates also behaved the same way.

The results on *LastMsgPos* was interesting (Figure 8.6). The t-Test results for both power and gender were significant, although the magnitude of the difference was relatively small. The last message from superiors tend to come later than those of subordinates. Similarly, males tend to send their last messages later than females. The ANOVA results show that the factorial groups of power and gender also differ significantly ($p < 0.01$). Upon Tukey's HSD test we find that male managers are the only group that differs from everyone else. The differences between all other groups were not statistically significant. But male managers differed from every other group significantly ($p < 0.01$). We have not been able to determine an explanation for this correlation. For Power, this result is in contrast with our findings in Chapter 7 (Section 7.4.2, page 104), again an artifact of removing broadcast email threads.

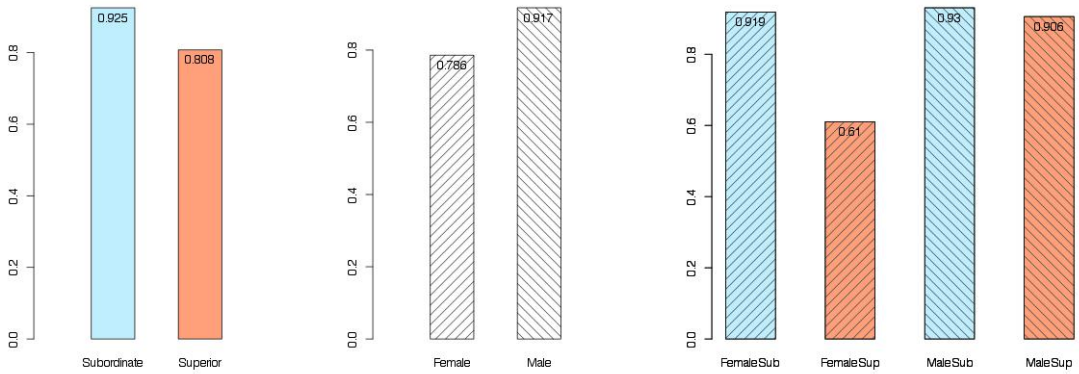



Figure 8.7: Mean value differences along Gender and Power: MsgCount

### 8.4.2 Verbosity Features

There are five features in this category — *MsgCount*, *MsgRatio*, *TokenCount*, *TokenRatio*, and *TokenPerMsg*. The first two features measure verbosity in terms of messages, whereas the third and fourth features measure verbosity in terms of words. The last feature measure how terse or verbose on average were the messages.

*MsgCount* and *MsgRatio* behaved similarly, so did *TokenCount* and *TokenRatio*. Figure 8.7 and Figure 8.8 show the mean values of each groups for the feature *MsgCount* and *TokenCount*. Superiors tend to send fewer of messages in the thread than subordinates ($p < 0.001$), and women tend to send fewer messages than men ($p < 0.001$). The ANOVA results for both *MsgCount* and *MsgRatio* were significant ($p < 0.001$). Tukey's HSD test reveals an interesting picture. Female superiors send significantly fewer messages than everyone else, almost 25% fewer than other groups. In fact, they are the only single group that is different from anyone else. Difference between none of the other groups were significant. For *TokenCount* and *TokenRatio*, the results were similar. Superiors tend to contribute fewer words in the thread than subordinates ($p < 0.001$). Women tend to contribute fewer words than men ($p < 0.01$). The ANOVA test of both features returned to be not significant.

*TokenPerMsg* behaved differently. Gender was not significant at all. That is, men and women did not differ in how long their messages were. In terms of Power, as we saw in Chapter 7, subordinates sent significantly longer emails. The ANOVA test was highly significant. It turns out that among

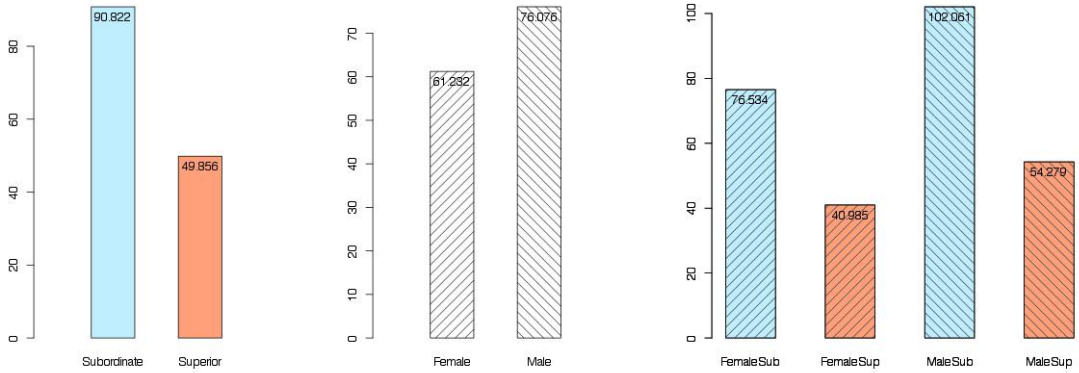



Figure 8.8: Mean value differences along Gender and Power: TokenCount

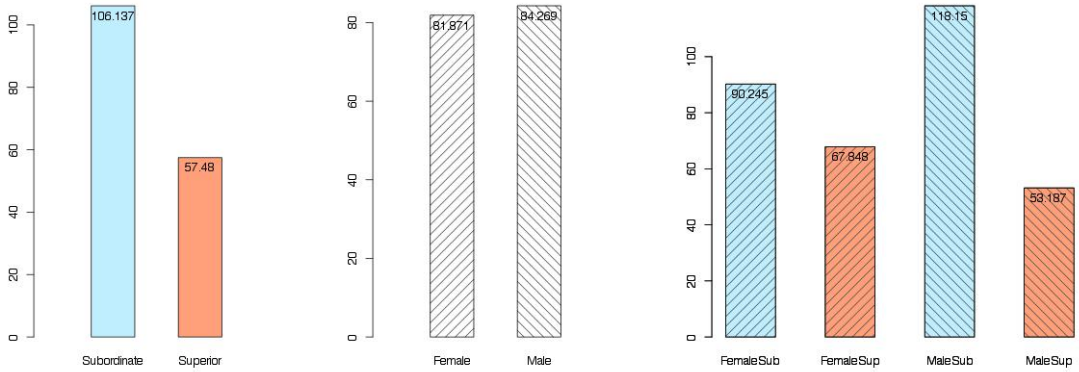



Figure 8.9: Mean value differences along Gender and Power: TokenPerMsg

superiors, there was no significant difference. But among subordinates, male subordinates sent significantly longer emails than female subordinates ($p < 0.01$) as per the Tukey's HSD test. In summary, power is a deciding factor in the difference between the verbosity exhibited by men and women. Female managers send significantly fewer messages than all other groups; both female and male managers send significantly shorter messages than subordinates. On the other hand, female subordinates send significantly shorter emails than male subordinates, although they do not differ in how many messages they send.

### 8.4.3 Thread Structure Features

While the verbosity and positional features measure behavioral aspects, thread structure features are in general dealing with functional aspects (e.g., is a participant in CC (carbon copy) a lot?). While being in CC as a feature might be significantly related to power relations, it is unlikely that someone keeps a person in CC based on their gender. Similarly, adding or removing people to the conversation is also a functional aspect of workplace interactions, and we do not expect gender to play a role there. As expected there was no significant difference between women and men for *InToList%*, *AddPerson*, and *RemovePerson*. The ANOVA test also returned not significant. In other words, gender does not affect the way superiors and subordinates behave in terms of these aspects.

The results from our analysis of *ReplyRate* is interesting. Figure 8.10 shows the mean values for each group. Females get significantly more replies to their messages $p < 0.001$. However, power did not have a significant effect. The ANOVA result was also significant. On further analysis, we find that the female superiors get the highest reply rate ($p < 0.05$). The difference between the *ReplyRate* for male and female subordinates was not significant. It appears to be an interesting finding, since it is an instance of gender of a person with power affecting how others behave towards them. However, on combining this finding with the analysis of *AvgRecipients* and *AvgToRecipients* (Figure 8.11), we find that female superiors on average had more recipients in their messages than any other groups. The difference in *ReplyRate* might also be a manifestation of the fact that female superiors send emails to larger number of people.

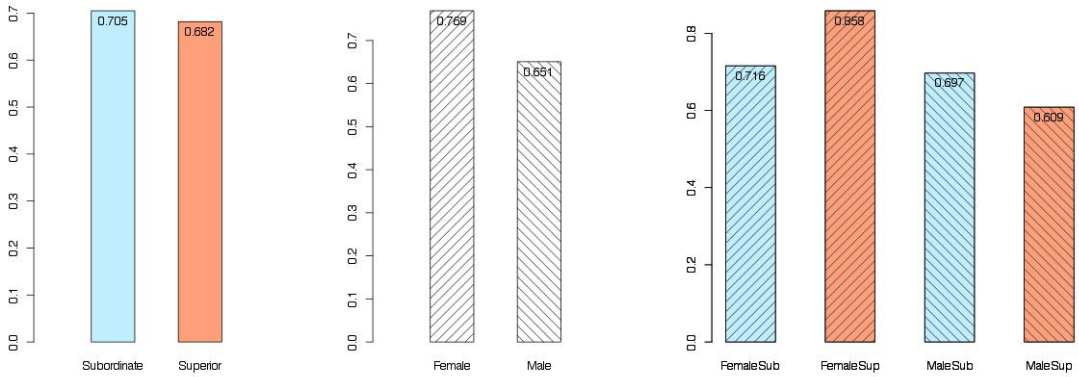



Figure 8.10: Mean value differences along Gender and Power: ReplyRate

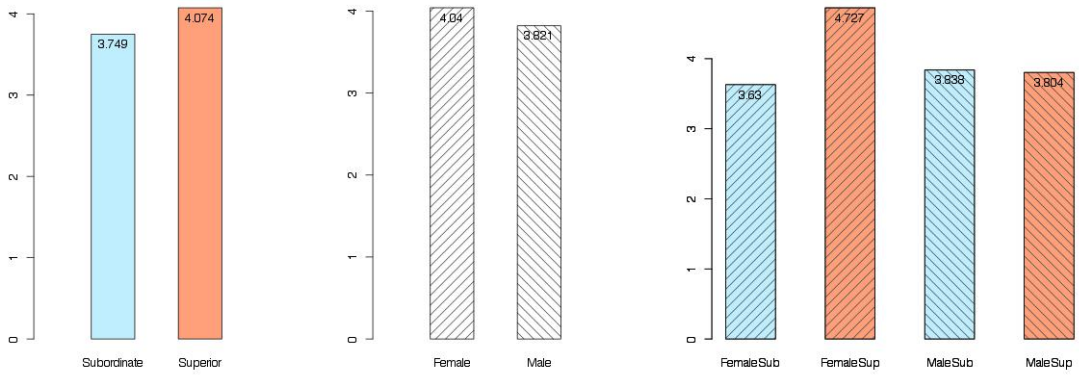



Figure 8.11: Mean value differences along Gender and Power: AvgToRecipients

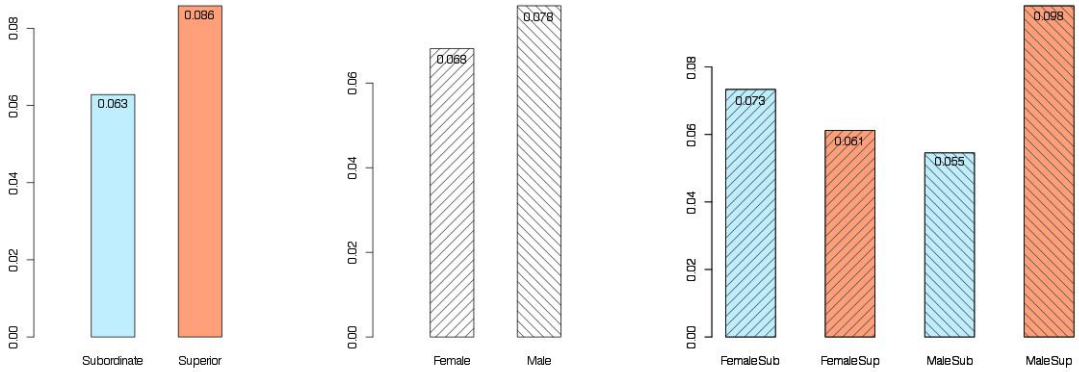



Figure 8.12: Mean value differences along Gender and Power: ReqActionCount

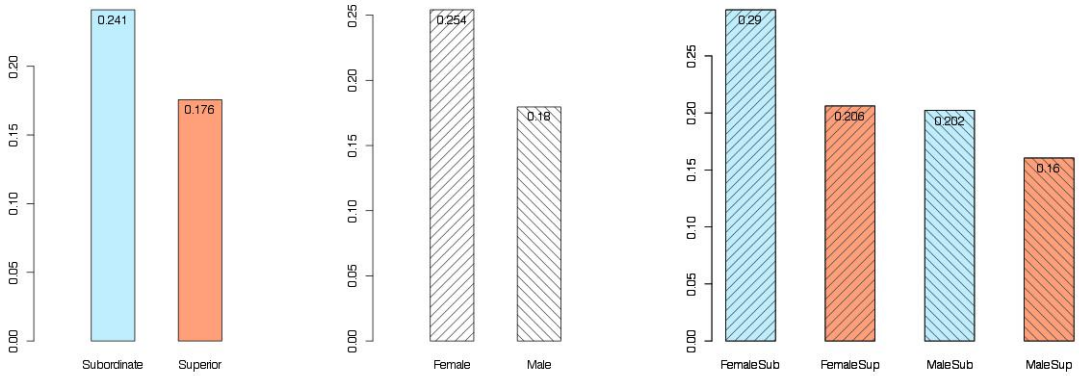



Figure 8.13: Mean value differences along Gender and Power: ReqInformCount

### 8.4.4 Dialog Act Features

We now discuss the finding in terms of dialog act counts. *InformCount* and *ConventionalCount* behaved similarly for all three tests. However, the magnitude of difference between superiors and subordinates for *InformCount* was much higher than that of *ConventionalCount* (superiors had 42.4% lower value than subordinates for *InformCount* as opposed to 13.8% in the case of *ConventionalCount*). The ANOVA test returned not significant, which means that the Gender did not affect the way superiors or subordinates use either conventional or inform dialog acts.

On the other hand, the finding on *ReqActionCount* and *ReqInformCount* are very interesting.

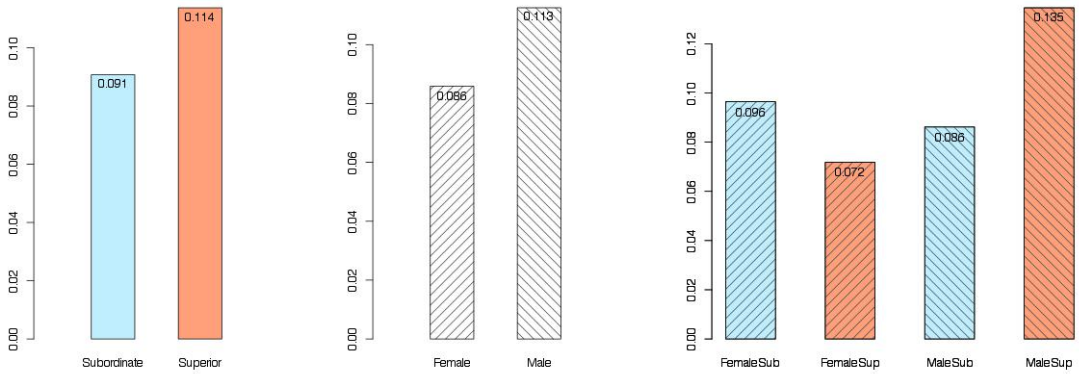



Figure 8.14: Mean value differences along Gender and Power: ODPCount

There was no significant difference between men and women in how often they make requests for action (Figure 8.12), whereas they differed significantly ($p < 0.001$) in terms of how often they request for information. Women issue almost 41% more requests for information than men. The ANOVA test for *ReqActionCount* returned significance ($p < 0.01$), but not for *ReqInformCount*. In other words, Gender affects how superiors and subordinates issue requests for actions, but not requests for information. Male superiors issued more requests for actions than male subordinates, whereas female subordinates held back from making requests. In fact, there was not significant difference between male superiors and female subordinates in terms of *ReqActionCount*. For *DanglingReq%*, there was no significant difference with respect to gender or gender and power together.

### 8.4.5 Overt Displays of Power

Figure 8.14 shows the mean values of ODP counts in each group of participants. The results obtained were similar to what we found for *ReqActionCount*. Both Power and Gender were significant on their own. 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 Power and Gender, however, several differences were very highly significant. Male Superiors use the most ODPs, with an average of 0.135 counts. Somewhat surprisingly, Female Superiors use the *least* of the entire group, with an average of 0.072 counts. However, the differences among Female Su-

periors, Female Subordinates, and Male Subordinates are not significant, as per the Tukey's HSD test.

### 8.4.6 Summary and Discussion

In summary, we find that gender affects the manifestations of power significantly along many structural aspects of interactions. Overall, the gender of the participants do not affect the manifestations of power in positional features (only one ANOVA test returned significance), even though those features differ significantly with respect to both gender and power separately. On the other hand, gender do significantly affect the manifestations of power in verbosity features (of the ANOVA tests we performed on the five verbosity features, three returned to be highly significant) as well as some of the thread structure features (reply rate and number of recipients). Power manifestations on the dialog act based features, especially the request features and overt displays of power were also affected highly significantly by the gender of the participants.

The findings presented in this section do not exhaust the possibilities of this corpus. However, it shows how computational techniques can aid in performing large-scale sociolinguistics analysis. In order to demonstrate this point, we attempted to verify a hypothesis derived from the sociolinguistics literature we consulted. The 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 (Kendall 2003).

We recall that our notion of overt display of power (ODP) is a face-threatening communicative strategy (Chapter 6, Section 6.2, page 75). An ODP limits the addressee's range of possible responses, and thus threatens his or her (negative) face.[4] 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 saw in the results presented in Section 8.4.5 that this hypothesis is indeed true. We found that female superiors used the least number of ODPs among all groups. The results confirmed our hypothesis: female superiors use fewer ODPs than male superiors. However, we also see that among

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

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. In other words, the fact that superiors use more ODPs than subordinates is entirely due to male superiors using more ODPs. Similarly, the fact that men use more ODPs than women is also entirely due to superiors among men using significantly more ODPs.

## 8.5 Notion of Gender Environment

The notion of "gender environment" refers to the gender composition of a group who are communicating. Holmes and Stubbe (2003) use the term gender environment to refer 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 $t$, $P_t$, and exclude $p$ from it: $P_t \setminus \{p\}$. We then calculate the percentage of females in this set.[5] We obtain three gender environments by setting thresholds on these percentages (dividing equally): Female Environment, Mixed Environment, and Male Environment. Across all threads, we have 791 female, 2087 mixed and 1642 male gender environments.

[5]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.

- **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%

## 8.6 Statistical Analysis: Gender Environment and Power

In this section, we present our investigation on whether the manifestations of power differs based on the gender environment. As in Section 8.4, we use the ANOVA test to assess the statistical significance of differences. We perform ANOVA tests on all features keeping both Power and Gender Environment (GenderEnv, hereafter) as independent variables. We also perform ANOVA keeping GenderEnv alone as the independent variable; since GenderEnv has more than two groups, we cannot use Student's t-Test. We list the results obtained in all the statistical tests in Appendix B and discuss the main findings in each set of features below.

### 8.6.1 Positional Features

For the positional features, any difference that we see in the feature values between different gender environments is not interesting. For example, it is not sensible to investigate whether the value of *Initiator* is different between gender environments (all threads had to be initiated by someone). However, it is still interesting to see whether there is any connection between the gender environment and how the superiors and subordinates differed in terms of when they started and stopped participating in the threads.

As we saw in Section 8.4, subordinates initiate more emails than superiors (*Initiator*) and overall start participating earlier in the thread (*FirstMsgPos*). The ANOVA test keeping Power and GenderEnv as independent variables was highly significant ($p < 0.001$). In other words, the gender environment does affect the initiative shown by subordinates in starting email threads. Figure 8.15 shows the mean values of each group. Subordinates do start participating in the threads significantly earlier than superiors. However, the magnitude of this difference was dependent on the gender environment. This suggests that subordinates tend to show more initiative in female environments than other gender environments, and that superiors tend to start participating in the threads much

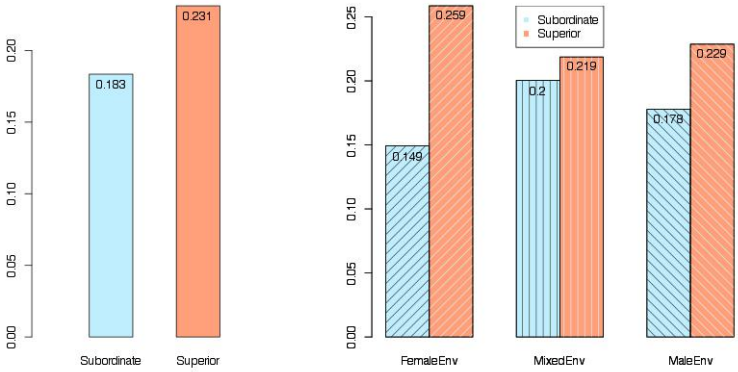



Figure 8.15: Mean value differences along Gender Environment and Power: FirstMsgPos

later in female environments. For the relative position of last message, the ANOVA results were not significant.

### 8.6.2 Verbosity Features

As per the ANOVA results, the gender environment has no significance in *MsgCount* or in how Power is manifested in *MsgCount*. On the other hand, in terms of *TokenCount*, there was a significant difference ($p < 0.01$) across gender environments (Figure 8.16). The ANOVA test keeping Power and GenderEnv as independent variables also returned significance ($p < 0.001$). In fact, in male environments, there was no significant difference in *TokenCount* between superiors and subordinates. Subordinates behaved more or less the same across the gender environments, but superiors contributed much less in female and mixed environments. A similar pattern is also observed in *TokenPerMsg* across different gender environments.

### 8.6.3 Thread Structure Features

The effect of gender environment on *ReplyRate* was minimal. We observed that the number of recipients (both *AvgRecipients* and *AvgToRecipients*) was significantly higher in the mixed environment than others. This, however, is another artifact of how our corpus is constructed. In a thread with large number of participants, it is more likely to have a mixed environment than either male or female environment. The ANOVA test keeping Power and GenderEnv also returned no signifi-

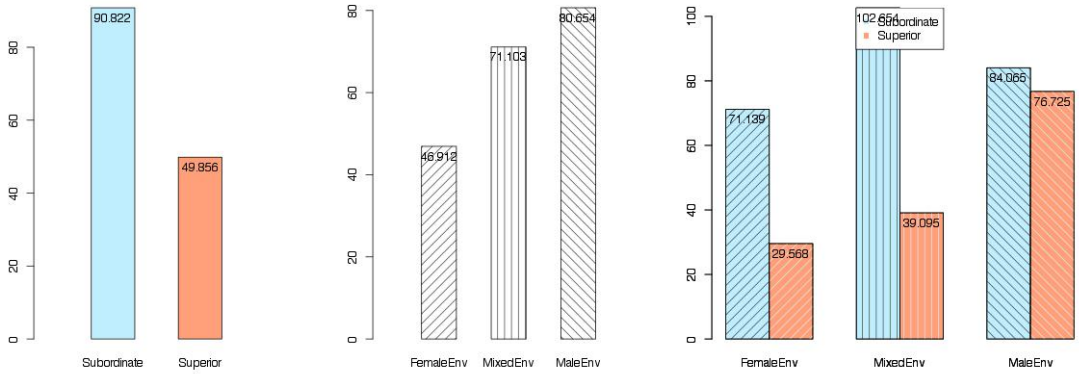



Figure 8.16: Mean value differences along Gender Environment and Power: TokenCount

cance for *AddPerson* and *RemovePerson*. In summary, the effect of gender environment on thread structure features was minimal.

### 8.6.4 Dialog Act Features

The results obtained on the ANOVA tests for the dialog act features were interesting. Lets start with the *ConventionalCount*. Figure 8.17 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). While the averages by GenderEnv differ, the differences are not significant. However, the groups defined by both Power *and* GenderEnv 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. In the Tukey HSD test, the only significantly different pairs are exactly the set of subordinates in female environments paired with each other group. 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. We interpret this result to mean that subordinates are more comfortable in female environments to use a style of communication which includes more conventional dialog acts than outside the female environments.

The ANOVA tests for *InformCount* also returned high significance. The difference between

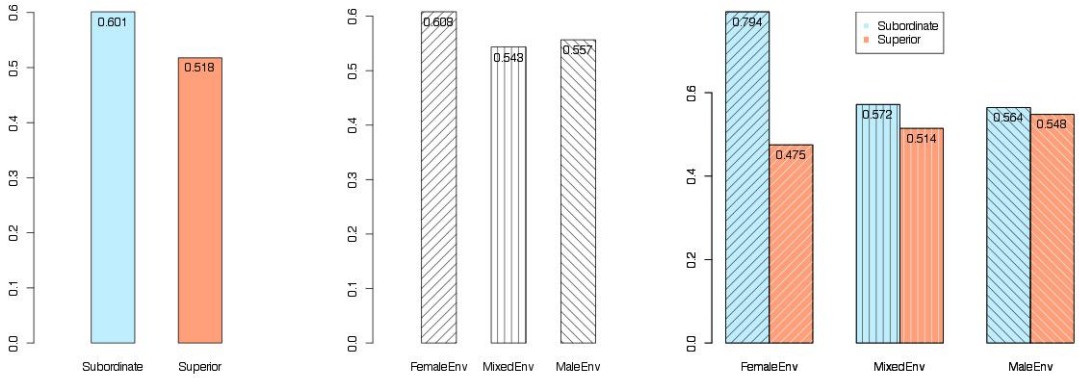



Figure 8.17: Mean value differences along Gender Environment and Power: ConventionalCount

mean values of *InformCount* feature in male environments and mixed environments were not significant; but it differed significantly between female environments and both male and mixed environments. The groups defined by both Power *and* GenderEnv also have highly significant differences. There was no significant difference between superiors' and subordinates' count of inform dialog acts when operating in a male environment. In other words, the finding that subordinates use more inform dialog acts holds true only in female and mixed environments, but not in male environments. However, on comparing this result with our findings in terms of verbosity features (Figure 8.16), we find that this is in fact an artifact of most of the contributions being inform statements (the findings in *InformCount* mirror that of *TokenCount*).

The ANOVA results for both *ReqActionCount*, *ReqInformCount*, and *DanglingReq%* were not significant when tested using Power *and* GenderEnv. The male environment had a significantly ($p < 0.05$) lower *DanglingReq%*.

### 8.6.5 Overt Displays of Power

The results of the ANOVA analysis on *ODPCount* are interesting. Figure 8.19 shows the mean values of each group. As we saw already in Chapter 7 (Section 7.3) and in Section 8.4, superiors use significantly more overt displays of power than subordinates. However, this pattern varied across gender environments significantly. The same relationship holds only in a mixed gender environment, where also most of the ODP occur. In male environments, there was no significant

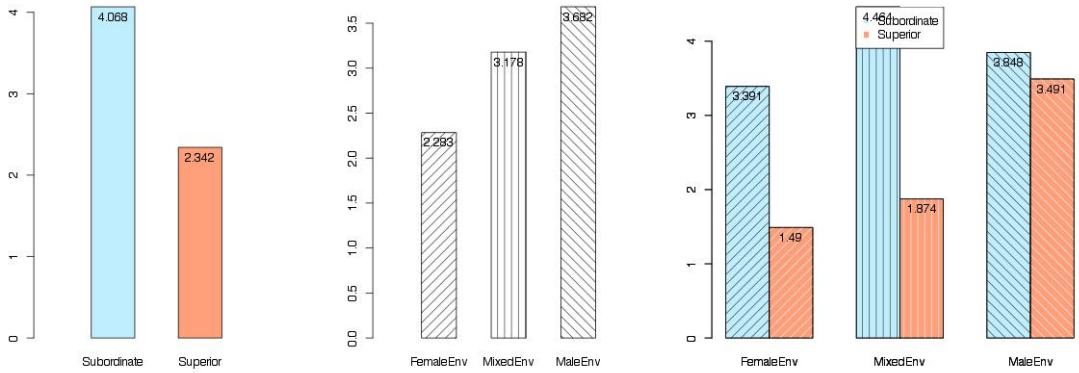



Figure 8.18: Mean value differences along Gender Environment and Power: InformCount

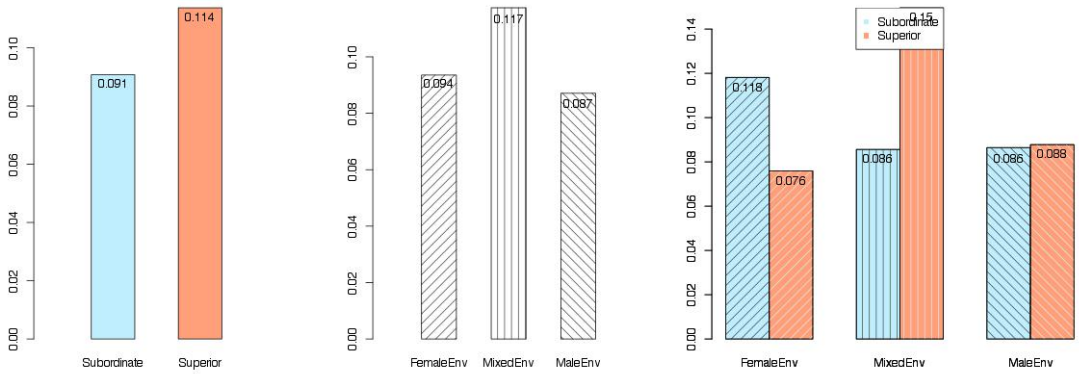



Figure 8.19: Mean value differences along Gender Environment and Power: ODPCount

difference in *ODPCount* between superiors and subordinates, whereas in female environments, the value of *ODPCount* for superiors was significantly lower than that of subordinates. This goes in line with our finding in Section 8.4 that female managers use fewer overt displays of power.

### 8.6.6 Summary and Discussion

In summary, we find that gender environment also affects the manifestations of power significantly along different structural aspects of interactions. The gender environment significantly affects the difference between how often superiors and subordinates initiate email threads. In terms of verbosity, we found that the gender environment affected how much the subordinates contribute in the email threads. The effect of gender environment on manifestations of power along thread structure features was minimal. The power manifestations on the inform and conventional dialog act features also differed significantly across different gender environments. We also found that the frequency of overt displays of power was significantly lower in female environment.

Similar to what we did in Section 8.4.6, we attempt to verify a hypothesis derived from the sociolinguistics literature we consulted in relation to the notion of gender environment. The hypothesis we investigate is:

- **Hypothesis 2**: Women when talking among themselves 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 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. We make our Hypothesis 2 more precise by saying that a higher number of conventional dialog acts will be used in female environments.

We presented the results of our analysis of *ConventionalCount* feature in Section 8.6.4. Our results first appears to be a negative result: while the averages by Gender Environment differ, the

differences are not significant. However, we found that 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. 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.

## 8.7 Utility of Gender Information in Predicting Power

In this section, we investigate the utility of the gender information in the problem of predicting the direction of power that we presented in Chapter 7. The SVM-based supervised learning system presented in Section 7.5 uses a quadratic kernel, which we expect will capture the interdependence between dialog structure features and gender features that we found in our statistical analysis presented in Section 8.4 and Section 8.6.

We perform our experiments on the ENRON-APGI subset, training a model using the same machine learning framework presented in Section 7.5 using the related interacting participant pairs in the *Train* subset of ENRON-APGI, and choosing the best model based on performance on the *Dev* subset. We experimented using all subsets of features discussed in Chapter 7 (Section 7.3). In addition, we add two gender-based feature sets: GENDER containing the gender of both persons of the pair and GENDERENV which is a singleton set with the gender environment as the feature. Table 8.5 presents the results obtained using various feature combinations. Note that the numbers presented in Table 8.5 are not directly comparable to the results presented in Table 7.6 (Chapter 7, Section 7.5, page 107), since the results presented there are on the *Dev* set of the ENRON-LARGE corpus, whereas here we discuss results obtained on the *Dev* set of the ENRON-APGI, which is a subset of around 50% of the ENRON-LARGE corpus.

The majority baseline obtains an accuracy of 55.8%. Using the gender-based features alone performs only slightly better than the majority baseline, posting an accuracy of 57.6%. The best performance is obtained using a combination of LEXICAL, THREAD STRUCTURE, GENDER and GENDERENV, which posts an accuracy of 70.7%. Removing the GENDERENV feature set de-

| | Description | Accuracy |
|---|---|---|
| Baselines | Majority | 55.83 |
| Using gender features alone | GEN | 57.59 |
| | GEN + ENV | 57.59 |
| Best feature sets | LEX + THR + GEN + ENV | **70.74** |
| | LEX + THR + GEN | 70.46 |
| | LEX + THR | 68.24 |
| | LEX + THR + PST + VRB | 68.33 |
| Best without LEXICAL | DA + ODP + THR + GEN | 67.31 |
| | DA + ODP + THR | 64.63 |
| Best with no content | PST + VRB + THR + GEN | 66.57 |
| | PST + VRB + THR | 62.96 |

Table 8.5: Results on using gender features for power prediction.
PST: POSITIONAL, VRB: VERBOSITY, THR: THREAD STRUCTURE,
DA: DIALOG ACTS, ODP: OVERT DISPLAY OF POWER, LEX: LEXICAL,
GEN: GENDERENV: GENDERENV

creases the accuracy marginally to 70.5%, whereas removing the GENDER features as well reduces the performance significantly to 68.2%. This reduction of 2.4% percentage points in accuracy shows that gender features are in fact useful for this power prediction task. The best performance feature set without using any gender information is the combination of LEXICAL, THREAD STRUCTURE, POSITIONAL and VERBOSITY, which reports an accuracy of 68.3%. The best performing feature set without using LEXICAL is the combination of DIALOG ACTS, OVERT DISPLAY OF POWER, THREAD STRUCTURE and GENDER (67.3%). Removing the gender features from this reduces the performance to 64.6%. Similarly, the best performing feature set which do not use the content of emails at all is POSITIONAL + VERBOSITY + THREAD STRUCTURE + GENDER (66.6%). Removing the gender features decreases the accuracy by a larger margin (5.4% accuracy reduction to 63.0%).

It is interesting to look at the error reduction obtained by adding gender features to different feature sets. Using gender features alone obtains only an error reduction of 4.0% over the majority baseline (i.e., without using any other features). However, the predictive value of gender features improve considerably when paired with other features. For the best feature set we obtained, the gender features contributed to an error reduction of 7.9% (68.2% to 70.7%). For the best feature set without using LEXICAL also the gender features contributed a similar error reduction of 7.6% (64.63% to 67.3%). For the setting where no content features were used, gender features obtained an even higher error reduction of 11.0% (63.0% to 66.6%). In other words, 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. 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.

On our blind test set, the majority baseline obtains an accuracy of 57.9% and the baseline system that does not use gender features obtains an accuracy of 68.9%. On adding the gender-based features, the accuracy of the system improves to 70.3%.

## 8.8 Conclusion

The first contribution of this chapter is the new, freely available resource — Gender Identified Enron Corpus, an extension to the Enron email corpus with 87% of the email senders' gender identified. We used the Social Security Administration's baby-names database to automatically assess the gender ambiguity of first names of email senders and assigned the gender to those whose names were highly unambiguous. Our gender identified corpus is orders of magnitude larger than other existing resources in this domain that capture gender information. We expect it to be a rich resource for social scientists interested in the effect of power and gender on language use.

Our second contribution is the detailed statistical analysis of the interplay of gender, gender environment and power in how they affect the dialog behavior of participants of an interaction. We introduced the notion of gender environment to capture the gender makeup of the discourse participants of a particular interaction. We showed that gender and gender environment affect the ways power is manifested in interactions in complex ways, resulting in patterns in the discourse

that reveal the underlying factors. While our findings pertain to the Enron email corpus, 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.

Finally, we showed the utility of gender information in the task of predicting the direction of power between pairs of participants based on single threads of interactions. We obtained statistically significant improvements by adding the gender of both participants of a pair as well as the gender environment as features to a system trained using lexical and dialog structure features alone.

# Chapter 9

# Levels of Expressed Beliefs and Power

There is a rich tradition of modeling dialog participants' cognitive states (e.g., (Bratman 1987, Rao and Georgeff 1991)) and relating this modeling to the way their cognitive states are expressed in language through extensions to speech act theory (e.g., (Perrault and Allen 1980, Clark 1996, Bunt 2000)). This line of study has also benefited the task of dialog act tagging (e.g., (Stolcke et al. 2000)), which is one of the ways we model dialog behavior of interactants in this thesis. In this chapter, we investigate further into how the dialog participants signal their beliefs using language, and the strength of their beliefs; this latter point is not usually included in dialog act tagging. Our notion of belief stems from the idea that there is more to "meaning" expressed in language than just propositional content. Consider the following sentences.

(1) a. *John will submit the report on-time.*

b. *John may submit the report on-time.*

c. *Sara says John will submit the report on-time.*

d. *I wish John would submit the report on-time.*

e. *Will John submit the report on time?*

All the above sentences contain the proposition SUBMIT(JOHN,REPORT,ON-TIME) However they allow us different inferences about the level of belief the author expresses towards the truth value of the proposition. The author is committed to the proposition in (1a) and wants the reader/hearer to believe that John will submit the report on-time, whereas in (1b) and (1c) the author is not

committing to it. In (1b), she is explicitly signaling her lack of commitment using the word *may*, where as in (1c), she is attributing the belief to someone else. In (1d) and (1e), the author does not tell us anything about whether anyone believes whether John will submit the report on-time or not. Diab et al. (2009) released a corpus called Language Understanding (LU) corpus that contains annotations for different types of beliefs expressed in text.

In sociolinguistics studies, there is evidence that expressions of non-committedness in text do correlate with social power relations. For example, O'Barr (1982) identifies linguistic markers such as hedges as indicators of power-less language, i.e., language used by people with lower power in an interaction. In this chapter, our main objective is to study how the different expressions of beliefs as captured in annotations by Diab et al. (2009) correlate with the social power relations between interactants. For example, do subordinates express relatively more non-committed beliefs in their messages, or do they use more reported beliefs? Since Diab et al. (2009) model belief in a more general semantic framework than hedges, it allows us to capture differences in expressions of belief that goes beyond just non-committedness. We first build an automatic committed belief tagger using the annotations by Diab et al. (2009) that can detect different levels of belief expressed in text, and use the automatically obtained belief tags to perform our analysis of how they correlate with power. We also show how we can incorporate belief information to improve the performance of a power prediction system.

In this chapter, we describe in detail the general notion of committed belief, as well as our investigation of how it correlates with power. Although the research on detecting belief was initiated as part of this thesis (Prabhakaran et al. 2010), it is currently an active research area (Werner et al. 2015, Prabhakaran et al. 2015) involving researchers interested in its applications to traditional NLP problems such as semantics, information extraction, and knowledge-base population. Consequently, there are multiple versions of committed belief analysis frameworks, developed as part of different parallel efforts, some of which are only partially part of this thesis. When studying its correlates with power, we use both the original tagging framework we developed (which follows a 3-way belief distinction) as well as a more advanced tagging framework (which follows a 4-way belief distinction); the latter has contributions from other researchers as well.

Section 9.1 situates the notion of committed belief among other closely related notions, and discusses studies on how these notions correlate with power. Section 9.2 and 9.3 describe the commit-

ted belief annotations and tagging frameworks we use. Section 9.4 presents the statistical analysis of how different types of beliefs correlate with power relations. Section 9.5 discusses different ways of incorporating the belief tags into the machine learning framework for power prediction from Chapter 7. Section 9.6 concludes the chapter.

## 9.1 Related Work

Our notion of belief is closely related to factuality, hedging, veridicality, and modality. We first discuss these concepts and how our notion of belief relates to them.

**Relation with factuality** A closely related corpus is FactBank (FB; Saurí and Pustejovsky (2009)), which captures factuality annotations on top of event annotations in TimeML. FactBank is annotated on the genre of newswire. FactBank models the factuality of events at three levels: certain (CT), probable (PB) and possible (PS), and distinguishes the polarity (e.g., CT- means certainly not true). Moreover it marks an unknown category (Uu), which refers to uncommitted or underspecified belief. It also captures the source of the factuality assertions, thereby distinguishing the SW's factuality assertions from those of a source introduced by the author. Despite the terminology difference between FactBank ("factuality") and LU Corpus ("committed belief"), they both address the same type of linguistic modality phenomenon, namely level of committed belief. FactBank differs from the LU corpus in two major respects (other than the granularity in which they capture annotations): 1) FactBank is roughly four times the size of the LU corpus, and 2) FactBank is more homogeneous in terms of genre than the LU corpus as it consists primarily of newswire.

**Relation with Hedging** Hedging and uncertainty are very closely related to our notion of non-committed belief. There is much work within NLP community on uncertainty detection such as hedging and use of weasel words. There has been an open evaluation as part of the CoNLL shared task in 2010 to detect uncertainty in language (Farkas et al. 2010). Prokofieva and Hirschberg (2014) define hedges as words or phrases that add ambiguity or uncertainty (Propositional Hedges) or show the speaker's lack of commitment to a proposition (Relational Hedges). For example, *The ball is* **sort of** *blue* contains a Relational Hedge (*sort of*) and *I* **think** *the ball is blue* includes a propositional hedge (*think*). Propositional hedges indicate non-committed belief. While belief and

hedging are closely related, we see the belief/factuality annotation as more general than hedging (since it does not only include non-committed belief), and also more semantic (since we are not identifying language use but underlying meaning). The later version of our committed belief tagger uses hedge based features and shows that it improves the performance of identifying non-committed beliefs.

**Relation with modality** The term "modality" is used in formal semantics as well as in descriptive linguistics. Many semanticists (e.g. (Kratzer 1991, Kaufmann et al. 2006)) define modality as quantification over possible worlds. Modality can be of two types: epistemic, which qualifies the speaker's commitment, and deontic, which concerns freedom to act. Belief/factuality falls under epistemic modality. Another view of modality relates more to a speaker's attitude toward a proposition (e.g. (McShane et al. 2004, Baker et al. 2010, Prabhakaran et al. 2012a)), which is closer to the way we model belief. For us, belief is one of the modalities a speaker/writer expresses.

**Relation with veridicality** We interpret the term "veridical" as referring to a property of certain words (usually verbs), namely to mark the proposition expressed by their syntactic complement clause as firmly believed (committed belief) by the writer (Kiparsky and Kiparsky 1970). Veridicality as a property of lexical or lexico-syntactic elements is thus a way of relating belief/factuality to linguistic means of expressing them, but we take the notion of belief/factuality as being the underlying notion.

## 9.2 Committed Belief Annotations

For the work presented in this chapter, we use two different sets of committed belief annotations. The first one uses a 3-way belief distinction presented in the LU (Language Understanding) Corpus (Diab et al. 2009) — COMMITTEDBELIEF (CB), NONCOMMITTEDBELIEF (NCB), and NONAPPLICABLE (NA). In later work, this was extended to a 4-way distinction presented in the DEFT[1] Corpus (Prabhakaran et al. 2015) — COMMITTEDBELIEF (CB), NONCOMMITTEDBELIEF (NCB), REPORTEDBELIEF (ROB), and NONAPPLICABLE (NA) — in which the original NONCOMMITTEDBELIEF class was split into NONCOMMITTEDBELIEF and REPORTEDBELIEF. In this chapter,

[1]The DARPA program on Deep Exploration and Filtering of Text

| | LU Corpus | DEFT Corpus |
|---|---|---|
| Genre | Newswire, Emails, Instruction manuals | Discussion forums |
| Size | 13K words | 850K words |
| Tags | {CB, NCB, NA} | {CB, NCB, ROB, NA} |

Table 9.1: Differences between LU and DEFT bommitted belief annotations.

we describe both corpora and the associated annotation schema. The difference between both corpora is summarized in Table 9.1.

### 9.2.1 LU Corpus Annotations

The LU Corpus annotations capture whether a speaker/writer (SW) intends the reader to interpret a stated proposition as the writer's strongly held belief, as a proposition which the writer does not believe strongly (but could), or as a proposition towards which the writer does not express a belief, but rather a different cognitive attitude, such as desire or intention.

#### 9.2.1.1 Data and Source

The LU Corpus is relatively small in size, but spans different domains and genres such as newswire, blogs, instruction manuals, email threads, letters, and transcribed dialog data. The corpus contains around 13K word tokens annotated for speaker belief of stated propositions. Around 70% of the corpus was doubly annotated and they report an inter-annotator agreement of 95.8%. For more details on the data, see (Diab et al. 2009).

#### 9.2.1.2 Annotations

The corpus annotates each verbal proposition (clause or small clause), by attaching one of the following three tags to the head of the proposition (verbs and heads of nominal, adjectival, and prepositional predications).

**Committed belief (CB):** the writer strongly believes that the proposition is true, and wants the reader/hearer to believe that. Examples:

(2) a. John will **submit** the report on-time.

**Non-committed belief (NCB):** the writer identifies the proposition as something which he or she could believe, but he or she happens not to have a strong belief in. There are two sub-cases. First, the writer makes clear that the belief is not strong, for example by using an epistemic modal auxiliary (3a). Second, in reported speech, the writer is not signaling to the reader what he or she believes about the reported speech (3b). Examples:

(3) a. John may **submit** the report on-time.

b. Sara says John will **submit** the report on-time.

**Non-belief propositions (NA):** – the writer expresses some other cognitive attitude toward the proposition, such as desire or intention (4a), or expressly states that he or she has no belief about the proposition (e.g., by asking a question (4b)). In other words, the proposition does not have a truth value in this world (be it in the past or in the future). Examples:

(4) a. I wish John would **submit** the report on-time

b. Will John **submit** the report on-time?

### 9.2.2 DEFT Corpus Annotations

The DEFT Corpus extends the 3-way belief distinction in LU Corpus to a 4-way scheme. The DEFT Corpus annotations capture whether a speaker/writer (SW) intends the reader to interpret a stated proposition as the writer's strongly held belief, as a proposition which the writer does not believe strongly (but could), as a proposition the writer is reporting someone else's belief about, or as a proposition towards which the writer does not express a belief, but rather a different cognitive attitude, such as desire or intention.

#### 9.2.2.1 Data and Source

The DEFT corpus consists of English text from discussion forum threads from a wide variety of sites collected as part of the DARPA BOLT program. The discussions are usually about current events

or personal anecdotes. The corpus contains around 850K words. They report an inter-annotator agreement on headword selection of 93% and agreement on belief type labeling of 84%.

#### 9.2.2.2 Annotations

Same as in LU Corpus, the DEFT corpus annotates heads of all (clausal) propositions in each document with a four-way belief type distinction, with the following categories.

**Committed belief (CB):** the writer strongly believes that the proposition is true, and wants the reader/hearer to believe that. Examples:

(5) a. John will **submit** the report on-time.

**Non-committed belief (NCB):** the writer identifies the proposition as something which he or she could believe, but he or she happens not to have a strong belief in, for example by using an epistemic modal auxiliary. Examples:

(6) a. John may **submit** the report on-time.

**Reported belief (ROB):** the writer attributes belief (either committed or non-committed) to another person or group. Note that this label is only applied when the writer's own belief in the proposition is unclear. Examples:

(7) a. Sara says John will **submit** the report on-time.

**Non-belief propositions (NA):** – the writer expresses some other cognitive attitude toward the proposition, such as desire or intention (8a), or expressly states that he or she has no belief about the proposition (e.g., by asking a question (4b)). In other words, the proposition does not have a truth value in this world (be it in the past or in the future). Examples:

(8) a. I wish John would **submit** the report on-time

b. Will John **submit** the report on-time?

### 9.2.3 Annotation Details

For both the LU Corpus and the DEFT corpus, the following are true:

- The annotations capture only the belief of the speaker/writer (SW).
- The annotations identify the target propositions about which a belief is expressed, not the linguistic markers that trigger a particular kind of belief.
- The annotations do not mark the text spans of propositions; they propositional heads as one of the belief classes and all other tokens as 'O' (Other).
- Event nominals (such as *the on-time submission by John was unexpected*) are not annotated for belief and are always marked 'O'.
- The syntactic form does not determine the annotation, but the perceived writer's intention – a question will usually be an NA, but sometimes a question can be used to convey a belief (for example, a rhetorical question), in which case it would be labeled CB.
- The annotations do not capture any cognitive attitudes expressed about a proposition other than belief. A proposition tagged as CB may also have other cognitive attitudes expressed about them (e.g., in 'John managed to submit the report on-time", the author is expressing CB towards the proposition *submit*, but also the *success* modality (Prabhakaran et al. 2012a)); but the annotations capture only the former.
- The annotations do not annotate subjectivity (Wiebe et al. 2004, Wilson and Wiebe 2005), nor opinion (e.g., (Somasundaran et al. 2008)).
- The annotations do not evaluate the truth value of the propositions, only the expressed level of belief in them held by the writer. Thus a strongly held false belief would not appear any different from a strongly held true belief.
- The annotations take expressed beliefs at "face value" and do not capture deception, sarcasm, irony, and other cases where the writer's internal belief may differ from the expressed belief.

The distribution of belief tags in both corpora are summarized in Table 9.2. In the LU corpus, around 10.4% of words were identified as propositional head words, whereas in the DEFT corpus, this was around 16.8%. The per-class distributions also vary between both corpora. For example, NCB accounted for 12.6% of the propositional heads in the LU Corpus, whereas this was much

| Corpus | # of Words | CB | NCB | ROB | NA |
|---|---|---|---|---|---|
| LU Corpus | 13,485 | 1396 (10.4%) | | | |
| | | 631 (45.2%) | 176 (12.6%) | | 589 (42.2%) |
| DEFT Corpus | 852,836 | 143,240 (16.8%) | | | |
| | | 79,995 (55.8%) | 3,890 (2.7%) | 7,150 (5.0%) | 52,205 (36.4%) |

Table 9.2: Belief tag distribution in the LU and DEFT corpora.

lower in the DEFT corpus: for NCB and ROB combined, it was only 7.7% (in the LU corpus, these two were together called NCB). The difference in proportions of belief tags is not surprising because both corpora differ in terms of their genres. The LU Corpus is a multi-genre corpus which has mostly well written or edited text (newswire, letters, instruction manuals, etc.), whereas the DEFT corpus is from discussion forums.

## 9.3 Automatic Committed Belief Tagging

In this section, we describe the three different committed belief tagging systems we use for analysis in this chapter — CB3-TAGGER, CB3-TAGGERPLUS, and CB4-TAGGER. CB3-TAGGER is the initial tagger developed using the committed belief annotations in the LU Corpus and was described in detail in (Prabhakaran et al. 2010). CB3-TAGGERPLUS is a reimplementation of that system, which uses the same data and more or less the same features. The reimplementation was done in order to seamlessly integrate the system into the power analysis framework that is the core of this thesis. CB3-TAGGERPLUS reports better performance than the original published results reported by CB3-TAGGER in (Prabhakaran et al. 2010). CB4-TAGGER is an extension of CB3-TAGGERPLUS developed by (Werner et al. 2015). CB4-TAGGER is trained on the much bigger DEFT Corpus, and has the capability to detect the fourth category of committed beliefs (i.e., ROB). We describe CB3-TAGGER and CB3-TAGGERPLUS in detail in Section 9.3.1 and Section 9.3.2. We summarize CB4-TAGGER in Section 9.3.3; for more details refer to (Werner et al. 2015). Table 9.3 summarizes the three taggers and highlights in what aspects they differ. A "$\longrightarrow$" sign after a factor denotes that it is a factor that changed for the next iteration.

| CB3-TAGGER<br>Original 3-way CB Tagger | CB3-TAGGERPLUS<br>Re-implemented Tagger | CB4-TAGGER<br>Re-trained 4-way Tagger |
|---|---|---|
| **Task:**<br>3-way {CB, NCB, NA}<br>**Data:**<br>LU Corpus (13K words)<br>**Implementation:** ⟶<br>Perl<br>**Parser:** ⟶<br>MICA<br>**ML Algorithm(s):** ⟶<br>Yamcha, CRF (Mallet)<br>**Features:** ⟶<br>• lexical features<br>• dependency features | **Task:** ⟶<br>3-way {CB, NCB, NA}<br>**Data:** ⟶<br>LU Corpus (13K words)<br>**Implementation:**<br>Java, UIMA, ClearTk<br>**Parser:**<br>Stanford CoreNLP<br>**ML Algorithm(s):**<br>SVMLight<br>**Features:** ⟶<br>• lexical features<br>• dependency features<br>• conditional features | **Task:**<br>4-way {CB, NCB, NA, ROB}<br>**Data:**<br>DEFT Corpus (850K words)<br>**Implementation:**<br>Java, UIMA, ClearTk<br>**Parser:**<br>Stanford CoreNLP<br>**ML Algorithm(s):**<br>SVMLight<br>**Features:**<br>• lexical features<br>• dependency features<br>• conditional features<br>• word cluster features<br>• hedge features |

Table 9.3: Comparison of different committed belief taggers.

### 9.3.1 CB3-TAGGER

We applied a supervised learning framework to the problem of identifying committed belief in context. Our task consists of two conceptual subtasks: identifying the propositions, and classifying each proposition as CB, NCB, or NA. For the first subtask, we could use a system that cuts a sentence into propositions, but we are not aware of such a system that performs at an adequate level. Instead, we tag the heads of the proposition, which amounts to the same in the sense that there is a bijection between propositions and their heads. Practically, we have the choice between a joint model, in which the heads are chosen and classified simultaneously, and a pipeline model, in which heads are chosen first and then classified. We consider the joint model in detail. Section 9.3.1.3, we present results of the pipeline model; they support our choice.

In the joint model, we define a four-way classification task where each token is tagged as one

of four classes – COMMITTEDBELIEF, NONCOMMITTEDBELIEF, NONAPPLICABLE, or OTHER. In this section, we describe the experiments we conducted using two machine learning algorithms for this tagging task: Support Vector Machines (SVM) and Conditional Random Fields (CRF). For SVM, we use the YAMCHA (Kudo and Matsumoto 2005) sequence labeling system,[2] which uses the TinySVM package for classification.[3] For CRF, we used the linear chain CRF implementation of the MALLET (McCallum 2002) toolkit.[4] We start by describing the features we use for building the CB3-TAGGER.

#### 9.3.1.1 Features

We divide our features into two types - Lexical and Syntactic. Lexical features are at the token level and can be extracted without any parsing with relatively high accuracy. We expect these features to be useful for our task. For example, *isNumeric*, which denotes whether the word is a number or alphabetic, is a lexical feature. Syntactic features of a token access its syntactic context in the dependency tree. For example, *parentPOS*, the POS tag of the parent word in the dependency parse tree, is a syntactic feature. For the experiments and results discussed in this section, we used the MICA deep dependency parser (Bangalore et al. 2009) for parsing in order to derive the syntactic features. We use MICA because we expect that the predicate-argument structure of the verbs, which is explicit in the MICA output, would be helpful for this task.

The list of features we used in our experiments are summarized in Table 9.4. The column 'Type' denotes the type of the feature. 'L' stands for lexical features and 'S' stands for syntactic features. For finding the best performing features, we did a search on the entire feature space, incrementally pruning away features that are not useful. For example, the token's supertag (Bangalore and Joshi 1999), the parent token's supertag, and a binary feature *isRoot* (Is the word the root of the parse tree?) were deemed not useful. We list the features we experimented with and decided to discard in bottom section of Table 9.4.

Table 9.5 presents some of these dependency features for the sentence *Republican leader Bill Frist said the Senate was hijacked*. Figure 9.1 shows the corresponding dependency parse. For this

[2]http://chasen.org/ taku/software/YAMCHA/

[3]http://chasen.org/ taku/software/TinySVM/

[4]http://mallet.cs.umass.edu/

| No | Feature | Type | Description |
|---|---|---|---|
| | | | Features that performed well |
| 1 | isNumeric | L | Word is Alphabet or Numeric? |
| 2 | POS | L | Word's POS tag |
| 3 | verbType | L | Modal/Aux/Reg ( = 'nil' if the word is not a verb) |
| 4 | whichModalAmI | L | If I am a modal, what am I? ( = 'nil' if I am not a modal) |
| 3 | amVBwithDaughterTo | S | Am I a VB with a daughter *to*? |
| 4 | haveDaughterPerfect | S | Do I have a daughter which is one of *has, have, had*? |
| 5 | haveDaughterShould | S | Do I have a daughter *should*? |
| 6 | haveDaughterWh | S | Do I have a daughter who is one of *where, when, while, who, why*? |
| 7 | haveReportingAncestor | S | Am I a verb/predicate with an ancestor whose lemma is one of *tell, accuse, insist, seem, believe, say, find, conclude, claim, trust, think, suspect, doubt, suppose*? |
| 8 | parentPOS | S | What is my parent's POS tag? |
| 9 | whichAuxIsMyDaughter | S | If I have a daughter which is an auxiliary, what is it? ( = 'nil' if I do not have an auxiliary daughter) |
| 10 | whichModalIsMyDaughter | S | If I have a daughter which is a modal, what is it? ( = 'nil' if I do not have a modal daughter) |
| | | | Features that were not useful |
| 1 | Lemma | L | Word's Lemma |
| 2 | Stem | L | Word stem (Using Porter Stemmer) |
| 3 | Drole | S | Deep role (drole in MICA features) |
| 4 | isRoot | S | Is the word the root of the MICA Parse tree? |
| 5 | parentLemma | S | Parent word's Lemma |
| 6 | parentStem | S | Parent word stem (Using Porter Stemmer) |
| 7 | parentSupertag | S | Parent word's super tag (from Penn Treebank) |
| 8 | Pred | S | Is the word a predicate? (pred in MICA features) |
| 9 | wordSupertag | S | Word's Super Tag (from Penn Treebank) |

Table 9.4: Features used for training CB3-TAGGER.

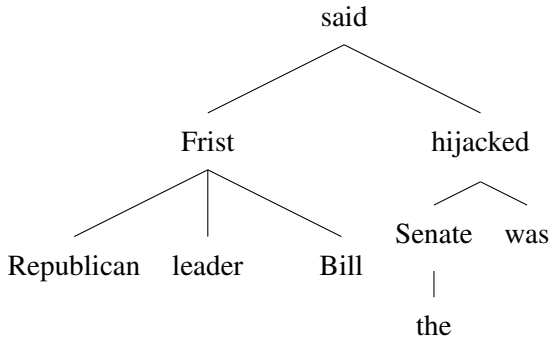



Figure 9.1: Dependency tree for example sentence: *Republican leader Bill Frist said the Senate was hijacked*

sentence, *said* and *hijacked* are the propositional heads that should be tagged. Let's look at *hijacked* in detail. The feature *haveReportingAncestor* of *hijacked* is 'Y' because it is a verb with a parent verb *said*. Similarly, the feature *whichAuxIsMyDaughter* gets the value *was*. Values of all features we use are listed in Table 9.5.

#### 9.3.1.2 Evaluation

For evaluation, we use 4-fold cross validation on the LU corpus. The data was divided into 4 folds of which 3 folds were used to train a model which was tested on the 4th fold. We did this with all four configurations and all the reported results in this paper are micro-averaged results across 4 folds. We report recall, precision, and F-measure on word tokens in our corpus for each of the three tags. It is worth noting that the majority of the words in our data will not be tagged with any of the three classes.

#### 9.3.1.3 Experiments and Results

This section describes different experiments we conducted. We explain the experimental setup as well as results obtained using two learning frameworks — YAMCHA and MALLET. We also explain the pipeline model which uses YAMCHA as the underlying learning framework and the results obtained using it. The best performing feature configuration and corresponding precision, recall and F-measure for each experimental setup is presented in Table 9.7. The best F-measure for each category under various experimental setups is presented in Table 9.8.

| Feature Name | Value |
|---|---|
| isNumeric | N |
| POS | VBN |
| verbType | Reg |
| WhichModalAmI | nil |
| amVBwithDaughterTo | N |
| haveDaughterPerfect | N |
| haveDaughterShould | N |
| haveDaughterWh | N |
| haveReportingAncestor | Y |
| parentPOS | VBD |
| whichAuxIsMyDaughter | *was* |
| whichModalIsMyDaughter | nil |

Table 9.5: Values of representative features for the token *hijacked* in the example sentence.

**YAMCHA Experiments** We categorized our YAMCHA experiments into different experimental conditions as shown in Table 9.6. For each class, we did experiments with different feature sets and (linear) context widths. Here, context width denotes the window of tokens whose features are considered. For example, a context width of 2 means that the feature vector of any given token includes, in addition to its own features, those of 2 tokens before and after it, as well as the tag prediction for 2 tokens before it. For $L_N S_N$, the context width of all features is set to 0. A context width of 0 for a feature means that the feature vector includes that feature of the current token only. For $L_C S_N$, the context width of syntactic features alone was set to 0. When context width was non-zero, we varied it from 1 to 5, and we report the results for the optimal context width. We tuned the SVM parameters, and the best results were obtained using the *One versus All* method for multi-class classification on a quadratic kernel with a $c$ value of $0.5$. All results presented for YAMCHA here use this setting.

The first set of rows in Table 9.7 presents the best performing feature sets and context width configuration for each class. For all experiments with context, the best result was obtained with a

| Class | Description |
|---|---|
| $L_C$ | Lexical features with Context |
| $L_N S_N$ | Lexical with No-context and Syntactic features with No-context |
| $L_C S_N$ | Lexical features with Context and Syntactic features with No-context |
| $L_C S_C$ | Lexical and Syntactic features with Context |

Table 9.6: YAMCHA experiment sets.

context width of 2, except for $L_C$, where a context width of 3 gave the best results. The results show that syntactic features improve the classifier performance considerably. The best model obtained for $L_C$ has an F-measure of 56.9%. In $L_N S_N$ it improves marginally to 59.9%. Adding back context to lexical features improves it to 62.4% in $L_C S_N$ while also adding context to syntactic features further improves this to 64.0%. We observe that the feature *parentPOS* has the most impact on increased context widths, among syntactic features.

The improvement pattern of precision and recall across the classes is also interesting. Syntactic features with no context improve recall by 4.8 percentage points over only lexical features with context, whereas precision improves only by 0.6 points. However, adding back context to lexical features further improves precision by 4.9 points while recall just improves by 0.6 points. Finally, adding context of syntactic features improves both precision and recall moderately. We infer that syntactic features (without context) help identify more annotatable patterns thereby improving recall, whereas linear context helps removing the wrong ones, thereby improving precision.

The per-category F-measure results presented in Table 9.8 are also interesting. The CB F-measure improves 8.1 points and NCB improves 18.9 points from $L_C$ to $L_C S_C$. But, the improvement in NA F-measure is only a marginal 1.3 points between $L_C$ and $L_C S_C$. Furthermore, the F-measure decreases by 3.3 points when syntactic and lexical features with no context are used. On analysis, we found that NAs often occur in syntactic structures like *want to find* or *should go* (deontic *should*), in which the relevant words occur in a small linear window. In contrast, NCBs are often signaled by deeper syntactic structures. For example, in *He said that his visit to the US will mainly focus on the humanitarian issues*, a simplified sentence from our training set, the verb *focus* is an NCB because it is in the scope of the reporting verb *said* (specifically, it is its daughter). This

| Class | Feature Set | Param | P | R | F |
|---|---|---|---|---|---|
| | YAMCHA - Joint Model | | | | |
| $L_C$ | POS, whichModalAmI, verbType, isNumeric | CW=3 | 61.9 | 52.7 | 56.9 |
| $L_N S_N$ | POS, whichModalAmI, parentPOS, haveReportingAncestor, whichModalIsMyDaughter, haveDaughterPerfect, whichAuxIsMyDaughter, amVBwithDaughterTo, haveDaughterWh, haveDaughterShould | CW=0 | 62.5 | 57.5 | 59.9 |
| $L_C S_N$ | POS, whichModalAmI, parentPOS, haveReportingAncestor, whichModalIsMyDaughter, whichAuxIsMyDaughter, haveDaughterShould | CW=2 | 67.4 | 58.1 | 62.4 |
| $L_C S_C$ | POS, whichModalAmI, parentPOS, haveReportingAncestor, whichModalIsMyDaughter, haveDaughterPerfect, whichAuxIsMyDaughter, haveDaughterWh, haveDaughterShould | CW=2 | 68.5 | 60.0 | 64.0 |
| | MALLET - Joint Model | | | | |
| $L$ | POS, whichModalAmI, verbType | GV=1 | 55.1 | 45.0 | 49.6 |
| $LS$ | POS, whichModalAmI, parentPOS, haveReportingAncestor, whichModalIsMyDaughter, haveDaughterPerfect, whichAuxIsMyDaughter, haveDaughterWh, haveDaughterShould | GV=1 | 64.5 | 54.4 | 59.0 |
| | Pipeline Model | | | | |
| $L_C S_C$ | POS, whichModalAmI, parentPOS, haveReportingAncestor, whichModalIsMyDaughter, haveDaughterPerfect, whichAuxIsMyDaughter, haveDaughterWh, haveDaughterShould | CW=2 | 49.8 | 42.9 | 46.1 |

Table 9.7: Overall CB tagging results.

CW = Context Width, GV = Gaussian Variance, P = Precision, R = Recall, F = F-Measure

| Setup | Class | CB | NCB | NA |
|---|---|---|---|---|
| Joint-YAMCHA | $L_C$ | 61.5 | 15.2 | 63.2 |
| Joint-YAMCHA | $L_N S_N$ | 67.0 | 28.3 | 59.9 |
| Joint-YAMCHA | $L_C S_N$ | 67.6 | 33.2 | 64.5 |
| Joint-YAMCHA | $L_C S_C$ | 69.6 | 34.1 | 64.5 |
| Joint-MALLET | $L$ | 53.9 | 7.5 | 54.1 |
| Joint-MALLET | $LS$ | 65.8 | 40.6 | 59.1 |
| Pipeline | $L_C S_C$ | 55.2 | 16.5 | 51.3 |

Table 9.8: CB tagging results: micro-averaged F-measures per category

could not be captured using the context because *said* and *focus* are far apart in the sentence. But a correct parse tree gives *focus* as the daughter of *said*. So, a feature like *haveReportingAncestor* could easily capture this. It is also the case that the root of a dependency parse tree would mostly be a CB. This is captured by the feature *parentPOS* having value 'nil'. This property also cannot be captured by lexical features alone.

NCB performs much worse than the other two categories. NCB is a class which occurs rarely compared to CB and NA in our corpus. Out of the $13,485$ word tokens, only $176$ were NCB; i.e., only $1.3\%$. We assume that this could be a main factor of its poor performance. However, it is worth noting that we obtain significant improvement on the performance of NCB by using syntactic features. In fact, NCB obtained the highest F-measure improvement of $124\%$, compared to $12\%$ for CB and $2\%$ for NA.

**MALLET Experiments** We categorized our MALLET experiments into two classes as shown in Table 9.9. We experimented with varying orders and the best results were obtained for order= "0,1", which makes the CRF similar to Hidden Markov Model. All results reported here use the order= "0,1". We also conducted experiments varying the Gaussian variance parameter from 1.0 to 10.0 using the same experimental setup (i.e. we did not have a distinct tuning corpus) and observed that best results were obtained with a low value of $1$ to $3$, instead of MALLET's default value of $10.0$.

The second set of rows in Table 9.7 presents the best performing feature sets for both classes.

| Class | Description |
|---|---|
| $L$ | Lexical features only |
| $LS$ | Lexical and Syntactic features |

Table 9.9: MALLET experiment sets

These results again show that syntactic features improve the classifier performance considerably. The best model obtained for $L$ class has an F-measure of 49.6%, whereas addition of syntactic features improves this to 59.0%. Both precision and recall are improved by 9.4 percentage points as well.

However, MALLET-CRF's performance was comparatively worse than YAMCHA's SVM. The best model for MALLET ($LS$) obtained an F-measure of 59.0% which is 5.0 percentage points less than that of the best model for YAMCHA ($L_C S_C$). It is interesting to note that MALLET performed well on predicting NCB. The highest NCB F-measure of MALLET — 40.6% — is 6.5 percentage points higher than the highest NCB F-measure for YAMCHA. However, corresponding CB and NA F-measures were 65.8% and 59.1% which are much lower than YAMCHA's performance for these categories. However, MALLET was more time efficient than YAMCHA. On an average, for our corpus size and feature sets, MALLET ran 3 times faster than YAMCHA in a cross validation setup (i.e. training and testing together).

**Pipeline Model** We also did experiments to support our choice of the joint model over the pipeline model. We chose the best performing feature configuration of the $L_C S_C$ class and set up the pipeline model. We trained a sequence classifier using YAMCHA to identify the head tokens, where tokens are tagged as just propositional heads without distinguishing between CB/NA/NCB. The predicted head tokens were then classified using a 3-Way SVM classifier trained on gold data. The head prediction step of the pipeline obtained an F-measure of 83.9% with precision and recall of 86.7% and 81.2%, respectively, across all 4 folds. The 3-way classification step to classify the belief of the identified head obtained an accuracy of 72.7% across all folds. In the pipeline model, false positives and false negatives adds up from step 1 and step 2, whereas only the true positives of step 2 is considered as the true positives overall. In this way, the overall precision was only 49.8% and

recall was 42.9% with an F-measure of 46.1% as shown in Table 9.7. The results for CB/NCB/NA separately are given in Table 9.8. The per-category best F-measure was decreased by 14.4, 17.6 and 13.2 percentage points from the YAMCHA joint model for CB, NCB and NA, respectively. The performance gap is big enough to conclude that our choice of joint model was right.

### 9.3.2 CB3-TAGGERPLUS

In this section, we describe the CB3-TAGGERPLUS system, which is a reimplementation of the CB3-TAGGER described in Section 9.3.1. We reimplemented the system in order to seamlessly integrate the belief tagging process into the power analysis framework that is the core of this thesis. CB3-TAGGERPLUS also uses the LU Corpus with the 3-way belief distinction for training. CB3-TAGGERPLUS obtains better precision, recall and F-measure than CB3-TAGGER. CB3-TAGGERPLUS differs from CB3-TAGGER in terms of four aspects — a different underlying software framework, a different parser to obtain dependency parses and part-of-speech tags, a different machine learning algorithm, and an extended set of features. We describe these differences below.

#### 9.3.2.1 Difference 1: Underlying Software Framework

The original CB3-TAGGER system, which was the first implementation of a belief tagger, was built using the Perl programming language and uses different off-the-shelf NLP products to perform basic NLP tasks such as tokenization and lemmatization. However the UIMA-ClearTk framework that we use in the rest of the thesis provides a scaleable and extensible platform that has been proven very useful for building complex NLP systems. Hence we reimplemented the CB3-TAGGER using the UMIA-ClearTk framework, which we call CB3-TAGGERPLUS. The CB3-TAGGERPLUS system also makes it easier to incorporate belief tagging into the power analysis framework which is the core of this thesis.

#### 9.3.2.2 Difference 2: Dependency Parser

The original CB3-TAGGER system uses the MICA dependency parser (Bangalore et al. 2009) to derive the syntactic features. The motivation for using MICA was that the supertag information that MICA provides will be useful for the task of belief tagging. However, in our experiments (Section 9.3.1.3) we found that MICA's supertag features are not useful for this task. The features

that turned out to be useful were the ones that could also be extracted out of a dependency parser that does not use supertagging. This allowed us to choose the Stanford dependency parser as part of the Stanford CoreNLP package that is integrated into the UIMA framework through ClearTk. Hence, CB3-TAGGERPLUS uses the dependency parses created by the Stanford CoreNLP package in order to extract syntactic features.

#### 9.3.2.3 Difference 3: Machine Learning Algorithm

In our experiments (Section 9.3.1.3), we found the SVM based joint-prediction approach performed better than both the CRF based approach and the pipeline model. Hence, CB3-TAGGERPLUS uses the SVMlight wrapper in ClearTk to perform the training and testing steps. There was no YAMCHA wrapper available in the ClearTk framework.

#### 9.3.2.4 Difference 4: Reimplemented Features with More Features

In CB3-TAGGERPLUS, we use both lexical and syntactic features. We reimplemented the features that were shown useful for CB3-TAGGER. In addition, we also introduced a set of new features, especially some features that capture conditional structures (*if/when*), which helped improve the prediction performance of NA. The complete list of features used by CB3-TAGGERPLUS is listed below.

- **tokenLemma** & **tokenPOS**: Lemma and part-of-speech tag of token.
- **depRel**: Dependency relation of the current token.
- **whichModalAmI**: If the current token is a modal verb, its identity, otherwise 'nil'.
- **parentLemma** & **parentPOS**: Lemma and part-of-speech tag of the parent of the current token in the dependence tree.
- **parentVerbClass**: If the parent token is a verb, the verb classes that verb could be part of as per VerbNet (Schuler 2005). In this step, we did not do any disambiguation; we used all classes a verb may belong to.
- **siblingLemma** & **siblingPOS**: Lemma and part-of-speech tag of each sibling of the current token in the dependency tree.

- **childLemma** & **childPOS**: Lemma and part-of-speech tag of each child of the current token in the dependence tree.
- **ancestorLemma** & **ancestorPOS**: Lemma and part-of-speech tag of each ancestor of the current token in the dependence tree.
- **depRoleOfClosestNounAncestor** & **lemmaOfClosestNounAncestor**: dependency relation and lemma of the closest ancestor whose POS is a noun (i.e., whose POS tag starts with NN)
- **depRoleOfClosestVerbAncestor** & **lemmaOfClosestVerbAncestor**: dependency relation and lemma of the closest ancestor whose POS is a verb (i.e., whose POS tag starts with VB)
- **isTokenUnderConditional**: binary feature denoting whether the token is under the scope of a conditional (i.e., has a child node in the dependency tree that corresponds to tokens *if/when*)
- **isParentUnderConditional** & **parentUnderConditionalPOS**: binary feature denoting whether the token's parent token is under the scope of a conditional (i.e., has a child node in the dependency tree that corresponds to tokens *if/when*), and if so, that parent's part-of-speech tag.

### 9.3.3 CB4-TAGGER

In this section, we describe the CB4-TAGGER system, which is pretty much the same tagger framework as CB3-TAGGERPLUS, but with some additional features, and trained on the DEFT Corpus instead of the LU Corpus. We use CB4-TAGGER to see if the 4-way distinction of belief tags is more useful than the 3-way distinction for the power prediction task. CB4-TAGGER is the system described in (Werner et al. 2015), which we retrained using the DEFT corpus and presented in (Prabhakaran et al. 2015) as the System C. We describe the system briefly in this section. We refer the reader to (Werner et al. 2015) for more details on features and implementation. CB4-TAGGER differs from CB3-TAGGERPLUS in terms of three aspects — the belief tag-set, the training corpus, and the set of features. We describe these differences below.

#### 9.3.3.1 Difference 1: Belief Tag-set

CB4-TAGGER extends the CB3-TAGGERPLUS to perform a 4-way belief distinction — COMMITTEDBELIEF, NONCOMMITTEDBELIEF, REPORTEDBELIEF, NONAPPLICABLE (refer to Sec-

tion 9.2.2, page 162). This results in a 5-way classification (four belief classes and the OTHER class) for the underlying machine learning framework.

#### 9.3.3.2 Difference 2: Corpus

CB4-TAGGER is retrained on the DEFT Corpus (Section 9.2.2), which is much bigger than the LU corpus.

#### 9.3.3.3 Difference 3: Extended feature set

CB4-TAGGER uses a larger feature set than that of CB3-TAGGERPLUS. We summarize the new features below. For more details, refer to (Werner et al. 2015).

- **specialAncestorLemma** & **specialAncestorPOS**: Lemma and part-of-speech tag of each special ancestor of the current token in the dependence tree. A special ancestor is a verb whose lemma is one of the special lemma list. The special lemma list contains the set of reporting verb lemmas used in CB3-TAGGER (*tell, accuse, insist, seem, believe, say, find, conclude, claim, trust, think, suspect, doubt, suppose*) as well as six additional lemmas (*treat, prevent, induce, cause, contain, consist*).
- **bareInfinitive**: binary feature denoting whether the token is a bare infinitive.
- **modalInfinitive**: binary feature denoting whether the token is a modal infinitive.
- **questionwords**: binary features indicating whether the token, parent token, siblings, or children are question words.
- **WordSense** features: the word sense mapping of nouns and verbs.
- **word2vec** features: the class assigned to the token by the word2vec software (Mikolov et al. 2013).[5]
- **hedge** features: binary features denoting whether the token, parent token, siblings, or children are propositional hedges or a relational hedges (Prokofieva and Hirschberg 2014). The hedge

[5] https://code.google.com/p/word2vec/

list contains two types of hedge words/phrases — propositional hedges and relational hedges. Propositional hedges add ambiguity or uncertainty to stated propositions (e.g., *I **think** it is late*), whereas relational hedges show the speaker's lack of commitment to a proposition (e.g., *It is **sort of** late*). The list we use contain 28 propositional hedge phrases and 55 relational hedge phrases.

### 9.3.4 Summary

Here, we summarize the three taggers. Table 9.10 lists the best F-measures obtained by each of the three belief taggers. These numbers are listed to have a sense of how well these taggers perform, overall. They are not directly comparable to each other, since the task, data, and experiment setup are all different for them. Our reimplemented belief tagger (CB3-TAGGERPLUS) posted a much higher F-measure than CB3-TAGGER overall on cross validation, although these numbers cannot be compared directly since the folds have changed. CB4-TAGGER reported an overall F-measure of 69.1 on blind test (Prabhakaran et al. 2015). In the rest of this chapter, we use CB3-TAGGERPLUS and CB4-TAGGER to study how belief tags correlate with power and whether they help in the task of power prediction.

| | F-measures | | | | | |
|---|---|---|---|---|---|---|
| Tagger | CB | NCB | NA | ROB | Overall | Remarks |
| CB3-TAGGER | 69.6 | 34.1 | 64.5 | n/a | 64.0 | LU Corpus cross validation |
| CB3-TAGGERPLUS | 74.4 | 48.7 | 74.0 | n/a | 71.3 | LU Corpus cross validation |
| CB4-TAGGER | 73.1 | 38.0 | 69.9 | 23.0 | 69.1 | DEFT corpus blind test |

Table 9.10: F-measures of different committed belief taggers.

## 9.4 Beliefs, Hedges and Power: A Statistical Analysis

In the rest of this chapter, we investigate whether a participant's usage of different kinds of belief tags is related to the power relations he/she has with the other participants, and how the belief tags can help in the problem of power prediction.

### 9.4.1 Analysis Framework

We use the same problem formulation introduced in Chapter 7. We study the pairs of participants $(p_1, p_2) \in RIPP_t$; i.e., the set of interacting participant pairs in an email thread $t$ who share a superior-subordinate relationship. We are interested in how superiors and subordinates differ in their relative use of different kinds of belief expressions and how the belief tags could help identify who is the superior/subordinate of the pair. For a detailed account of the problem formulation, refer Chapter 7, Section 7.2, page 96 & Section 7.3, page 99.

In order to obtain the belief tags, we had to first obtain the dependency parses of each sentence in the emails. (We did not use any deep syntactic features for making the prediction in the original system presented in Chapter 7 and hence did not have to perform the parsing step.) For this purpose, we use the Stanford CoreNLP package as the underlying NLP stack for the analysis presented in this chapter. The Stanford CoreNLP pipeline failed to process 117 of the email threads in our entire corpus; we excluded them from our analysis. In other words, we use the same data setup (train/dev/test splits) as in Chapter 7, except for the removed 117 threads. The removed threads accounted for only 0.3% of the corpus in terms of number of threads. More over, the number of related interacting participant pairs that got removed as result of this was even smaller. In out training set, this resulted in removing 11 pairs (0.2%) and in the development set, this resulted in the removal of only 1 pair (0.03%). On randomly checking five of these threads, we found that the body of all of those email threads contained non-parse-able text such as dumps of large tables, system logs, or unedited dumps of large legal documents.

### 9.4.2 Features

In this section, we describe the different belief tag based features we compute. For each participant of each pair in our corpus, we aggregate the belief tags (if any) in their messages. As we have already seen, the amount of contribution by a participant is correlated with whether they are superior or subordinate (i.e., subordinates contribute more). Hence, using the raw counts of belief tags will not be useful. Instead, we use the percentage of each belief tags in a participant's messages as the set of features denoted by BELIEFAGGREGATES. There are two versions of this feature set depending on which tagger was used to generate the belief tags — BELIEFAGGREGATES$^{3\text{WAY}}$ and BELIEFAGGREGATES$^{4\text{WAY}}$.

- BELIEFAGGREGATES$^{3WAY}$: belief tag features using CB3-TAGGERPLUS
  - *CBPercent*: percentage of propositional heads tagged as CB
  - *NCBPercent*: percentage of propositional heads tagged as NCB
  - *NAPercent*: percentage of propositional heads tagged as NA
- BELIEFAGGREGATES$^{4WAY}$: belief tag features using CB4-TAGGER
  - *CBPercent*: percentage of propositional heads tagged as CB
  - *NCBPercent*: percentage of propositional heads tagged as NCB
  - *ROBPercent*: percentage of propositional heads tagged as ROB
  - *NAPercent*: percentage of propositional heads tagged as NA

In addition, we also use a set of features called HEDGEAGGREGATES based on a list of hedge words and phrases (Prokofieva and Hirschberg 2014). CB4-TAGGER uses this list to compute a set of hedge features, which improved its performance on NCB by 2.2 percentage points. We use the features to capture each type of hedge occurrences separately – propositional hedges and relational hedges. We use raw counts instead of percentages.

- HEDGEAGGREGATES: hedge features using list of relational and propositional hedges
  - *HPropCount*: number of propositional hedges
  - *HRelCount*: number of relational hedges

### 9.4.3 Analysis

In this section, we describe the findings from the statistical analysis we performed on the set of features described in the previous section. Our hypothesis is that power relations do correlate with the kinds of beliefs people express in their messages. For our analysis, each participant of the pair $(p_1, p_2)$ is a data instance. We perform a two-sample two-tailed Student t-Test to determine if the superiors and subordinates had significantly different mean values for each feature.

| | Feature | Superiors | Subordinates | Significance |
|---|---|---|---|---|
| BELIEFAGGREGATES$^{3WAY}$ | *CBPercent* | 49.8% | 50.1% | - |
| | *NAPercent* | 47.1% | 46.2% | - |
| | *NCBPercent* | 3.1% | 3.7% | $p < 0.001$ |
| BELIEFAGGREGATES$^{4WAY}$ | *CBPercent* | 51.8% | 54.4% | $p < 0.001$ |
| | *NAPercent* | 46.0% | 42.5% | $p < 0.001$ |
| | *NCBPercent* | 1.5% | 2.0% | $p < 0.01$ |
| | *ROBPercent* | 0.7% | 1.2% | $p < 0.001$ |
| HEDGEAGGREGATES | *HPropCount* | 1.1% | 1.1% | - |
| | *HRelCount* | 1.3% | 2.1% | $p < 0.001$ |

Table 9.11: Student t-Test results of CB percentages: *Superiors* vs. *Subordinates*.

**BELIEFAGGREGATES$^{3WAY}$ features:** In our data, it is possible that for some participants there are no propositional heads tagged by the belief tagger. This can be caused by one of three causes: the participant didn't send any messages in the thread (remember that some of the pairs in our set of related interacting participant pairs contain only one-way communication; i.e, one of the persons in the pair may not have sent any messages in the thread), the participant's messages were empty (e.g., forwarding messages), or the participant's messages didn't contain any propositions (e.g., short messages such as "Thanks."). In such cases, the percentage values are undefined and hence we remove such instances from the analysis. This resulted in 7701 instances, out of which 4249 were superiors and 3452 were subordinates. The results obtained on the analysis are presented in Table 9.11 (first set of three rows). We did not find any significant difference in the percentage of propositional heads that are tagged as CB or NA between superiors and subordinates. However, subordinates use significantly more NCBs in their messages. The average percentage of NCBs in subordinates' messages was almost 20% more than that in superiors' messages.

**BELIEFAGGREGATES$^{4WAY}$ features:** As in the case of BELIEFAGGREGATES$^{3WAY}$, we excluded instances for which no proposition heads were identified by the CB4-TAGGER from our analysis. This resulted in 7762 instances, out of which 4285 were superiors and 3477 were subordinates. It

is worth noting that CB4-TAGGER identifies more propositional heads than CB3-TAGGERPLUS in the same text. The results obtained on the analysis of CB4-TAGGER-based belief tag percentages is presented in Table 9.11 (second set; four rows). We found significant differences in the percentage of all belief tags. Subordinates use significantly more CB, NCB and ROB, whereas superiors use significantly more NAs. The relative difference is also worth noting. The relative difference was relatively small for CB and NA. Subordinates use 5.0% more CBs and 7.7% fewer NAs than superiors. For NCB and ROB, the differences were much higher. Subordinates' average percentage of NCBs is 31.4% more than that of superiors, and the average percentage of ROBs is 66.7% more than that of superiors.

**HEDGEAGGREGATES features:** For the hedge feature analysis, we excluded the instances for which the message count was zero; i.e., the participant sent no messages in the thread. This resulted in 8323 instances, out of which 4635 were superiors and 3688 were subordinates. The results obtained on the analysis of hedge feature counts is presented in Table 9.11 (third set; two rows). We found that there was no significant difference between superiors' and subordinates' use of propositional hedges. However, subordinates used 58.5% more relational hedges than superiors.

### 9.4.4 Summary

The results from our statistical analysis of belief tags validates our original hypothesis that power relations do correlate with the kind of beliefs people express in their messages. The finding that superiors use more NAs goes in line with the finding from Chapter 7, Section 7.4 that superiors issue more requests. The difference in NCB and ROB is striking. Not only do the subordinates express more non committed beliefs, they also report others' beliefs more often than superiors. The significant difference we find in the usage of hedges is also interesting. It shows the importance of distinguishing between propositional and relational hedges; subordinates use significantly more relational hedges than superiors, but they do not differ in their relative use of propositional hedges. It is also important to point out that hedges are one of the ways NCBs are expressed and our findings in both these features confirm that subordinates use more non-committedness in their language.

## 9.5 Utility of Belief Tags for Power Prediction

Having established that the committed belief labels significantly correlate with whether or not the author has social power or not, our next step is to explore whether we can use the belief tags to improve the performance of the task of automatic power prediction. In this chapter we investigate different ways of incorporating the belief tag information into the power prediction system presented in Chapter 7.

### 9.5.1 Implementation

As discussed in Section 9.4.1, the analysis framework we use in this chapter differs from the one in Chapter 7 (Section 7.5) in terms of the underlying NLP stack we use. In this chapter, we use the Stanford CoreNLP pipeline which performs tokenization, part-of-speech tagging, lemmatization, as well as dependency parsing, whereas in Chapter 7 (Section 7.5) we use CklearTk's default tokenizer, part-of-speech tagger and lemmatizer (no parsing). This change affects the results presented in this chapter in two ways. First, we had to exclude 117 threads from our corpus since Stanford CoreNLP failed on processing them, which resulted in a slightly different train and test sets, and hence the results are not directly comparable with what was presented in Chapter 7. Second, the source of part-of-speech tags and word lemmas are different in this chapter, which might affect the performance of the dialog act tagger and overt display of power tagger. Except for the underlying NLP stack, we use the same experimental framework as in Chapter 7 for the results presented in this section.

### 9.5.2 Baseline Results

We use the two best performing feature sets from Chapter 7 as baseline systems here. The performance obtained using the baseline systems are shown in Table 9.12 (first row in each set of rows). The first baseline using the combination of THREAD STRUCTURE, DIALOG ACTS, OVERT DISPLAY OF POWER and LEXICAL obtain an accuracy of 70.3%. The second baseline system using THREAD STRUCTURE, POSITIONAL and LEX obtain an accuracy of 70.9%.

The difference in their performance is notable. We attribute the lower performance of the first baseline to the issue of DIALOG ACTS and OVERT DISPLAY OF POWER features being affected by the change in the source of part-of-speech tags and word lemmas that they use to make the

| | System | Accuracy | |
|---|---|---|---|
| | | 3Way | 4Way |
| Baseline 1 | THR + DA + ODP + LEX | 70.25 | |
| | THR + DA + ODP + LEX + HDG | 70.25 | |
| | THR + DA + ODP + LEX + CBA | 70.90 | 70.67 |
| Baseline 2 | THR + PST + LEX | 70.87 | |
| | THR + PST + LEX + HDG | 70.87 | |
| | THR + PST + LEX + CBA | 70.70 | 70.70 |

Table 9.12: Power prediction results using CB counts and percentages as features.

predictions. One way to fix this issue is to retrain the dialog act tagger and overt display of power tagger using the Stanford CoreNLP generated part-of-speech tags and word lemmas. We have not done this step for the experiments presented in this section.

### 9.5.3 Incorporating Counts and Percentages of Belief Tags

One straightforward way of using the belief labels in the machine learning experiments is by using their percentages as features. As we saw in the Section 9.4, superiors and subordinates differ significantly in what percentage of belief tags they use in their messages, and how often they use hedges. We added the BELIEFAGGREGATES (CBA) and HEDGEAGGREGATES (HDG) for each participant of the pair to the features used by the machine learning system. The results obtained in these experiments are shown in Table 9.12. We report the results obtained using both BELIEFAGGREGATES$^{3WAY}$ and BELIEFAGGREGATES$^{4WAY}$ versions.

For both baselines, adding HEDGEAGGREGATES features did not have any effect. This is not surprising, since the features as part of the HEDGEAGGREGATES are already captured by LEXICAL features. Adding BELIEFAGGREGATES$^{3WAY}$ and BELIEFAGGREGATES$^{4WAY}$ improved the accuracy from 70.3% of the first baseline to 70.9% and 70.7% respectively. However, adding either of the BELIEFAGGREGATES decreased the performance of the second baseline. This result is surprising, especially since the BELIEFAGGREGATES differ significantly between superiors and subordinates

(Section 9.4). This is possibly because BELIEFAGGREGATES features add complementary information to dialog act based features (and hence the improvement in the first baseline), whereas the combination of positional and belief features cause confusion to the machine learning algorithm. However, the decrease in performance is only marginal (0.17 percentage points).

### 9.5.4 Incorporating Belief Tags into Lexical Features

In this section, we investigate a more sophisticated way of incorporating the belief tags into the power prediction framework. As we have seen in previous chapters, LEXICAL features are very useful for the task of power prediction. However, it is often hard to capture contextual information of words and phrases using ngram features. We hypothesize that incorporating belief tags into the ngrams will enrich the representation and will help disambiguate different usages of same words/phrases. For example, let us consider two example sentences: *I need the report by tomorrow* vs. *If I need the report, I will let you know*. The former is likely coming from a person who has power, whereas the latter does not give any such indication. Applying the belief tagger to these two sentences will result in *I need(CB) the report ...* and *If I need(NA) the report ...* Capturing the difference between *need(CB)* vs. *need(NA)* will help the machine learning system to make the distinction between these two usages and in turn improve the power prediction performance.

We use two ways to incorporate the belief tags into the ngram features. In building the ngram features, whenever we encounter a token that is assigned a belief tag, then we have two options — *Append* the belief tag to the corresponding lemma or part-of-speech tag in the ngram or *Replace* the corresponding lemma or part-of-speech tag in the ngram with the belief tag. In the example discussed above, we have shown the *Append* method. These two approaches are applied to each type of ngram features. We list the different versions of each type of ngram features below.

- *LN*: the original word lemma ngram; e.g., *i_need_the*
- $LN^{CBAppend}$: word lemma ngram with appended belief tags; e.g., *i_need(CB)_the*
- $LN^{CBReplace}$: word lemma ngram with replaced belief tags; e.g., *i_(CB)_the*
- *PN*: the original part-of-speech ngram; e.g., *PRP_VB_DT*
- $PN^{CBAppend}$: part-of-speech ngram with appended belief tags; e.g., *PRP_VB(CB)_DT*

| Feature Configuration | Accuracy | |
|---|---|---|
| | 3Way | 4Way |
| THR + DA + ODP + LEX (*LN +PN +MN*) | 70.25 | |
| THR + DA + ODP + LEX ($LN^{CBAppend}$+*PN +MN*) | 70.65 | 70.70 |
| THR + DA + ODP + LEX ($LN^{CBReplace}$+*PN +MN*) | 70.48 | 70.31 |
| THR + DA + ODP + LEX (*LN +*$PN^{CBAppend}$+*MN*) | 69.25 | 69.61 |
| THR + DA + ODP + LEX (*LN +*$PN^{CBReplace}$+*MN*) | 69.30 | 70.76 |
| THR + DA + ODP + LEX (*LN +PN +*$MN^{CBAppend}$) | 69.42 | 69.97 |
| THR + DA + ODP + LEX (*LN +PN +*$MN^{CBReplace}$) | 69.36 | 70.39 |
| BEST = THR + DA + ODP + LEX ($LN^{CBAppend}$+*PN +MN*) | **70.65** | 70.70 |
| BEST = THR + DA + ODP + LEX (*LN +PN +MN +*$MN^{CBAppend}$+$MN^{CBReplace}$) | 69.42 | **71.29** |
| THR + PST + LEX (*LN +PN +MN*) | 70.87 | |
| THR + PST + LEX ($LN^{CBAppend}$+*PN +MN*) | 71.01 | 70.98 |
| THR + PST + LEX ($LN^{CBReplace}$+*PN +MN*) | 70.62 | 70.84 |
| THR + PST + LEX (*LN +*$PN^{CBAppend}$+*MN*) | 69.84 | 70.11 |
| THR + PST + LEX (*LN +*$PN^{CBReplace}$+*MN*) | 70.20 | 71.04 |
| THR + PST + LEX (*LN +PN +*$MN^{CBAppend}$) | 69.75 | 70.53 |
| THR + PST + LEX (*LN +PN +*$MN^{CBReplace}$) | 69.95 | 70.93 |
| BEST = THR + PST + LEX ($LN^{CBAppend}$+*PN +MN*) | **71.01** | 70.98 |
| BEST = THR + PST + LEX (*LN +* $PN^{CBReplace}$+ *MN +* $MN^{CBAppend}$) | 69.50 | **71.71** |

Table 9.13: Power prediction results after incorporating CB tags into lexical features.

The two sections of the table shows results using two different baselines, one for each winning feature set in the experiments presented in Chapter 7 (Section 7.5))

THR: THREAD STRUCTURE, DA: DIALOG ACTS, ODP: OVERT DISPLAY OF POWER, PST: POSITIONAL, LEX: LEXICAL

- $PN^{CBReplace}$: part-of-speech ngram with replaced belief tags; e.g., *PRP_(CB)_DT*
- *MN*: the original mixed ngram; e.g., *i_VB_the*
- $MN^{CBAppend}$: mixed ngram with appended belief tags; e.g., *i_VB(CB)_the*
- $MN^{CBReplace}$: mixed ngram with replaced belief tags; e.g., *i_(CB)_the*

In Table 9.13, we show the results obtained by incorporating the belief tags in this manner to the LEXICAL features of the original baseline feature sets. The first row in both sets of results indicate the baseline results and the following rows show the impact of incorporating belief tags as *Append* or *Replace* to each type of ngram. For both baselines, $LN^{CBAppend}$improved the results over *LN*. In other words, the distinctions such as the one we discussed earlier in the example (*i_need(CB)* vs. *i_need(NA)*) are helpful for the power prediction system. For the other two types of ngrams (part-of-speech and mixed), the improvement pattern was not clear. For both types of ngrams, the *Replace* version of incorporating the 4-way belief tags reported improvements of varying degrees across the board.

We then experimented with different combinations of the types of ngrams. For each type of ngram, we try six different settings. For example, in the case of word ngrams, we have *LN*, $LN^{CBAppend}$, $LN^{CBReplace}$, *LN* + $LN^{CBAppend}$(using both the regular and belief *Append* versions), *LN* + $LN^{CBReplace}$, and *LN* + $LN^{CBAppend}$+ $LN^{CBReplace}$. Similarly we have six different settings for part-of-speech and mixed ngrams. Altogether we have $6^3 = 216$ different settings for LEXICAL for each belief tag source (3-way vs. 4-way). We performed these $432$ experiments on both baselines. The best performing configuration obtained for both 3-way and 4-way belief tags are presented in the last two rows of each section of rows.

For the first baseline, the best accuracy obtained using 3-way belief tags was 70.7%, whereas using 4-way belief tags, we obtain an accuracy of 71.3%. This improvement is more than a 1 percentage point improvement of accuracy. For the second baseline, the best accuracy obtained using 3-way belief tags was 71.0%. The 4-way belief tags obtained an overall best accuracy of 71.7%, a significant improvement over not using any belief information (70.9%). We use the approximate randomization test (Yeh 2000) for testing statistical significance of the improvement. An interesting observation is that, for 3-way belief tags, the *Append* approach applied to the word lemma ngrams

works the best. However, in the 4-way distinction, the *Append* applied to either word lemma or mixed ngrams is helpful.

## 9.6 Conclusion

In this chapter, we studied how the levels of beliefs expressed by participants of an interaction correlates with the power relations they have with other participants. The chapter has three major contributions — a system to automatically tag levels of belief in running text, correlation analysis of how the proportion of belief tags correlate with power, and different ways of incorporating the belief tag information into an automatic system that predicts power relations.

We built our automatic belief tagger based on existing committed belief annotations that labels heads of propositions with one of the three belief tags — committed belief, non-committed belief and non-belief. We experimented using SVM-based and CRF-based machine learning techniques, as well as lexical and syntactic features. We obtained the best performance using the SVM-based approach using both syntactic and lexical features.

We then applied the tagger we built to the Enron email corpus in order to study how belief tags correlates with power. In this part of the study, we also used another belief tagger that makes a 4-way distinction of — committed belief, non-committed belief, reported belief, and non-belief. In our analysis, we found that superiors and subordinates use significantly different proportions of different types of beliefs in their messages. In particular, subordinates use significantly more non-committed beliefs than superiors.

Finally, we investigated different ways to incorporate the belief tag information to the machine learning system that automatically detects the direction of power between pairs of participants in an interaction. We found that while the relative proportions are helpful to improve the prediction performance, a better way to incorporate this information into the machine learning framework is to include this information in the lexical features, either by appending the belief tags to the propositional heads or by replacing the heads with the belief tags.

# Chapter 10

# How Types of Power Manifest Differently

Power is not a singular concept; it stems from different bases (French and Raven 1959). However, most computational approaches towards analyzing power in interactions rely only on a single notion of power, often based on static power structures. In Chapters 7 and 8, we also used a single notion of power; one that is based on organizational hierarchy. Although recently there have been some studies that looked into dynamic notions of power such as influence (Biran et al. 2012), not much work has been done to understand how different types of power differ in the ways they are manifested in dialog

In this chapter, we introduce a new typology of power in a workplace setting — hierarchical power, situational power, influence and power over communication. The main contribution of this chapter is to describe this typology and show that these four types of power we introduce are in fact different from one another and that they affect the dialog participant's behavior in different but predictable ways. We investigate how these four types of power differ in the ways they are manifested along the different aspects of dialog behavior we analyze in this thesis. We also present a system to automatically detect the dialog participants with one of these types of power.

## 10.1 A New Power Typology for Organizational Email

In social sciences, different typologies of power have been proposed. The five bases of power proposed by French and Raven (1959) (Coercive, Reward, Legitimate, Referent, and Expert) and its extensions are widely used to study power. More recently, Wartenberg (1990) makes the distinction between power-over and power-to in the context of interactions. We find these definitions and typologies helpful as general background, but not specific enough for a data-oriented study of how they are expressed in written dialogs such as emails. In this chapter, we propose a new typology of four types of power in the context of organizational email. We describe the core distinguishing factors of each type of power below.

- Hierarchical Power (HP): A is said to have hierarchical power over B if A is above B in a static power structure/hierarchy.
- Situational Power (SP): A is said to have situational power over B if A has the power or authority to direct and/or approve B's actions in the current situation or while a particular task is being performed.
- Power over Communication (PC): A is said to have power over communication if A actively attempts to achieve the intended goals of the communication.
- Influence (INFL): A is defined to have influence over B if A has credibility and/or persists in attempting to convince B, even if some disagreement occurs.

To understand these types of power better, let us consider a hypothetical situation (Table 10.1) — a meeting held in an organizational setting regarding enforcing a Human Resource (HR) policy regarding a new technology X. The HR head Hannah who is in charge of enforcing the policy and the two project managers Matt and Mary are part of the meeting. An expert in technology X, Evan, who's a subordinate of Mary, is also present at the meeting. Matt volunteered to be the chair of the meeting. Mary has hierarchical power over Evan since she is above Evan in the organizational hierarchy. Hannah has situational power over both Matt and Mary since she is in charge of the task of enforcing the policy. Evan, being the expert, has the power of influence over Hannah. Matt, being the chair of the meeting, has the power over conversation; i.e., the power to decide how the conversation is structured and whether and when its objectives are met.

| Meeting Topic: | Power Relations Scenario |
|---|---|
| Technology X enforcement by HR | HierarchicalPower(Mary,Evan) |
| Attendees: | SituationalPower(Hannah,Mary) |
| Hannah - HR head | SituationalPower(Hannah,Matt) |
| Matt - Manager (Project A) | Influence(Evan,Hannah) |
| Mary - Manager (Project B) | PowerOverCommunication(Matt) |
| Evan - Team Member (Project B); Expert (Technology X) | |

Table 10.1: Hypothetical meeting to illustrate different types of power.

Hierarchical power is the most commonly used notion of power in the domain of organizational interactions. Situational power is a closely related notion; i.e., both stem from a role/position. However, the role is dynamic in the case of situational power, whereas it is static in the case of hierarchical power. Situational power is also an active form of power (i.e., being in charge in the current situation), whereas hierarchical power need not be active. Power over communication, like situational power, is an active form of power. However, the fundamental difference between them is that situational power stems from being in charge of a task, whereas power over communication stems from being in charge of the conversation. We adopt the notion of influence from the IARPA Socio-Cultural Content in Language (SCIL) program. Many of the researchers participating in the SCIL program contributed to the scope and refinement of the definition of influence. Influence is also an active form of power. We define each type of power more formally and describe them in more detail in Section 10.3 where we discuss the annotation process.

## 10.2 Relation to Other Power Typologies in Literature

In social sciences, different typologies of power have been proposed. Wartenberg (1990) makes the distinction between power-over and power-to in the context of interactions. Power-over refers to relationships between interactants set by external power structures, while power-to refers to the ability an interactant possesses within the interaction, even if it is temporary. Our notions of hierarchical power and influence are special cases of power-over. Hierarchical power is determined by organizational hierarchy, while influence is determined by knowledge, expertise etc. Similarly, our

notions of situational power and power over communication are special cases of power-to. Situational power applies to the situation or task at hand, while power over communication applies to the interaction itself. French and Raven (1959) proposed five bases of power: Coercive, Reward, Positional, Referent, and Expert. They are widely used to study power in sociology. We consider hierarchical power, situational power and power over communication to be positional in nature; although the former two can also have bases in coercion and rewards. The bases of influence are mainly referent and expert power.

Within the dialog community, researchers have studied notions of control and initiative in dialogs (e.g. (Walker and Whittaker 1990, Jordan and Di Eugenio 1997)). Walker and Whittaker (1990) define "control of communication" in terms of whether the discourse participants are providing new, unsolicited information. They use utterance level rules to determine which discourse participant (whether the speaker or the hearer) is in control, and extend it to segments of discourse. Their notion of control differs from our notion of power over communication. They model control locally over discourse segments. What we are interested in (and what our annotations capture) is the possession of controlling power by one (or more) participant(s) across the entire dialog, i.e. which participant controls the communication in a dialog thread and tries to achieve its intended goals. Despite this difference in definition, we show in Section 10.8 that our notion of power over communication correlates with Walker and Whittaker (1990)'s notion of control over discourse segments.

## 10.3 Power Annotations

In this section, we describe the procedure followed to obtain manual annotations for power relations. We start by describing the corpus we use. We then describe the annotation instructions in detail as well as give example annotations. We obtain different types of annotations apart from the power annotations we use in this chapter (for example, we obtain annotation for intention).

### 10.3.1 Corpus

We use the ENRON-SMALL corpus presented in Chapter 3 Section 3.1.4 for our study in this chapter. The corpus contains manual dialog act annotations by Hu et al. (2009), which enable us to perform reliable analysis of how different types of power affects dialog behavior. The corpus contains 122

email threads with a total of 360 messages and 20,740 word tokens. There are about 8.5 participants per thread. There are 221 active participants (participants of a thread who has sent at least one email message in the thread) in the corpus.

### 10.3.2 Annotation for Intention

We ask the annotator to first identify the intention of the email thread. This step forces the annotator to consider the email thread as a whole and think about the thread from a high level. The identified intention is also used as an artifact in determining the person with the power over communication. The exact instruction given to the annotator is as follows:

> *What, in general, is the purpose, or content type, of this discussion thread?*

The annotator was asked to choose an intention from one of the following eight options (verbatim from the instructions). We also asked the annotator to enter a very short description of the topic of the thread.

- Knowledge-Acquisition: The thread purpose is mainly to convey or exchange information.
- Argumentation: The thread purpose is mainly to argue or explore the pros and cons of a position or claim.
- External-event-planning: Planning events that will take place outside of the email exchange, such as a meeting, or performance of a task.
- Collaboration-on-information-product: Collaboration on a document or information. Mark this if the work will be done inside the email communication channel.
- Problem-solving: The main purpose is solving a problem. The solution to the problem need not happen in the email thread; it could also be the discussion of possible solutions. An example for Problem-solving is a thread discussing how to fix a wrong price on a deal.
- Social: The main purpose of the thread is simply being social.
- Approval: The main purpose of the thread is approval of a task or document

- Other or Unsure: None of the above applies, or it is too complicated to decide (explain briefly why).

A thread can be annotated with more than one intention, if those intentions are simultaneously present. For example, in a thread discussing putting together the company's 10 year anniversary (*event planning*), if the communication in the thread is primarily to argue over which of the two suggested venues for the event is better, then the thread's intention is *event planning/argumentation*. In cases where there are different intentions at different parts of the thread, we asked the annotator to annotate the dominating intention. For example, if there is a side discussion by two participants about going out for drinks after work (*social*), we asked the annotator to annotate the thread's intention as *event planning*, since that is the main objective of the thread. An example for intention annotation from the annotated corpus is given below.

```
Example for Intention Annotation (Ref thread - A)
Intention:  Collaboration-on-information-product
Topic:  updating trading rotations document
Intention:  Knowledge-Acquisition
Topic:  discussion of interview plan
```

### 10.3.3 Power Narrative Annotation

After the intention annotation, we asked the annotator to provide a power narrative: a brief summary of the power relation scenario. By power relation scenario, we mean "a consistent assignment of power relations between participants of the thread". For a given thread there could be different such possible power scenarios one could think of. We asked the annotator to select the one that appears most probable to them and that the narrative should specify the scenario they have chosen. Their remaining power annotations should be in sync with the scenario they have chosen. We asked the annotator to also specify the reasoning of specific power relations within a scenario, in cases where that is not obvious. The purpose of the power narrative annotation is to force the annotator to think about the thread-level picture when making the power annotations, and thereby preventing any contradicting annotations within a thread. This annotation is obtained in free form text. An example for power narrative annotation from the annotated corpus is given below.

Example for Narrative (Ref thread - A)
Narrative: Karen has situational power over Lloyd, John, Kimberly and others because she gives specific directives to everyone. However, it seems like John have some sort of power/influence over Karen, based on the way he asks about outsiders and Karen giving detailed report of who those outsiders are. Karen seems to be someone from HR.

### 10.3.4 Situational Power

Based on the power narrative that was formed after looking at the entire thread, we asked the annotator to list the instances of situational power in the thread. The exact instruction given to the annotator is as below:

> List (Person_1,Person_2,Confidence,Reason) tuples such that based on the communication in the current thread, person 1 has power (authority to direct / approve other people's actions) in the current situation or while a particular task is being performed. Situational power is not necessarily aligned with organizational hierarchy: Person_1 with situational power may or may not be above Person_2 in the organizational hierarchy (or there may be no organizational hierarchy at all). The Confidence states whether you are Certain or Almost Certain about this power relation. The Reason is supporting evidence, ideally referring to specific parts of the thread. Sometimes it is hard to give a specific reason, or to point to specific text passages; in this case, just be as specific as possible.

An example for situational power annotation from the annotated corpus is given below.

Example for Situational Power (Ref thread - A)
Situational Power:
Person_1: Karen Buckley
Person_2: John Lavorato / Lloyd Will
Confidence: Certain
Reason: In messages 1 and 3 from Karen (specifically M1.2, M1.5, M3.3), she gives specific directives to the others; in Message 2 from Lloyd Will to Karen, he complies with her request.

Example for Situational Power (Ref thread - A)
Situational Power:
Person_1: Shelley Corman
Person_2: Rick Dietz
Confidence: Certain
Reason: Rick is informing Shelley about his absence and ensuring work is not affected while he is away. His messages have an apologetic tone, while Shelley's message are short and precise.

### 10.3.5 Power over Communication

The third type of power we asked the annotator to identify was the power over communication. The exact instruction given to the annotator is as below:

> Identify the people who actively attempt to achieve the intended goals (Intention annotation of the thread) by controlling the dialog. This would be people who ask questions, request others to take action, etc. and not people who simply respond to questions or perform actions when directed to do so. A prototypical example is the chair of a meeting or the anchor of a debate. A person with power over communication should be present (as the recipient/CC) in the part of thread where the corresponding intention is pursued. For each such person, provide a `Reason`, ideally referring to specific parts of the thread that show their active participation in achieving the intention. Power over communication is a power one has over thread and hence, there is no `Person_2`.

Example for Power over Communication (Ref thread – A)
Annotation for Power over Communication:
Person_1: Karen Buckley
Confidence: Certain
Reason: Karen initiates the thread and sends out directives in Message 1 and Message 3 to achieve the intention of updating the trading rotations document. She sends out information on the interview plan in Message 5.

Someone initiating the conversation does not automatically give them the power over communication. Initiating the thread implies that the person has topic control (in other words, he or she introduced the topic of the thread), but if the person doesn't follow up on ensuring topic control, he is not in control. Also, someone else could take over and carry forward the conversation, in which case that other person should be judged to be having the power over communication. Example below.

Example for Power over Communication (Ref thread – A)
Annotation for Power over Communication:
Person_1: Kimberly Watson
Confidence: Certain
Reason: Kimberly disburses the details sent by Kevin to different people and clarifies their queries, sometimes checking back with Kevin.

Power over communication doesn't have to align with situational power or hierarchical power. In below example, even though Sara seem to have situational/hierarchical power over Kim, the conversation is controlled by Kim.

In a social email (one that doesn't necessarily have a task to do), if there is one participant who persists in keeping the conversation continuing (by asking questions, etc.), he or she should be judged as the person with power over communication.

```
Example for Power over Communication (Ref thread - A)
Annotation for Power over Communication:
Person_1: Kim S Ward
Confidence: Certain
Reason: Kim initiates the thread, tells Sara Shackleton about the
Glendale city attorney's questions, and asks the city attorney, Steve
Lins, for the bond counsel's information.
```

```
Example for Power over Communication (Ref thread - A)
Annotation for Control:
Person_1: Michelle Nelson
Confidence: Certain
Reason: Michelle initiates the conversation, carries forward the
conversation by asking things or saying things that urges Mike to
respond (for example, name calling). And at the end, she chose to
stop the conversation saying she's bored.
```

### 10.3.6 Influence

If there were any influence relations the annotator identified in the thread, we asked the annotator to list them. The exact instruction given to the annotator is as below:

> Identify (Person_1,Person_2,Confidence,Reason) tuples where both of the below conditions hold.
>
> - Person_1 is afforded credibility by Person_2
> - Person_1 is able to change or affect Person_2's ideas or opinions in a positive way.
> - Person_1 does not expect Person_2 to accept his ideas or opinions readily.

It could be a scenario where Person_1 emits ideas or opinions and Person_2 picks up on it and support it; sometimes readily and sometimes after Person_1 convinces Person_2 (in case of some low level disagreements). Another scenario is where Person_2 asks Person_1 for advice or opinions and that advice or those opinions are adopted or supported by Person_2. The affordance

of credibility could be explicit (for example, by asking for an opinion) or implicit (for example, by adopting Person_1's ideas or language). The third point — *no expectations* — is important to distinguish influence from situational power. The person with situational power does expect the other person to take his advice and opinions, although he might be willing to negotiate. However, the person with influence would not have this expectation that his advice and opinions are to be readily accepted. He might have an intention to make them accepted (the case where he tries to convince others in case of disagreements), but that still qualifies as Influence. The Confidence states whether you are Certain or Almost Certain about this power relation. The Reason is supporting evidence, ideally referring to specific parts of the thread where Person_1 seem to have influence over Person_2. An example for situational power annotation from the annotated corpus is given below.

```
Example for Influence (Ref thread - A)
Influence Annotation:
Person_1:  Mark Taylor
Person_2:  Sara Shackleton
Confidence:  Certain
Reason:  In message 1, Sara asks Mark for his advice on using a law
firm for a second time, and after he expresses his view in message 3,
she concurs in message 4, to Martin Rosell.
```

### 10.3.7 Annotation Guidelines

The following guidelines were given to the annotator to ensure the quality fo annotations. These guidelines were formed as a result of the first few rounds of annotations as well as annotator clarification questions.

- Consider only people present in at least one from, to, or cc field (In other words, disregard relations between people that are merely mentioned in the messages, but are not one of the participants).
- Annotations should be made based solely on the communication within the thread, disregarding annotator's world knowledge about any participant or knowledge about participant

relations from previously annotated threads.

- For each message, you should find its parent message (the message to which it is the reply of), to understand the dialog structure. Since the messages are listed in the chronological order in the threads in our data, a message may not always be a reply to the one listed just before it.
- Power relations should not be judged purely based on speculations such as "the person used his personal email address, he must not be the boss" etc.
- Power annotations should be made independently of each other. For example, one person can have situational power over another, without being the one in control, and vice versa.
- Situational power and Influence are annotations on pairs of participants, while Power over Communication is an annotation on a participant over the conversation in the thread.

### 10.3.8 Known Issues in the Data

The email threads on which the annotations were obtained are created automatically and hence there were some issues in a few cases. We informed the annotators of such issues and suggested what to be done if you encounter such a case.

- There are some threads which have messages with no from or to tags. In such cases, even if you can guess who the sender is, from the conversation, please consider him or her as not a participant (unless he is present in the `from`/`to`/`cc` in another message within the same thread).
- There are some threads which have messages arranged in wrong order chronologically. For example, message A comes in the thread before message B, whereas message B was sent 2 hours earlier to message A. You should look at the time tag of every message to make sure messages are in the correct order. If you find cases where messages are arranged in the wrong order, please point out such threads to us via email and do the annotation as if the messages were in the correct order.
- For some messages, person name field will be empty, use the ID field in the `to`/`from`/`cc` fileds and mention the person as `EmptyName(id:"9999")`

### 10.3.9 Other Annotations

As part of the annotation effort, we also obtained annotations for hierarchical power, which we do not use in this thesis. The purpose for those annotations were to capture the instances where hierarchical power relation is apparent from the interaction (e.g., from seeing a subordinate asking for a leave approval). However, we use a more reliable source of hierarchical power relations (Agarwal et al. 2012) for the annotations presented in this corpus. In addition, we also asked the annotator to mark text segments that are instances of attempts to exercise each type of power. However the annotations were sparse and we do not use those annotations in this thesis.

### 10.3.10 Annotations Statistics

Table 10.2 presents the counts and percentages of active participants with each type of power in the corpus. We have very few instances of hierarchical power and influence, whereas we have around 37% active participants judged to be having situational power and around 58% to be judged to have power over communication.

| Type of power | Count | Percentage |
|---|---|---|
| Hierarchical Power (HP) | 18 | 8.1 |
| Situational Power (SP) | 81 | 36.7 |
| Power over Communication (PC) | 127 | 57.5 |
| Influence (INFL) | 11 | 5.0 |

Table 10.2: Power annotation statistics.

## 10.4 Subjectivity of Power Annotations

The power annotations in the corpus are performed by a single annotator and capture her perception of the overall power structure among the participants of the interaction. Although we take this as the gold annotations for our analysis and experiments, it is possible that the judgment of power, observed by a third party (our annotator) is subjective. In this section, we investigate how subjective the perception of power is, by doing an inter-annotator agreement study. We performed an inde-

pendent study of annotator perceptions of power on a subset of 47 threads from the corpus. We trained two annotators — AnnA and AnnB — using the same annotation manual and compared the annotations they produced for power relations on the selected threads. Both AnnA and AnnB were undergraduates, one from the Arts Department and the other from the Engineering Department.

The cognitive process behind labeling a participant to have any type of power is not a binary decision the annotator makes for each participant. Annotators read the entire thread before performing the annotations. As mentioned in Section 10.3.3, page 195, they are also asked to provide, in free-form English, a short "power narrative" which describes their perception of the overall power structure among the discourse participants of that thread. Annotators build a fairly consistent mental image of a power narrative — an outline of the power structure between the participants — based on various indicators from across the thread. Their individual power annotations are based on this power narrative. Evaluating agreement on such a task is not trivial. For the purposes of this study, we port this task into a binary decision task of identifying whether participant X has power of type P or not.

| Type of power | Round 1 | Round 2 |
|---|---|---|
| Situational Power (SP) | 0.47 | 0.47 |
| Power over Communication (PC) | 0.27 | 0.76 |
| Influence (INFL) | 0.50 | 0.79 |

Table 10.3: Inter annotator agreement ($\kappa$) of power annotations.

There were 289 participants in the selected 47 threads. The Cohen's $\kappa$ (Cohen 1960) values obtained for each type of power is shown in Table 10.3 under Round 1. The $\kappa$ values obtained in round 1 were moderate to substantial (Landis and Koch 1977). Upon further analysis, we found that this was caused by a misunderstanding on the part of AnnA in understanding the annotation instructions. We performed another round of training and inter annotator study. For this round, AnnB was not available, and we hired another annotator AnnC. The $\kappa$ values obtained between AnnA and AnnC on another set of 10 threads is presented in Table 10.3 under Round 2. The $\kappa$ values obtained in both round 1 and round 2 are in the range of those previously reported for similar tasks (e.g., 0.18 for managerial influence and 0.52 for establishing solidarity (Bracewell et al. 2012);

0.72 for influence (Biran et al. 2012)). The agreement in round 2 improved considerably for both Power over Communication and Influence after the second round of training, however the $\kappa$ for Situational Power stayed 0.47, which is considered only moderate agreement.

The fact that we don't obtain a higher agreement, especially for situational power, could be due to many reasons. Firstly, in porting the task to a binary labeling task, we are unnecessarily penalizing the annotators by introducing instances to represent judgments that the annotator never actually made. For example, if an email invite to a party was sent to 50 recipients, the annotator will not have considered each single recipient individually and made a choice about him or her. However, these 50 recipients will be added as data points in our $\kappa$ calculation, thereby increasing the expected agreement and decreasing the $\kappa$ value. Another reason could be just that the task by itself is subjective. The indicators that are noticed by each annotator may under-specify how they can be interpreted in the power narrative (and subsequently the power annotations). The annotator's choices will then vary depending on the annotator's familiarity with corporate culture, or with other individual characteristic of the annotators.

We investigated the annotations further to confirm this. We found that there were many instances where different valid power narratives could be built based on the same email thread. For example, consider the example thread in Table 10.4. The message from Bill (first message) could be interpreted in isolation as a request from a peer or even a subordinate. However, if you take into consideration that Barry delegated the task to Stephanie upon receiving the message from Bill, the first message could be considered as Bill assigning a task to Barry. Either judgment is valid depending on the power narrative that one builds around the interaction within the thread. The original annotations adopted the latter narrative whereas both AnnA and AnnB adopted the former. In our investigation of the cases where AnnA and AnnB disagreed, we found many cases where both scenarios (person $X$ having power and not having power) are plausible based on the annotators' power narrative.

The original annotations that were in the corpus are the perception of one particular annotator. The moderate agreement obtained in our inter-annotator agreement study suggests that there must be some core indicators of different types of power that we could obtain by combining multiple perceptions. We leave that to future work. For the rest of this chapter, we rely on the original annotations for the perception we are modeling.

| |
|---|
| From: William S Bradford; To: Barry Tycholiz; CC: Michael Tribolet<br>_____<br>Barry,<br>Let me know if you have any time to review.<br>Bill<br><br>From: Barry Tycholiz; To: Stephanie Miller<br>_____<br>Steph,<br>further to our discussion, Pls review.<br>I took a quick look at the locations and most appear to be East based. You might want to use an analyst to figure this out. Also, they have valued the inventories off of the Nymex only ( or so it appears) and I would have to believe that the value of these molecules is materially different than this.<br>Pls review and let's discuss asap.<br>BT |

| | |
|---|---|
| Person with SP | William S Bradford (over Barry Tycholiz); Barry Tycholiz (over Stephanie Miller) |
| Person with CNTRL | William S Bradford |
| Person with INFL | N/A |

Table 10.4: Example email thread with power annotations.

## 10.5 Problem: Predicting Persons With Power

In this chapter, we look at the problem of detecting persons with power. This problem is different from the problem of detecting the direction of power between pairs of participants that we looked at in Chapter 7 to Chapter 9. Here, we are interested in predicting whether a participant has power over someone else in the thread. More formally,

> *For each type of power $P$, for each participant $X$, we would like to predict whether $X$ has power of type $P$ over some other participant in the thread.*

We look at this problem in the context of each of the four types of power — hierarchical power, situational power, influence, and power over communication. For hierarchical power, we use the

same Enron gold organizational hierarchy from (Agarwal et al. 2012) as we did in Chapter 7. We labeled a participant to have hierarchical power within a thread if there exist a dominance pair in the gold hierarchy such that he/she dominates any other participant in the same thread. For other types of power, we use the annotations described in Section 10.3 to label whether a participant has a particular type of power.

We start by describing the features we use for this analysis before we present the results obtained on the statistical analysis performed for each type of power on each feature. We then describe the machine learning system we built and discuss the results obtained using it in various experiments conducted.

## 10.6 Features

We analyze features along six different aspects of interactions in this chapter: POSITIONAL, VERBOSITY, DIALOG ACTS, DIALOG LINKS, OVERT DISPLAY OF POWER, and LEXICAL. We describe these features below. The features are also summarized in Table 10.5. POSITIONAL, VERBOSITY, and LEXICAL are same as what we used in Chapter 7. For DIALOG ACTS, DIALOG LINKS, and OVERT DISPLAY OF POWER, we use gold annotations present in our corpus.

### 10.6.1 Positional Features

We use features that denote the placement of the participant's messages relative to the thread.

- *Initiator*: is a binary feature denoting whether the participant was the initiator of the thread.
- *FirstMsgPos*: denotes the position where the participant sent his or her first message normalized by the total number of messages in the thread.
- *LastMsgPos*: denotes the position where the participant sent his or her last message normalized by the total number of messages in the thread.

### 10.6.2 Verbosity Features

We use features denoting how verbose the participant is within the thread.

- *MsgCount*: denotes the number of messages sent by the participant.

| Aspects | Features | Description |
|---|---|---|
| PST | *Initiator* | did $p$ sent the first message in the thread? |
| | *FirstMsgPos* | relative position of $p$'s first message in the thread |
| | *LastMsgPos* | relative position of $p$'s last message in the thread |
| VRB | *MessageCount* | Count of messages sent by $p$ in the thread |
| | *MessageRatio* | Ratio of messages sent in the thread |
| | *TokenCount* | Count of tokens in messages sent by $p$ in the thread |
| | *TokenRatio* | Ratio of tokens across all messages in the thread |
| | *TokenPerMsg* | Number of tokens per message in messages sent by $p$ in the thread |
| DA | *ReqAction%* | % of Request Action dialog acts in $p$'s messages |
| | *ReqInform%* | % of Request Information dialog acts in $p$'s messages |
| | *Inform%* | % of Inform dialog acts in $p$'s messages |
| | *InformOffline%* | % of Inform-Offline dialog acts in $p$'s messages |
| | *Conventional%* | % of Conventional dialog acts in $p$'s messages |
| | *Commit%* | % of Commit dialog acts in $p$'s messages |
| DL | *FLinkCount* | Number of forward links in $p$'s messages |
| | *SFLinkCount* | Number of secondary forward links in $p$'s messages |
| | *BLinkCount* | Number of backward links in $p$'s messages |
| | *CLinkCount* | Number of connected links in $p$'s messages |
| | *DLinkCount* | Number of dangling links in $p$'s messages |
| | *DLinkRatio* | Ratio of dangling links to forward links in $p$'s messages |
| ODP | *ODPCount* | Number of instances of overt displays of power by $p$ |
| LEX | *LemmaNGram* | Word lemma ngrams |
| | *POSNGram* | Part of speech (POS) ngrams |
| | *MixedNGram* | POSNGrams, with closed classes replaced with lemmas |

Table 10.5: Aspects of interactions analyzed to study different types of power.

- *MsgRatio*: denotes the proportion of messages sent by the participant compared to the total number of messages in the thread.
- *TokenCount*: denotes the number of tokens used by the participant.
- *TokenRatio*: denotes the proportion of tokens used by the participant compared to the total number of tokens in the thread.
- *TokenPerMsg*: denotes the average number of tokens per messages sent by the participant.

### 10.6.3 Dialog Act Features

We use features to denote the count of dialog acts. We accumulate the counts over all the messages sent by the participant and then represented them as a percentage of all dialog acts by the same participant. We use the gold annotations for dialog acts in our corpus to extract these features. Various dialog act percentage features we considered are listed below.

- *ReqAction%*: percentage of Request-Action dialog acts
- *ReqInform%*: percentage of Request-Information dialog acts
- *Inform%*: percentage of Inform dialog acts
- *InformOffline%*: percentage of Inform-Offline dialog acts (the gold annotations contained 3 cases of Inform that were tagged as Inform offline, which capture inform statements that are in response to questions that appears to have happened offline)
- *Conventional%*: percentage of Conventional dialog acts
- *Commit%*: percentage of Commit dialog acts

### 10.6.4 Dialog Link Features

We use counts of various types of dialog structure links between DFUs as features. We use absolute counts here rather than relative counts since there is no obvious maximal number of links against which to compare. We use the gold annotations for these links in our corpus to extract these features.

- *FlinkCount*: denotes the total number of forward links in messages sent by the participant. Forward link is placed on a dialog functional unit if there is an explicit request for action or information.

- *SFlinkCount*: denotes the total number of SFlinks in messages sent by the participant. Secondary forward links are cases where other people interpret a segment of text as a request and respond to it, even though there was no explicit request.

- *BLinkCount*: denotes the total number of back links in messages sent by the participant. Back links denote instances when participant is responding to some one else's request or informs.

- *CLinkCount*: denotes the number of Blinks by other people connected back to DFUs in messages sent by the participant. This include both Flinks and SFlinks.

- *DLinkCount*: denotes the number of Flinks by the participant that were not connected back via Blinks by other people ("dangling links"). These are requests with no responses.

- *DLinkRatio*: denotes dangling links as a percentage of number of forward links by the participants.

### 10.6.5 Overt Displays of Power

This set is a singleton set with one feature capturing the instances of overt displays of power (Chapter 6, Section 6.1, page 74).

- *ODPCount*: Count of instances of overt displays of power in messages sent by the participant. We use the gold ODP annotations present in our corpus.

### 10.6.6 Lexical Features

As in Chapter 7, we use lexical features also in this chapter.

- *LemmaNGrams*: word lemma ngrams

- *POSNGrams*: part-of-speech ngrams

- *MixedNGrams*: a special case of *LemmaNGrams* where words belonging to open classes are replaced with their part-of-speech tags, thereby being able to capture longer sequences without increasing the dimensionality as much as *LemmaNGrams* do.

## 10.7 Statistical Analysis: Different Types of Power

In this section, we present the results of a statistical analysis of the various dialog behavior features with respect to people with the four types of power. For each type of power (HP, SP, INFL and PC), we consider two populations of people who participated in the dialog: $\mathcal{P}$, those judged to have that type of power, and $\mathcal{N}$, those not judged to have that power. Then, for each feature, we perform a two-sample, two-tailed Student's t-Test comparing means of feature values of $\mathcal{P}$ and $\mathcal{N}$. Table 10.6 presents means of each feature value for both populations $\mathcal{P}$ and $\mathcal{N}$ (as "mean($\mathcal{P}$)| mean($\mathcal{N}$)") along with the p-value associated with the t-Test as the subscript. For $p < 0.05$, we reject the null hypothesis and consider the feature to be statistically significant (boldfaced in Table 10.6).

We find many features which are statistically significant, which suggests that power types are reflected in the dialog structure. The t-Test results also show that significance of features differ considerably from one type of power to another, which suggests that different power types are reflected differently in the dialog structure, and that they are thus indeed different types of power.

### 10.7.1 Findings

For hierarchical power, we find that people with hierarchical power are less active in threads than those without. For example, persons with hierarchical power tend to talk less within a thread (*Token-Ratio*). They tend to start participating much later in the threads (*FirstMsgPos*) and do not initiate threads often (*Initiator*). This is in contrast to our findings in the analysis presented in Chapter 7 (Section 7.4) on how superiors and subordinates differ in their behavior. However, the problem formulation there was different; i.e., there we were classifying the direction of power in a pair of participants, whereas here we are classifying participants to have power over someone else or not. The data points in this formulation includes participants from threads that did not have any hierarchically related people, whereas the formulation in Chapter 7 does not. Reading these two findings together suggests that if a person starts an email thread, he's likely not to be the one who has power,

but if a thread includes a pair of people who are hierarchically related, then it is likely to be initiated by the superior and he/she tends to contribute more in such threads.

Situational power and power over communication manifest in stark contrast from hierarchical power. Persons with situational power and persons with power over communication both tend to contribute more within a thread (*TokenCount* and *TokenRatio*). They also tend to be the initiators of the thread (*Initiator*) or start participating in the thread closer to the beginning (*FirstMsg*). Situational power and power over communication have many other features which are also statistically significant. For example, they send significantly more messages (*MsgCount*). They also have significantly more instances of overt displays of power (*ODPCount*) than others. It is interesting to note that *ODPCount* was not a significant feature for HP. It suggests that bosses don't always display their power overtly when they interact. Situational power and power over communication also differ from one another. For example, those with situational power tend to request actions (*ReqAction%*) significantly more than those without. However, this was not significant in case of power over communication. Similarly, the number of back links (*BLinkCount*) was not a significant feature for situational power. But, people with power over communication tend to have significantly fewer back links (*BLinkCount*) than those without. Situational power and power over communication differ also in terms of the magnitude of differences in *Initiator*, *FirstMsgPos*, *TokenCount*, and *TokenRatio*. People with power over communication are the initiators of the threads almost eight times as often as those without power over communication, whereas people with situational power initiated threads only 42% more often than those without situational power. Similarly, the ratio of contributions by people with situational power (*TokenRatio*) was only 32% more than that of those without, whereas for power over communication, this difference was 180%. On the other hand, people with situational power use almost 5.6 times overt displays of power than those without situational power, whereas people with power over communication use only 2.3 times more overt displays of power than those without.

The finding that people with power over communication have fewer back links is interesting, since it aligns power over communication with the characterization of control by Walker and Whittaker (1990). According to them, control over a discourse segment is determined by whether the participant provide unsolicited information in the dialog or not. In the dialog act annotation scheme we use, solicited information (in other words, responses to requests and commands) places an oblig-

| Features | HP | SP | PC | INFL |
|---|---|---|---|---|
| | | POSITIONAL features | | |
| *Initiator* | $\mathbf{0.27}\vert\mathbf{0.57_{0.02}}$ | $\mathbf{0.68}\vert\mathbf{0.48_{3.3E-3}}$ | $\mathbf{0.88}\vert\mathbf{0.11_{3.4E-44}}$ | $0.64\vert0.55_{0.58}$ |
| *FirstMsgPos* | $\mathbf{0.34}\vert\mathbf{0.19_{0.02}}$ | $\mathbf{0.13}\vert\mathbf{0.24_{1.1E-3}}$ | $\mathbf{0.05}\vert\mathbf{0.40_{1.4E-28}}$ | $0.16\vert0.21_{0.55}$ |
| *LastMsgPos* | $0.47\vert0.37_{0.08}$ | $0.41\vert0.36_{0.21}$ | $\mathbf{0.31}\vert\mathbf{0.47_{1.9E-5}}$ | $0.32\vert0.38_{0.51}$ |
| | | VERBOSITY features | | |
| *MessageCount* | $1.33\vert1.46_{0.47}$ | $\mathbf{1.68}\vert\mathbf{1.32_{0.03}}$ | $\mathbf{1.62}\vert\mathbf{1.22_{1.3E-3}}$ | $1.45\vert1.45_{0.99}$ |
| *MessageRatio* | $0.48\vert0.52_{0.47}$ | $0.54\vert0.50_{0.18}$ | $\mathbf{0.61}\vert\mathbf{0.39_{2.8E-15}}$ | $0.45\vert0.52_{0.19}$ |
| *TokenCount* | $53.22\vert91.53_{0.06}$ | $\mathbf{113.04}\vert\mathbf{74.19_{0.02}}$ | $\mathbf{121.38}\vert\mathbf{43.90_{1.1E-8}}$ | $143.55\vert85.54_{0.10}$ |
| *TokenRatio* | $\mathbf{0.35}\vert\mathbf{0.54_{0.04}}$ | $\mathbf{0.62}\vert\mathbf{0.47_{2.1E-3}}$ | $\mathbf{0.72}\vert\mathbf{0.26_{1.0E-28}}$ | $0.63\vert0.52_{0.26}$ |
| *TokensPerMsg* | $39.73\vert63.45_{0.13}$ | $73.22\vert54.76_{0.07}$ | $\mathbf{78.27}\vert\mathbf{38.91_{1.3E-5}}$ | $118.94\vert58.52_{0.09}$ |
| | | DIALOG ACTS features | | |
| *ReqAction%* | $0.10\vert0.02_{0.23}$ | $\mathbf{0.07}\vert\mathbf{0.01_{0.01}}$ | $0.03\vert0.04_{0.48}$ | $\mathbf{0.0}\vert\mathbf{0.04_{6.9E-5}}$ |
| *ReqInform%* | $0.10\vert0.11_{0.87}$ | $0.10\vert0.12_{0.70}$ | $0.11\vert0.11_{0.91}$ | $0.09\vert0.11_{0.73}$ |
| *Inform%* | $0.56\vert0.60_{0.63}$ | $0.56\vert0.63_{0.10}$ | $0.60\vert0.61_{0.79}$ | $\mathbf{0.78}\vert\mathbf{0.59_{0.01}}$ |
| *InformOffline%* | $\mathbf{0.00}\vert\mathbf{0.005_{0.04}}$ | $0.003\vert0.005_{0.62}$ | $\mathbf{0.008}\vert\mathbf{0.0_{0.04}}$ | $\mathbf{0.0}\vert\mathbf{0.005_{0.04}}$ |
| *Conventional%* | $0.23\vert0.24_{0.96}$ | $0.25\vert0.23_{0.35}$ | $0.24\vert0.23_{0.81}$ | $\mathbf{0.13}\vert\mathbf{0.24_{0.04}}$ |
| *Commit%* | $0.0\vert0.002_{0.21}$ | $0.001\vert0.003_{0.51}$ | $0.001\vert0.004_{0.44}$ | $0.0\vert0.002_{0.21}$ |
| | | DIALOG LINKS features | | |
| *FLinkCount* | $0.56\vert0.74_{0.27}$ | $\mathbf{0.98}\vert\mathbf{0.59_{0.03}}$ | $\mathbf{0.91}\vert\mathbf{0.49_{6.2E-3}}$ | $0.45\vert0.74_{0.35}$ |
| *SFLinkCount* | $0.16\vert0.34_{0.09}$ | $\mathbf{0.49}\vert\mathbf{0.24_{0.02}}$ | $\mathbf{0.43}\vert\mathbf{0.21_{0.01}}$ | $0.64\vert0.32_{0.07}$ |
| *BLinkCount* | $0.94\vert0.61_{0.23}$ | $0.72\vert0.59_{0.40}$ | $\mathbf{0.41}\vert\mathbf{0.94_{1.7E-4}}$ | $1.00\vert0.61_{0.39}$ |
| *CLinkCount* | $\mathbf{0.27}\vert\mathbf{0.61_{0.04}}$ | $\mathbf{0.83}\vert\mathbf{0.44_{7.1E-3}}$ | $\mathbf{0.75}\vert\mathbf{0.35_{6.9E-4}}$ | $0.73\vert0.57_{0.46}$ |
| *DLinkCount* | $0.44\vert0.49_{0.79}$ | $0.64\vert0.39_{0.08}$ | $0.58\vert0.35_{0.06}$ | $0.36\vert0.49_{0.67}$ |
| *DLinkRatio* | $0.39\vert0.24_{0.24}$ | $\mathbf{0.33}\vert\mathbf{0.21_{0.05}}$ | $0.27\vert0.24_{0.57}$ | $0.18\vert0.26_{0.55}$ |
| | | OVERT DISPLAY OF POWER features | | |
| *ODPCount* | $0.50\vert0.36_{0.30}$ | $\mathbf{0.78}\vert\mathbf{0.14_{6.0E-8}}$ | $\mathbf{0.49}\vert\mathbf{0.21_{2.6E-3}}$ | $\mathbf{0.09}\vert\mathbf{0.39_{0.01}}$ |

Table 10.6: Variations in manifestations of power on feature values: mean($\mathcal{P}$)| mean($\mathcal{N}$)$_{p-value}$.
$\mathcal{P}$: people judged to have power; $\mathcal{N}$: people judged not to have power; Values with $p \leq 0.05$ are boldfaced
SP: Situational power, HP: Hierarchical power, PC: Power over Communication, INFL: Influence

atory Blink on the corresponding text segment. Hence, the fact that people with power over communication have significantly larger contributions to the dialog (VRB features), but with fewer back links, suggest that most of their contribution is unsolicited information. This is in line with Walker and Whittaker (1990)'s definition of control over discourse segments.

Although the power type influence (INFL) has fewer data points in our data, we found a few significant features for people with influence. People with INFL never request actions (*ReqAction*) as opposed to those with situational power who request for actions more frequently than others. Also, people with INFL tend to have significantly more inform utterances (*Inform*). They also have significantly fewer overt displays of power (*ODpower over communicationount*) than others, a stark contrast to those with situational power and power over communication.

### 10.7.2 Multiple Test Correction

The statistical measures presented in this section are exploratory in nature, presenting tests on all combinations of features and power types. We do not draw theoretical conclusions from the specific combination of interactions that are found statistically significant. Hence, we did not apply any corrections for multiple tests in statistical significance for individual features. When we apply the Bonferroni correction for multiple tests to adjust the p-value for number of test performed (threshold = 0.05/84 = 6.0E-4), 10 features would still remain statistically significant. Hence the global null hypothesis that the features we considered do not interact with the power types would still be rejected.

## 10.8 Experiments and Results

In this section, we present a system to predict whether a person has a given type of power in the context of an email thread. We show that different sets of features are helpful to detect different types of power. We build a separate binary classifier for each power type $\mathcal{P}$ predicting whether or not a given participant in a communication thread $\mathcal{X}$ has that type of power or not.

### 10.8.1 Implementation

As described in Chapter 4, we use the tokenizer, POS tagger, lemmatizer and SVMLight (Joachims 1999) wrapper in the ClearTK (Ogren et al. 2008) package. The ClearTK wrapper for SVMLight internally shifts the prediction threshold based on a posterior probabilistic score calculated using Lin et al. (2007)'s algorithm.

### 10.8.2 Experiments

We first find the best performing subset of features for each feature set by exhaustive search within the set. Once we have the best subset of each feature set, we do another round of exhaustive search combining best performers of each set to find the overall best performing feature subset. We experimented with a linear kernel and a quadratic kernel; the latter performed better. All results presented in this paper are obtained using a quadratic kernel.

### 10.8.3 Handling Class Imbalance

Since our dataset is skewed especially for hierarchical power and influence (with very few persons with power), we balanced our dataset by up-sampling minority class instances in the training step. This has proven useful in cases of unbalanced datasets (Japkowicz 2000). All results presented below have been obtained after balancing the training folds in cross validation; we balance only the training folds, the test folds remain unchanged.

### 10.8.4 Evaluation

We report results 5-fold cross validation on the data to evaluate the prediction performance for different feature subsets. The corpus was split into folds at the thread level. The corpus was divided into 5 folds at the thread level. Active participants from 4 folds were used to train a model which was then tested on active participants in the 5th fold. We did this with all five configurations and all the reported results in this paper are micro-averaged results across 5 folds.

### 10.8.5 Results

We now describe the results obtained in our experiments. Table 10.7-Table 10.10 show cross validation results for all four types of power for each set of features. For each power type, the table also lists (in the last row) the best performing feature subset combination and corresponding results.

**Baseline Approaches:** We present two simple baseline measures - **Random** and **AlwaysTrue**. In the Random baseline, we predict an active participant to have the particular type of power at random. In AlwaysTrue baseline, we always predict an active participant to have power.

**Hierarchical Power:** Hierarchical power is hard to predict, which could partly be due to the very small number of positive training examples in the corpus. All feature subsets except LEXICAL outperformed the other baselines of 11.3% and 15.0% (for Random and AlwaysTrue respectively). Overall, it was hard to obtain high precision; the highest precision posted by an individual feature set was 16.7% by VERBOSITY with a recall of 44.4%. POSITIONAL obtained the highest recall of 72.2%, however with a very low precision of 13.8%. DIALOG LINKS obtained the best F-measure of 24.4. A combination of VERBOSITY, POSITIONAL and OVERT DISPLAY OF POWER gave the best model obtaining an F measure of 29.5%, improving the precision to 20.9%.

**Situational Power:** For situational power, the random and AlwaysTrue baselines gave F measures of 42.1 and 53.6 respectively. The best performers of all feature sets except DIALOG ACTS outperformed these baselines. The best performing individual feature sets are OVERT DISPLAY OF POWER and DIALOG LINKS, both at or near 60.0%. While OVERT DISPLAY OF POWER gave a high precision (71.2%) model, DIALOG LINKS gave a high recall (75.3%) model, the combination of both gave the best performing system with an F measure of 64.4%.

**Power over Communication:** For power over communication, the best single feature was *FirstMsgPos* (relative position of first message). This is because the person with the power over communication is almost always the initiator of the thread. Note that the notion of PC is not defined in terms of positional features: annotators were asked to find the participants who "actively attempt to achieve the intended goals of the communication". It is our finding that those who are in PC were also the ones who did initiate the thread. It is also worth noting that OVERT DISPLAY OF POWER is

| Feature Set | Precision | Recall | F-measure |
|---|---|---|---|
| Random | 16.6 | 38.9 | 11.3 |
| AlwaysTrue | 8.1 | 100.0 | 15.0 |
| POSITIONAL (PST) | 13.8 | **72.2** | 23.2 |
| VERBOSITY (VRB) | **16.7** | 44.4 | 24.2 |
| DIALOG ACTS (DA) | 16.0 | 22.2 | 18.6 |
| DIALOG LINKS (DL) | 15.3 | 61.1 | **24.4** |
| OVERT DISPLAY OF POWER (ODP) | 15.3 | 50.0 | 23.4 |
| LEXICAL (LEX) | 0.0 | 0.0 | 0.0 |
| VRB +PST +ODP | 20.9 | 50.0 | 29.5 |

Table 10.7: Cross validation results on predicting persons with hierarchical power.

| Feature Set | Precision | Recall | F-measure |
|---|---|---|---|
| Random | 36.7 | 49.4 | 42.1 |
| AlwaysTrue | 36.7 | 100.0 | 53.6 |
| POSITIONAL (PST) | 45.1 | 67.9 | 54.2 |
| VERBOSITY (VRB) | 43.9 | 70.4 | 54.0 |
| DIALOG ACTS (DA) | 40.9 | 75.3 | 53.0 |
| DIALOG LINKS (DL) | 49.6 | **75.3** | 59.8 |
| OVERT DISPLAY OF POWER (ODP) | **71.2** | 51.9 | **60.0** |
| LEXICAL (LEX) | 54.9 | 55.6 | 55.2 |
| DL +ODP | 59.4 | 70.4 | 64.4 |

Table 10.8: Cross validation results on predicting persons with situational power.

| Feature Set | Precision | Recall | F-measure |
|---|---|---|---|
| Random | 57.5 | 51.2 | 54.2 |
| AlwaysTrue | 57.5 | 100.0 | 73.0 |
| POSITIONAL (PST) | **91.8** | 88.2 | **90.0** |
| VERBOSITY (VRB) | 78.7 | 84.3 | 81.4 |
| DIALOG ACTS (DA) | 60.5 | **92.9** | 73.3 |
| DIALOG LINKS (DL) | 74.3 | 81.9 | 77.9 |
| OVERT DISPLAY OF POWER (ODP) | 74.6 | 34.7 | 47.3 |
| LEXICAL (LEX) | 70.2 | 78.0 | 73.9 |
| PST | 91.8 | 88.2 | 90.0 |

Table 10.9: Cross validation results on predicting persons with power over communication.

the worst performer for power over communication which is in contrast with the case of situational power, supporting the claim that these two types of power are in fact different.

**Influence:** Influence is another very hard class to predict, again, possibly partly due to the very small number of positive training examples. The simple baseline F-measures were both 9.5. All feature sets except POSITIONAL and LEXICAL outperformed these baseline measures. The best performance was obtained by DIALOG LINKS with counts of *BLinkCount*, *FLinkCount*, *DLinkCount* and *SFLinkCount* as features. That system posted an F-measure of 22.6, with the highest precision of 13.7% and a recall of 63.6%. The best recall of 90.9% was obtained using OVERT DISPLAY OF POWER; remember that we had found in our statistical analysis that people with influence never use overt displays of power and it has turned out to be a very useful feature for the prediction task.

**Statistical Significance of Results:** For assessing statistical significance of F measure improvements over baseline, we used the Approximate Randomness Test (Yeh 2000). We found the improvements to be statistically significant for hierarchical power ($p < 0.001$), situational power ($p < 0.001$) and power over communication ($p < 0.01$). However, for influence, the improvement was not statistically significant (p = 0.3).

| Feature Set | Precision | Recall | F-measure |
|---|---|---|---|
| Random | 5.2 | 54.6 | 9.5 |
| AlwaysTrue | 5.0 | 100.0 | 9.5 |
| POSITIONAL (PST) | 4.6 | 45.5 | 8.4 |
| VERBOSITY (VRB) | 8.1 | 81.8 | 14.8 |
| DIALOG ACTS (DA) | 6.9 | 63.6 | 12.4 |
| DIALOG LINKS (DL) | **13.7** | 63.6 | **22.6** |
| OVERT DISPLAY OF POWER (ODP) | 6.2 | **90.9** | 11.6 |
| LEXICAL (LEX) | 0.0 | 0.0 | 0.0 |
| DL | 13.7 | 63.6 | 22.6 |

Table 10.10: Cross validation results on predicting persons with influence.

## 10.9 Conclusion

In this chapter, we introduced a new typology of four types of power in the context of organizational email — hierarchical power, situational power, power over communication, and influence. We described in detail the procedure we followed to obtain manual annotations of the different types of power. We then showed that these types of power are manifested very differently with respect to the features we are using, which validates our claim that these are indeed different types of power. We also presented a supervised learning system to predict persons with one of the types of power in written dialog yielding encouraging results. Like what we saw in previous chapters, we found that dialog features are very significant in predicting other types of power relations as well.

# Chapter 11

# Power of Confidence in Political Debates

Analyzing political speech has recently gathered great interest within the NLP community. Studies range from identifying markers of persuasion (Guerini et al. 2008), to predicting voting patterns (Thomas et al. 2006, Gerrish and Blei 2011), to detecting ideological positions (Sim et al. 2013). Studies have also analyzed how personal attributes of political personalities such as charisma affect their public discourse (Rosenberg and Hirschberg 2009). However, there has not been much work applying computational techniques to understand how external factors affect the political discourse. In this thesis, we investigate how the power of confidence that an election candidate has in terms of the support they are getting from the electorate is manifested in the ways they interact in political debates. Specifically, we study the US Republican party presidential primary debates conducted as part of the 2012 US Presidential election campaign. We model the power of confidence each candidate has based on their relative standings in the polls released prior to the debate.

We start by describing the domain in more detail (Section 11.1) and explaining how we model power in this domain (Section 11.2). In Section 11.3, we discuss the different aspects of interaction we analyze in this domain. Section 11.4 presents the statistical analysis studying how these aspects of interaction correlate with the candidates' power. Section 11.5 describes an automatic power ranker, a system that automatically ranks the candidates in terms of their power differential, and presents various experiments and results.

## 11.1 Domain: Political Debates

Before every United States presidential election, a series of presidential primary elections are held in each U.S. state by both major political parties (Republican and Democratic) to select their respective presidential nominees. These primaries are staggered between January and June before the general election in November. In recent times, it has become customary that candidates of both parties engage in a series of debates prior to and during their respective parties' primary elections. In this chapter, we explore how the power differential between these candidates manifests in these debates. Specifically, we use the 20 debates held between May 2011 and February 2012 as part of the 2012 Republican presidential primaries. (There were no Democratic presidential primary debates in 2012, since the incumbent President Barack Obama was the de-facto nominee.) There were a total of 10 candidates who took part in these primary debates; some of whom participated only in one or two debates. On an average, there were 6.6 participants per debate.

Presidential debates represent a domain of interactions which is fairly well structured. Most debates follow the pattern where the moderator asks questions directly to the candidates, or takes questions from the audience or people calling in via phone/video calls. Candidates to whom the questions were addressed then respond, after which the moderator asks other candidates for their responses or comments. This pattern of the moderator prompting and the candidates answering is often maintained across the debates with some disruptions due to out-of-turn talking and interruptions from other candidates.

Presidential debates serve an important role during the election process. It serves as a platform for candidates to discuss their stances on policy issues and contrast them with other candidates' stances. In addition, it also serves as a medium for the candidates to pursue and maintain power over other candidates. This makes it an interesting domain to investigate how power dynamics between participants are manifested in an interaction. In addition, the 2012 Republican presidential election campaign was one of the most volatile ones in recent times. Most candidates held the front runner position at some point during the campaign. This prevents the analysis of power dynamics in these debates from being biased by the personal characteristics of a single candidate or a small set of candidates.

## 11.2 Modeling Power in Debates

We use the term *Power Index* to denote the power or confidence with which a candidate comes into the debate. The Power Index of a candidate can be influenced by various factors. In Figure 11.1, we present the dependency flow diagram between some of the prominent components of the election process as they relate to the candidates' Power Index. Note that this diagram is not intended to be inclusive of all components in the election process; rather only the main ones. These components are as follows. (1) During the presidential primary election campaigns, candidates get endorsed by various political personalities, newspapers and businesses. We feel that such endorsements as well as the funds raised through campaigns positively affect the sense of power of the candidate. (2) The last few debates are held after the series of primary elections have begun. For example, the Iowa caucus is held in early January before the last 7 debates. For these debates, the results from the states where the primaries are over could also influence the level of power and confidence of the candidates. (3) However, a more important source of power or confidence is the relative standings in the recent poll scores. It gives the candidate a sense of how well he or she is successful in convincing the electorate of his/her candidature. Note that some of these components are inter-dependent (e.g., the funds raised may impact poll scores through advertising). We have not captured these inter-dependencies in Figure 11.1.

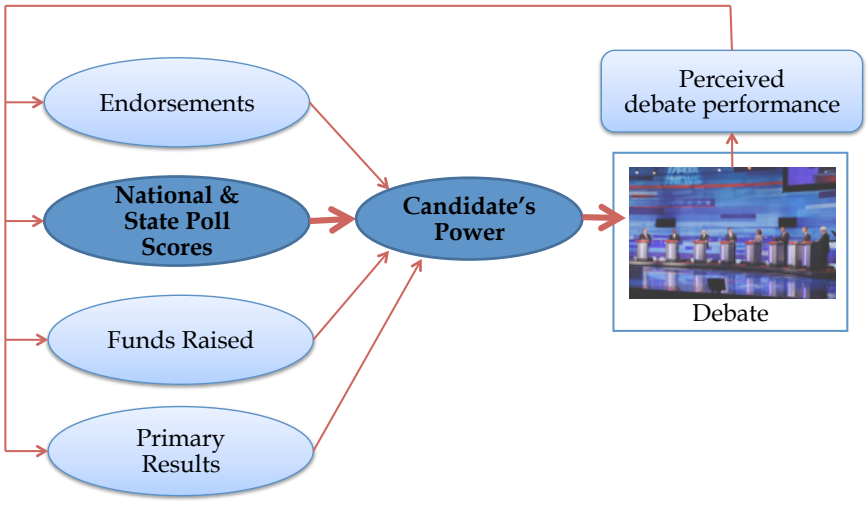



Figure 11.1: Dependency flow diagram of factors affecting the candidate's power index.

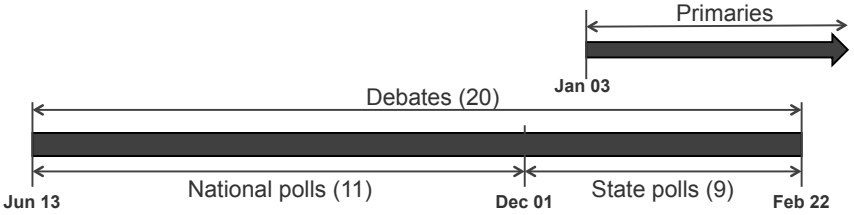



Figure 11.2: Timeline of debates and primaries.

We hypothesize that the sense of power or confidence the candidate draws from these various sources affect the form of interaction — how he/she interacts with others as well as how others interact with him/her. — in the debate. In turn, the interaction in the debates may also have a feedback effect on some of the components. The perception of how well or badly a candidate performed in a debate might trigger positive or negative opinions in the electorate and affect the different components such as endorsements, poll scores, funds, and the primary results. An example for such an effect is the lowering of poll scores for the Democratic candidate President Barack Obama after the first presidential debate in which his performance was largely considered sub-par.[1]

In this thesis, we focus on how the power or confidence of the candidate impacts the manner of interactions within the debate. We propose modeling the power of a candidate with a *Power Index* that is computed using the components in Figure 11.1. In this study, we model the Power Index of each candidate based solely on their recent state or national poll standings because we think that this is the most dominant factor. Other components such as the funds raised can be included in a similar fashion in the calculation of Power Index.

### 11.2.1 Timeline of Debates and Primaries

Figure 11.2 shows the timeline of debates and primaries held as part of the 2012 Republican primary election. Debates from December 2011 onwards were held in states where the primaries were to be held in the near future. For example, the debates on December 10th and 15th were held in the state of Iowa where the primary was scheduled for January 3rd 2012. Similarly, all debates in January and February were also held in states where the primaries were to be held few days after the debate.

[1] http://www.gallup.com/poll/157907/romney-narrows-vote-gap-historic-debate-win.aspx

In these debates, we assume that their standings in the respective state polls, rather than national polls, would be the dominating factor affecting the power or confidence of candidates. Hence, for those debates, we chose the respective state's poll scores as the reference. For others, we chose the national polls as the reference.

### 11.2.2 Power Index of a Candidate

For each debate $D$, we denote the set of candidates participating in that debate by $C_D$. Let $date(D)$ denote the date on which debate $D$ was held and $state(D)$ denote the state in which it was held. Let $refPollType$ denote the type of the reference poll we consider for debate $D$.

$$refPollType = \begin{cases} state(D), & \text{if } date(D) > 12/01/11 \\ NAT, & otherwise \end{cases}$$

We show the $refPollType$ for each debate in Figure 11.2. For the first eleven debates, we use the national poll scores as reference, whereas for the rest of the debates we use the states polls from the respective states in which they were held. For each debate, we find the poll results (national or state) released most recently and use the percentage of electorate supporting each candidate as the power index. If there are multiple polls released on the day the most recent poll was released, then we take the mean of poll scores from all those polls to find the power index. Let $RefPolls(D)$ be the set of polls of type $refPollType$ released on the most recent date on which one or more such polls were released before $date(D)$. We define the $PowerIndex$, $\mathcal{P}(X)$, of candidate $X \in C_D$ as below

$$\mathcal{P}(X) = \frac{1}{|RefPolls(D)|} \sum_{i=1}^{|RefPolls(D)|} p_i$$

where $p_i$ denote the poll percentage $X$ got in the $i^{th}$ poll in $RefPolls(D)$. Figure 11.3 shows the trend of how the power indices of candidates varied across debates. As the figure shows, almost every candidate was one of the top candidates during some of the debates. Table 11.1 lists the number of debates each candidate participated in and the number of times they were among the top three candidates as per the power indices.

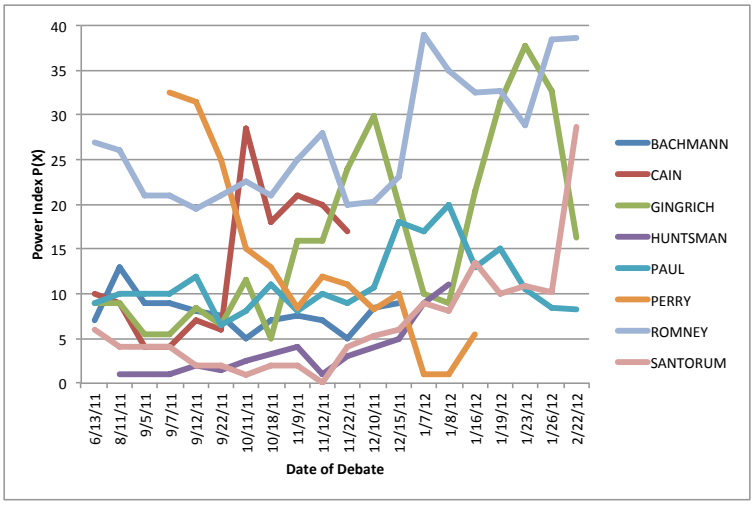



Figure 11.3: Power index P(X) variations across debates.

Note: Plots for Pawlenty and Johnson are not shown since they participated only in one or two debates.

| Candidate | # of debates | # of times in top 3 |
|---|---|---|
| Bachmann | 13 | 3 |
| Cain | 11 | 6 |
| Gingrich | 20 | 12 |
| Huntsman | 11 | 1 |
| Johnson | 1 | 0 |
| Paul | 20 | 9 |
| Pawlenty | 2 | 0 |
| Perry | 13 | 5 |
| Romney | 20 | 20 |
| Santorum | 19 | 4 |

Table 11.1: Candidates' participation and power standings based on their power indices

## 11.3 Aspects of Interactions Analyzed

In the work presented in this chapter, we analyze four different aspects of interaction to study the manifestations of power. We summarize these aspects in detail below. We use features to capture the language used in the debates as well as the structure of debates. Specifically, we analyze each debate participant in four dimensions — what they said (LEXICAL), how much they spoke (VERBOSITY), how they argued (TURN TAKING), and how they were talked about (MENTIONS). Some structural features such as turns information are readily available from the transcripts, while for some others like arguments and candidate mentions, we use simple heuristics or perform deeper NLP analysis. In addition to looking at how each candidate interacted with others, we also look at how others interacted with them. The features we use are described in detail below and are summarized in Table 11.2. Each feature $f$ is extracted with respect to a the candidate $X$.

### 11.3.1 Verbosity Features

We hypothesize that candidates' power will impact the proportion of turns they get to speak, the time duration they are allowed to speak, and the number of questions posed to them. The turns proportion is directly available from the transcripts. We approximated the time duration each speaker spoke by the total number of words spoken by him/her. To find the number of questions asked, we used the following heuristic — instances where the candidate spoke right after the moderator are answers to questions the moderator posed to the candidate. We verified this to be a reliable heuristic manually. The raw counts of turns, words and questions are dependent on the length of each debate, which varied from 90-120 minutes. One option to handle this is to consider the percentages of each of them. However, the percentage values of these features are again dependent on the number of participants in each debate, which varied from 9 to 4. To handle this, we measured the deviation of each candidate's percentage of turns, words and questions from their expected fair share percentage in the debate. We define the fair share percentage in a given debate to be $1/|C_D|$ — the percentage each candidate would have gotten for that feature if it was equally distributed. We calculate the deviation of each feature as the difference between observed percentage for that feature and $1/|C_D|$. In summary, we have three deviation features — *TurnDev*, *WordDev*, and *QstnDev*. We also investigated three additional structural features related to verbosity: number of words per turn — whether

| Aspects | Features | Description |
|---|---|---|
| VRB | *WordDev* | Deviation of $X$'s *WordPercent* from the mean |
| | *TurnDev* | Deviation of $X$'s *TurnPercent* from the mean |
| | *QstnDev* | Deviation of $X$'s *QstnPercent* from the mean |
| | *WordsPerTurn* | Number of words per turn for $X$'s turns |
| | *WordsPerSentence* | Number of words per sentence for $X$'s turns |
| | *LongestTurn* | Number of words in the longest turn by $X$ |
| TT | *SpeakingOutOfTurn* | Percentage of $X$'s turns that was spoken out of turn |
| | *SpokenToOutOfTurn* | Percentage of $X$'s turns after which others spoke out of turn |
| MNT | *MentionPercent* | Percentage of mentions of candidate $X$ |
| | *FirstNamePercent* | Percentage of first name mentions of candidate $X$ |
| | *LastNamePercent* | Percentage of last name mentions of candidate $X$ |
| | *FullNamePercent* | Percentage of full name mentions of candidate $X$ |
| | *TitlePercent* | Percentage of title mentions of candidate $X$ |
| LEX | *LemmaNGram* | Word lemma ngrams of $X$'s turns |
| | *POSNGram* | Part of speech (POS) ngrams of $X$'s turns |
| | *MixedNGram* | Mixed ngrams of $X$'s turns |

Table 11.2: Aspects of interactions analyzed in political debates.
All features are described with respect to candidate $X$

they had longer turns on average, words per sentence — whether they used shorter sentences, and the longest turn length in terms of number of words. Below, we list the verbosity features discussed above with respect to candidate $X$.

- *WordDev*: deviation of percentage of words in the debate spoken by $X$ from the mean
- *TurnDev*: deviation of percentage of turns in the debate spoken by $X$ from the mean
- *QstnDev*: deviation of percentage of questions in the debate asked to $X$ from the mean
- *WordsPerTurn*: number of words per turn for $X$'s turns

- *WordsPerSentence*: number of words per sentence for $X$'s turns
- *LongestTurn*: number of words in the longest turn by $X$

### 11.3.2 Turn Taking Features

Sociolinguistics studies have found correlations between turn-taking patterns and social power relations (Ng and Bradac 1993, Ng et al. 1993, Reid and Ng 2000). We compute features to capture the turn-taking patterns exhibited by the candidates. Debates follow a pattern where a candidate is expected to speak only after a moderator prompts him or her to either answer a question or to respond to another candidate. Hence, if a candidate talks immediately after another candidate, he/she is disrupting the expected pattern of the debate. This holds true even if such an out-of-turn talk may not have interrupted the previous speaker mid-sentence. We compute features to capture such instances where the candidates speaks out-of-turn after another candidate. In most cases, such out-of-turn speaking leads to back-and-forth exchanges between the candidates until a moderator steps in. We define the series of such exchanges between candidates where they talk with one another without the moderator intervening as an argument. Arguments can extend to many number of turns. In counting out-of-turn speech instances, we count only the first out-of-turn speaking by each candidate in the series of turns that constitute an argument. An example for an argument, which is an excerpt from the debate held at Myrtle Beach, South Carolina, on January 16, 2012, is given in Figure 11.4. From this argument, we count only one instance of out-of-turn speaking for Santorum and one instance for Romney.

We use features to capture out-of-turn speaking by the candidate $X$ as well as out-of-turn speaking by others just after/while candidate $X$ was speaking. Since the raw counts of these measures are dependent on the number of turns by each candidate, we use the normalized counts to find the *per-turn* value of these measures as features. We denote the two normalized features as follows:

- *SpeakingOutOfTurn*: number of times candidate $X$ speaks out of turn, divided by number of $X$'s turns
- *SpokenToOutOfTurn*: number of times others speak out of turn while/after candidate $X$ speaks, divided by number of $X$'s turns

…
SANTORUM: ... I would ask Governor Romney, do you believe people who have -- who were felons, who served their time, who have extended -- exhausted their parole and probation, should they be given the right to vote?

WILLIAMS: Governor Romney?

ROMNEY: First of all, as you know, the PACs that run ads on various candidates, as we unfortunately know in this --

SANTORUM: I'm looking for a question -- an answer to the question first. [applause]

ROMNEY: We have plenty of time. I'll get there. I'll do it in the order I want to do. I believe that, as you realize that the super PACs run ads. And if they ever run an ad or say something that is not accurate, I hope they either take off the ad or make it -- or make it correct. I guess that you said that they -- they said that you voted to make felons vote? Is that it?

SANTORUM: That's correct. That's what the ad says.

ROMNEY: And you're saying that you didn't?

SANTORUM: Well, first, I'm asking you to answer the question, because that's how you got the time. It's actually my time. So if you can answer the question, do you believe, do you believe that felons who have served their time, gone through probation and parole, exhausted their entire sentence, should they be given the right to have a vote?
…

Figure 11.4: Debate excerpt from the debate held at Myrtle Beach, SC on January 16 2012.

Even though some of the instances of out-of-turn speaking do not necessarily interrupt the previous speaker mid-sentence, they are nonetheless interruptions to the expected debate structure in which candidates are expected to speak only in response to the moderators. Hence, in prior work (Prabhakaran et al. 2013a), we used the term *interruption* to refer to the instances of out-of-turn speaking, and used the terms *InterruptOthersPerTurn* and *OthersInterruptPerTurn* for features. However, to avoid confusion with the traditional definition of the term interruption (i.e., instances where previous turn was incomplete), we use the terms *SpeakingOutOfTurn* and *SpokenToOutOfTurn* in this thesis for our features.

In addition, there has been work in the NLP community to detect arguments and interruptions in spoken as well as written interactions (e.g., (Somasundaran et al. 2007, Cabrio and Villata 2012, Ghosh et al. 2014)). The well-structured nature of interactions that is expected in presidential debates allows us to use the simple heuristics described above to detect arguments for the purposes of this study. Deeper NLP processing of candidate turns such as the work mentioned above might help detect arguments in the debates more reliably; we leave it to future work.

### 11.3.3 Mention Features

Intuitively, how often a candidate was mentioned or referred to by others in the debate is a good indicator of his or her power. The more a candidate is mentioned, the more central he or she is in the the context of that debate. We use the mention count normalized across the total number of mentions of all candidates in a given debate (*MentionPercent*) as a feature. In addition, we look at the form of addressing used while referring to each candidate. Previous studies in social sciences and linguistics have looked at the form of addressing in relation to the social relations (Brown and Ford 1961, Dickey 1997). Building on insights from these studies, we investigated if the modes of addressing candidates change with respect to their power. Specifically, we looked at four modes of addressing — first name mentions, last name mentions, full name mentions and title mentions. As titles, we included common titles such as Mr., Ms. etc. as well as a set of domain-specific titles: Governor, Speaker, Senator, Congresswoman and Congressman. The title mentions may sometimes would just be the title without the name. For example, Newt Gingrich was often referred to as just *Speaker* since it uniquely identified him among the candidates; he was the only person among the candidates who had ever been the Speaker. About 68.6% of total candidate mentions across debates

were title name mentions, whereas the other types of mentions accounted for close to 10% each. We list the features we use with respect to the candidate mentions below.

- *MentionPercent*: percentage of mentions of candidate $X$ out of of all the candidate mentions
- *FirstNamePercent*: percentage of first name only mentions of candidate $X$
- *LastNamePercent*: percentage of last name only mentions of candidate $X$
- *FullNamePercent*: percentage of full name mentions of candidate $X$ (without any title)
- *TitlePercent*: percentage of times candidate $X$ was mentioned using his title with or without their name

### 11.3.4 Lexical Features

As in previous chapters, we also used lexical bag-of-words features for this domain. Specifically, we use the word lemma ngrams, part-of-speech ngrams and mixed ngrams. Ngram based features have been used in previous studies to analyze power in written interactions (Bramsen et al. 2011, Gilbert 2012), as well as in our own study in other domains (Chapter 7 to Chapter 10). These features are expected to capture lexical patterns that denote power relations. We aggregated all turns of a participant and extracted the following ngram features:

- *LemmaNGram*: word lemma ngrams
- *PosNGram*: part-of-speech ngrams
- *MixedNGram*: mixed ngrams (lemma ngrams with open class words' lemmas replaced with part-of-speech tags)

## 11.4 Statistical Analysis

As a first step towards understanding the manifestations of candidates' power index on how they interacted in the debates, we computed the Pearson's product correlation between each candidate's power index ($P(X)$) and each feature. We perform this study on all features described in Section 11.3 other than LEXICAL features. Figure 11.5 shows the Pearson's product correlation between

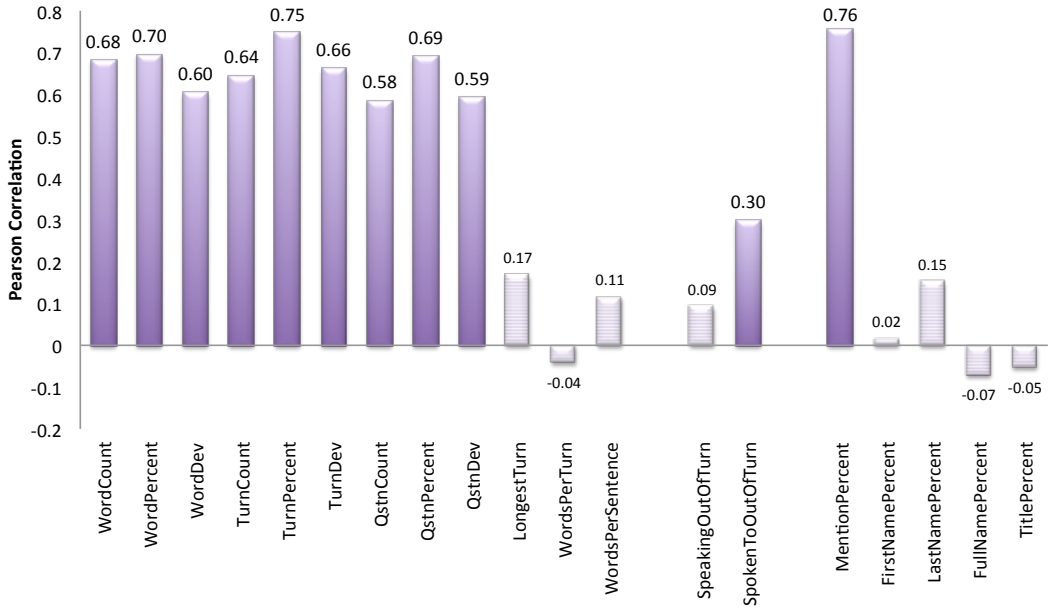



Figure 11.5: Correlations between power index and structural features.
Correlation windows: Weak (0.2 - 0.39); Moderate (0.4 - 0.69); High ( $\geq 0.7$)

each structural feature and candidate's power index $P(X)$. The darker bars denote statistically significant ($p < 0.05$) correlations. Applying Bonferroni correction for multiple tests, the threshold for p-value for significance would be reduced to 0.0025. Even then, the statistically significant features would retain their significance. We consider features with correlation values below 0.2 to have no correlation, between 0.2-0.39 to have weak correlation, 0.4-0.69 to have moderate correlation, and those above 0.7 to have high correlation.

**VERBOSITY:** We obtain statistically significant moderate positive correlation between all the word and turn features and candidates' power indices. Figure 11.5 shows the results only for the deviation measures of these features, but the correlation was similar for the raw counts as well as percentages. This suggests that candidates with higher power indices spoke for significantly more time than others (*WordDev*) and they they also got significantly more number of turns to talk (*TurnDev*). This finding is in line with the empirical findings in sociology literature (Ng et al. 1993, Reid and Ng 2000) that powerful people take the floor for more time during conversations. We also obtained moderate positive correlation between questions posed to the candidate and their power index, which suggests that the candidates with higher power indices were asked significantly more

questions by the moderators. So the correlation we obtain for the number of turns could also be due to the powerful candidates being asked more questions. This is a deviation from the expected behavior that the moderators treat each candidate uniformly. However, one could argue that this could also be the reflection of a conscious and predetermined allocation of questions to the candidates by the moderator(s) based on the candidates' poll standings. For the other structural features — *LongestTurn*, *WordsPerTurn*, *WordsPerSentence* — we did not obtain any significant correlation to the power indices of candidates.

**TURN TAKING:** In terms of the turn-taking patterns, there are some very interesting findings. We obtained no significant correlation between how powerful a candidate was and how often he/she spoke out-of-turn (*SpeakingOutOfTurn*). Instead, we found statistically significant positive correlation (although weak) for *SpokenToOutOfTurn*, which means that candidates spoke out of turn significantly more often after/while the candidates with higher power speaks, in effect interrupting them and/or the debate structure. This is counter-intuitive and in contrast with previous findings by (Ng et al. 1995) that those who interrupt are more influential or powerful. We believe that this is a manifestation of the participants pursuing power over each other rather than operating within a static power structure.

**MENTIONS:** We found statistically significant high positive correlation between the power indices of candidates and how often they were referenced/mentioned by others (*MentionPercent*). In other words, as candidates gain more power, they are referenced significantly more by others. However, the distribution of mentions of a candidate across different forms of addressing did not have any correlation with the power indices of the candidate. This suggests that while forms of addressing is found to be correlated with power relations by previous studies (Brown and Ford 1961, Dickey 1997), they are not affected by the short term variations of power as in our domain.

## 11.5 Automatic Power Ranker

In this section, we describe an automatic system we built to rank the participants of the debates based on their power indices. We use a supervised learning approach to solve this problem.

### 11.5.1 Problem Formulation

Given a debate $D$ with a set of participants $C_D = \{X_1, X_2, ...X_n\}$ and their corresponding power indices $P(X_i)$ for $1 < i < n$, we want to find a ranking function $r : C_D \rightarrow \{1...n\}$ such that for all $1 < i, j < n$,

$r(X_i) > r(X_j) \iff P(X_i) > P(X_j)$.

We use an SVM based supervised learning system to estimate the ranking function $r'$ that gives an ordering of participants $\{X'_1, X'_2, ...X'_n\}$, optimizing on the number of inversions between the orderings produced by $r'$ and $r$.

### 11.5.2 Implementation

As discussed in Chapter 4, we use the ClearTk's SVMrank (Joachims 2006) wrapper package to build the ranker. We also used the ClearTk wrapper for the Stanford CoreNLP package to perform basic NLP analysis on the speaker turn texts. The basic steps we performed include: tokenization, sentence segmentation, parts-of-speech tagging, lemmatization and named entity tagging.

### 11.5.3 Evaluation

We report results on 5-fold cross validation. We split our data into folds at the level of debates. That is, we divide our corpus of twenty debates into five folds; four debates in each fold. We report three commonly used evaluation metrics for ranking tasks — Kendall's Tau, NDCG and NDCG@3. NDCG based metrics are more suitable for our purposes since they provides a way to factor in the magnitude of ranking metric (in our case, power index) in the performance assessment. E.g., under NDCG, the penalty for swapping a pair of candidates with $P(X)$ values $35.0$ and $5.0$ will be higher than that for a pair with $P(X)$ values $12.0$ and $15.0$. Tau treats these mistakes equally if the swaps generate the same number of inversions. We describe the calculation of each metric below:

**Kendall's Tau:** This metric measures the similarity between two rankings based on the number of rank inversions (discordant pairings) between original and predicted ranking. The value of Tau varies between $-1$ and $+1$, the higher the value means better the performance. Kendall's Tau is calculated as follows:

$$Tau = (C - D)/(C + D)$$

where C = Concordant pairs; D = Discordant pairs

**Normalized Discounted Cumulative Gain (NDCG):** This metric employs a normalized version of discounted cumulative gain method which penalizes the inversions happening in the top of the ranked list more than those happening in the bottom. Given an ordering of the candidates $\{x'_1, x'_2...x'_n\}$, we first compute $DCG$ as below

$$DCG = P(x'_1) + \sum_{i=2}^{n} \frac{P(x'_i)}{log_2(i)}$$

$$NDCG = \frac{DCG}{IDCG}$$

NDCG normalizes DCG with respect to IDCG, the ideal DCG if the ranking was in perfect order. This allows us to compare the NDCG values across different debates with different number of participants and score values. The NDCG value varies from 0.0 to 1.0, with 1.0 representing the perfect ranking.

**Normalized Discounted Cumulative Gain at 3 (NDCG@3):** This metric is similar to NDCG, but focuses only on the performance of retrieving the top 3 candidates from each debate. NDCG@3 is calculated as follows.

$$DCG_k = P(x'_1) + \sum_{i=2}^{k} \frac{P(x'_i)}{log_2(i)}$$

$$NDCG_k = \frac{DCG_k}{IDCG_k}$$

### 11.5.4 Experiments and Results

In this section, we describe the different experiments we conducted as part of building the automatic power ranking system. The results obtained are presented in Table 11.3.

**Baseline results:** We use a baseline system that uses the same machine learning framework as the rest of the experiments, but uses only the word unigrams as features. In other words, this baseline system does not use any NLP preprocessing. This system obtained a Tau value of 0.17, NDCG of 0.836, and NDCG@3 of 0.686.

**Results using LEXICAL features alone:** We first find the best performing set of lexical features by varying the ngram length from 1 to 5. We found the best setting to be $n = 1$ for word lemma ngrams, $n = 2$ for part-of-speech ngrams, and $n = 5$ for mixed ngrams. We use this setting for the rest of the experiments conducted. The higher value of $n$ for mixed ngrams reaffirms our claim that they help capture longer patterns in text. In fact, mixed ngrams obtained the larger gain over the baseline than word lemma ngrams and part-of-speech ngrams, used separately. The system trained on mixed ngrams with $n = 5$ improved the Tau value to 0.29, NDCG to 0.881 and NDCG@3 to 0.756. The system trained on part-of-speech ngrams also posted similar improvements in NDCG and NDCG@3, but the improvement in Tau was only marginal. Improvement using word lemma ngrams over the baseline was only marginal. The best results were obtained when the combination of all three sets of ngrams were used. That system posted a Tau value of 0.30, NDCG of 0.895 and NDCG@3 of 0.820.

**Results using all features:** The next set of rows in Table 11.3 show the results obtained using all feature sets and the effect of removing each feature set separately. The system trained on all features obtained a significant improvement over LEXICAL alone. It posted a Tau value of 0.47, NDCG of 0.927 and NDCG@3 of 0.853. Removing the *MentionPercent* feature from this set decreased the results significantly to a Tau value of 0.30, NDCG of 0.875 and NDCG@3 of 0.755. This shows the utility of *MentionPercent* for this ranking task. Removing the *TurnDev* feature, on the other hand, improved the results to a Tau value of 0.49, NDCG of 0.942 and NDCG@3 of 0.896, suggesting that *TurnDev* adds noise to the ranker. The impact of removing any of the other features individually from the model affected the results only marginally.

**Best features:** The best performance was obtained using the features *QstnDev* and *MentionPercent*. The system trained using these two features obtain a Tau value of 0.55, NDCG of 0.968 and NDCG@3 of 0.939. Both of these features were very useful and were part of the top five systems measured in terms of NDCG. The next three rows show the best results obtained excluding either or both of these features. The best feature set without using *QstnDev* obtain an NDCG of 0.951 and an NDCG@3 of 0.920, with a marginal reduction in the Tau value. On the other hand, the best feature set without using *MentionPercent* obtain an NDCG of 0.954 and an NDCG@3 of 0.915, but the Tau value was reduced significantly to 0.43. The best result without using either of these features is was

obtained when using *TurnDev* and *SpokenToOutOfTurn*; it posted a Tau of 0.45, NDCG of 0.947 and NDCG@3 of 0.885.

### 11.5.5 Post-hoc Analysis

Although lexical ngram features turned out to be not very useful for the task of power ranking in the domain of political debates, the models trained using only lexical ngram features did perform significantly better than random prediction (e.g., Tau: 0.30 vs. -0.06; NDCG: 0.895 vs. 0.798; NDCG@3: 0.820 vs. 0.609). It would be worthwhile to investigate if there are any interesting patterns in the ngram features that the machine learning system finds. However, it is not feasible to perform a correlation study on each ngram features like we did for structural features in Section 11.4. Instead, we inspect the weights assigned to each of the ngram features in the best-performing linear kernel SVM model trained using all three types of ngrams, as a way to find the ngram features that are associated more with candidates with higher values for $P(X)$ and lower values for $P(X)$.

Table 11.4 lists the top 15 positive and negative weighted features from the best-performing lexical ngram model, along with corresponding weights. POS tags are capitalized and `_BOS_` stands for `beginning_of_sentence`. It is hard to infer strong conclusions based purely on the SVM feature weights, especially since the performance of this model was far from ideal. However, the SVM algorithm does pick up some interesting signals. For example, those with power used `agree` more, suggesting that they might be less contentious than others. `UH_.` which captures interjections/pauses was assigned a positive weight, which aligns with the finding that those with power get interrupted more. Another interesting pattern is in terms of types of verbs the candidates use. Eleven of the top fifteen positive weighted patterns contained a verb (part-of-speech tags starting with VB), out of which seven were in the past tense (VBD) or past participle (VBN) form. In contrast, none of the six patterns in the top negative weighted patterns that contained a verb was in the past tense or past participle form; four of these six were in the base form (VB). In other words, the candidates with higher power probably talk more about things that happened in the past (e.g., their accomplishments etc.) than those with lower power.

| Feature Set | Tau | NDCG | NDCG@3 |
|---|---|---|---|
| Baseline: Random | -0.06 | 0.798 | 0.609 |
| Baseline: Word Unigrams | 0.17 | 0.836 | 0.686 |
| *LemmaNGram*$^{n=1}$, *POSNGram*$^{n=3}$, *MixedNGram*$^{n=5}$ | 0.30 | 0.895 | 0.820 |
| *LemmaNGram* | 0.20 | 0.839 | 0.695 |
| *POSNGram* | 0.18 | 0.854 | 0.734 |
| *MixedNGram* | 0.29 | 0.881 | 0.756 |
| ALL: all features | 0.47 | 0.927 | 0.853 |
| ALL - *LemmaNGram* | 0.47 | 0.930 | 0.854 |
| ALL - *POSNGram* | 0.40 | 0.927 | 0.862 |
| ALL - *MixedNGram* | 0.50 | 0.930 | 0.858 |
| ALL - *WordDev* | 0.46 | 0.926 | 0.853 |
| ALL - *TurnDev* | 0.49 | 0.942 | 0.896 |
| ALL - *QstnDev* | 0.48 | 0.931 | 0.858 |
| ALL - *SpeakingOutOfTurn* | 0.51 | 0.929 | 0.868 |
| ALL - *SpokenToOutOfTurn* | 0.49 | 0.933 | 0.873 |
| ALL - *MentionPercent* | 0.30 | 0.875 | 0.755 |
| BEST: *QstnDev*, *MentionPercent* | 0.55 | 0.968 | 0.939 |
| *WordDev*, *MentionPercent* | 0.53 | 0.951 | 0.920 |
| *WordDev*, *QstnDev*, *SpokenToOutOfTurn* | 0.43 | 0.954 | 0.915 |
| *TurnDev*, *SpokenToOutOfTurn* | 0.45 | 0.947 | 0.885 |

Table 11.3: Automatic power ranker results.

| Positive Weighted Ngrams | | Negative Weighted Ngrams | |
|---|---|---|---|
| MixedNGram:that_it | 0.16 | WordNGram:do | -0.14 |
| MixedNGram:VBN_NN | 0.15 | PosNGram:DT_JJ | -0.12 |
| PosNGram:VBN_NN | 0.15 | PosNGram:VBP_: | -0.12 |
| WordNGram:agree | 0.15 | WordNGram:tell | -0.12 |
| MixedNGram:VBP_that | 0.15 | PosNGram:PRP_IN_PRP | -0.11 |
| PosNGram:VBP_IN_PRP | 0.14 | PosNGram:PRP_IN | -0.11 |
| PosNGram:VBP_IN | 0.14 | PosNGram:VB_PRP_IN | -0.11 |
| MixedNGram:VBD_a | 0.13 | PosNGram:PRP_VBP_: | -0.11 |
| PosNGram:UH | 0.13 | MixedNGram:... | -0.11 |
| MixedNGram:I_VBP_that | 0.13 | WordNGram:... | -0.11 |
| MixedNGram:it_VBD | 0.13 | MixedNGram:VB_I | -0.10 |
| PosNGram:VBD_DT | 0.13 | PosNGram:_BOS__VB | -0.10 |
| PosNGram:PRP_VBD | 0.13 | PosNGram:_BOS__VB_PRP | -0.10 |
| PosNGram:IN_PRP_VBD | 0.13 | MixedNGram:about | -0.10 |
| MixedNGram:NN_NN_. | 0.13 | PosNGram:JJ_NNS | -0.09 |

Table 11.4: Top weighted lexical features: top 15 positive and negative weighted features from the best-performing LEXICAL based model

### 11.5.6 Discussion

One of the main takeaway message from our experiments is that lexical features are not as helpful as structural features for the task of power ranking in political debates. This is in stark contrast with what we saw in the task of predicting power relations in organizational emails (Chapter 7-Chapter 9), where lexical features had great predictive power and structural features served only to improve the results obtained by lexical features; structural features by themselves performed much worse than using lexical features. There could be many reasons that contribute to this difference. We discuss two of them below:

**The differences in both types of power:** In Chapter 7, we study the manifestations of power that is sourced from a static power structure; the organizational hierarchy. In the domain of political debates, the power is more dynamic and changes from debate to debate. It is possible that such

dynamic forms of power do not impact the lexical choices much, resulting in the poor performance of lexical features.

**Discourse intentions of participants:** Hierarchical power affect the discourse intention each participant has in an interaction, especially in workplace interactions, and what the lexical features pick up on is in fact the discourse intentions that are manifested in the words and phrases a person use. For example, the finding that superiors issue significantly more requests for action than subordinates is most likely a manifestation of the discourse intentions they have in a workplace setting; i.e., superiors issuing more requests is not necessarily an artifact of their having power, but an artifact of the job roles they play in interactions with subordinates. Similarly, subordinates sending long emails with lots of information sentences is probably because their job roles are associated with more explaining, compared to superiors. In other words, the lexical features capture many patterns that are associated with the roles superiors and subordinates play in a workplace interaction. In contrast, in the domain of presidential debates, all candidates participating in the debates are playing the same role, with the same objective of convincing their electorate of their candidature. Since they all play the same role in the debates, the lexical features are less predictive.

## 11.6 Conclusion

In this chapter, we presented our analysis of how power of confidence is manifested in the genre of political debates. We used the 2012 Republican party presidential primary election debates for our analysis. We modeled power of confidence a candidate has coming in to a debate based on their most recent poll standings. We analyzed how different structural aspects of the interaction in the debates correlated with the power of confidence each candidate had. We found that the power affected both how a candidate behaved within the debates as well as how others behaved towards them. We then presented an automatic ranking system that rank the participants of a debate in terms of their relative power based on the language used in the debates and the structure of interactions. We found that, unlike the genre of organizational emails, lexical features do not help in the problem of power prediction in the genre of political debates.

# Chapter 12

# Topic Dynamics and Power

Understanding how topics of discussion evolve over the course of an interaction is an important way to model dialog behavior of its participants. Participants make choices in terms of what to talk about, or what to ask others to talk about, as well as whether to stay on topic or to attempt to start a new topic. In this chapter, we study how features that capture the topic dynamics in the presidential debates with respect to each candidate correlate with their power of confidence. We also show how these features could help improve the predictive performance of an automatic system that ranks the candidates in terms of their power.

As a first step in performing an analysis in this aspect, we need to assign topics to each turn of an interaction. One way to do this is to use the traditional topic modeling approaches such as LDA by considering each turn or message as a 'document' and assigning topic to it using the content of that turn. However, such approaches miss out on the important information about the sequence structure of turns; e.g., adjacent turns have higher probability to be of the same topic. Ideally, we would want a system that can reliably identify topical segments (sequence of turns) of the interaction as well as attempts (successful or not) to shift topics. In this chapter, we investigate three methods of assigning topics to the turns — LDA With Substantivity Handling, Speaker Identity for Topic Segmentation, and Variational Speaker Identity for Topic Segmentation.

## 12.1 Related Work

One of the NLP techniques that has gathered wide popularity is topic modeling using Latent Dirichlet Allocation (LDA) (Blei et al. 2003). LDA is a generative model that considers each document to be a mixture of a set of topics and each word in the document is attributed to one of the document's topics. After LDA gained wide acceptance in the research community, a variety of extensions to the basic formulation have been developed. Researchers have built techniques that model topics over time (Blei and Lafferty 2006, Wang and McCallum 2006), incorporate information about the author (Rosen-Zvi et al. 2004), and take into account hierarchical structures of topics (Blei et al. 2010).

There has been previous work that has studied the relation between a participant's power and the topical content of his/her utterances, beyond surface level lexical cues (e.g., (Reid and Ng 2000)). (Reid and Ng 2000) found that interruptions gained in prototypical utterances (i.e., utterances that provide information that defines speakers and listeners within a given social context) are more strongly correlated with the perception of influence than those in non-prototypical utterances. They study this in the context of small group discussions where two groups of opposing opinions debated about a controversial topic. In our domain of presidential debates, however, there is no one topic and two opinions; rather a multitude of topics and different opinions on them.

## 12.2 Topic Distribution and Power

As a first step towards understanding the topic dynamics in the debates, we investigated the distribution of topic across different candidates. Specifically, we look at whether the candidates' power indices have any correlation with the distribution of topics among themselves and within each of their set of turns. The hypothesis is that, those with higher power indices get to talk more on the central topics of the debate than those with lower power indices.

### 12.2.1 Assigning Topics to Turns

For the analysis presented in this section, we used the Topic Modeler in the Mallet (McCallum 2002) package to assign topics to each candidate turn. Mallet uses the Latent Dirchlet Allocation (LDA) (Blei et al. 2003) with Gibbs sampling in order to assign the topic of a given text sample. Under LDA, each speaker turn is considered as a mixture of a small number of topics and each

| Topic | Label | Top Ten Words |
|---|---|---|
| T1 | Space | state, south, space, carolina, national, move, administration, nasa, florida, air |
| T2 | Military | war, military, afghanistan, country, troops, world, pakistan, foreign, people, policy |
| T3 | US Issues | states, united, people, country, make, american, issue, back, state, problem |
| T4 | Energy | tax, job, percent, plan, energy, rate, country, create, economy, put |
| T5 | Election | fact, people, american, year, question, campaign, obama, reagan, bill, thing |
| T6 | Immigration | border, people, immigration, law, illegal, country, illegally, secure, legal, legally |
| T7 | Budget cuts | cut, program, budget, money, spending, security, social, debt, year, government |
| T8 | Banks | government, bank, money, market, housing, freddie, mac, company, people, street |
| T9 | Conservative | conservative, record, run, vote, issue, win, stand, fight, state, obama |
| T10 | Gay Rights | state, law, court, rights, people, issue, life, marriage, constitution, make |
| T11 | Health care | care, state, health, government, people, obamacare, federal, insurance, plan, mandate |
| T12 | Middle east | iran, israel, nuclear, united, weapon, states, world, ally, syria, sanction |
| T13 | Monetary policy | federal, reserve, money, policy, government, china, understand, fed, currency, trade |
| T14 | Economy | people, work, america, job, make, country, time, create, economy, business |
| T15 | Education | child, school, family, kid, education, parent, thing, home, good, language |

Table 12.1: Topics detected across debates (with manually assigned labels)

word in the turn is attributable to one of the turn's topics. LDA is an unsupervised algorithm which is parametrized by the number of topics, $N$. The reliability of the learned model greatly depends on the number of topics chosen. Choosing a large number would cause the learned model to be fragmented and a small number would create a model with incoherent topics. The number of topics are often chosen based on the domain knowledge. We chose $N$ to be 15 as a good estimate for the number of major topics during the 2012 Republican presidential primary campaign period. We selected the best model after running 2000 iterations, based on the maximum posterior probability. Table 12.1 lists the topics identified by our model from across the debates and the top ten words that represented each topic. We manually assign each topic a label (second column) after looking at the top ten words that represented the topic (third column).

### 12.2.2 Analysis

For each candidate $X$ of all the $C_D$ candidates of the debate D, we compute two measures — topic percent across candidates ($TP_{AC}(t_j)$) and across topics ($TP_{AT}(t_j)$) — for each topic $t_j$,

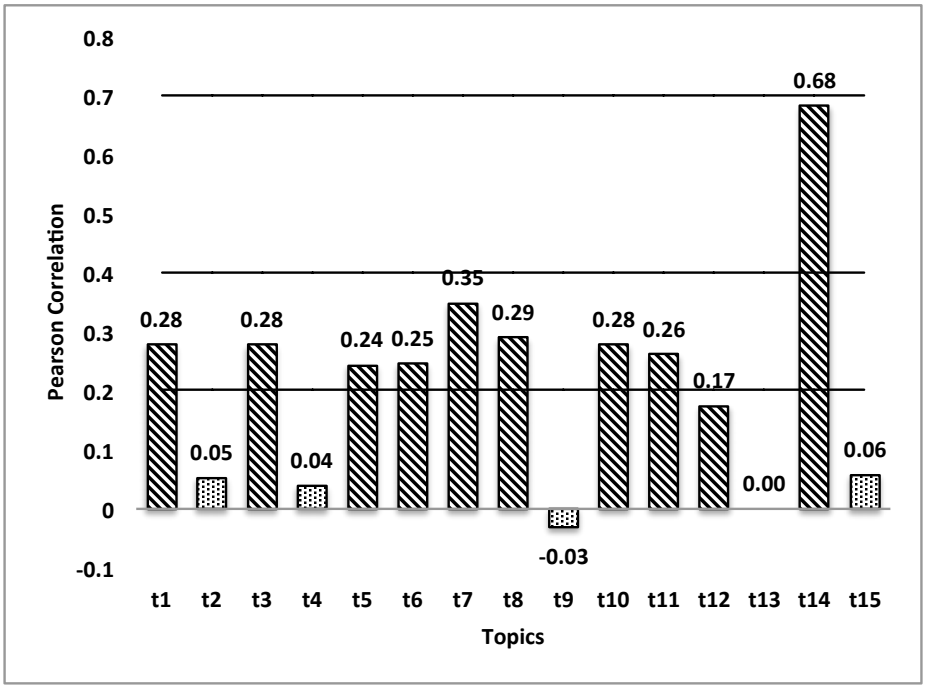



Figure 12.1: Distribution of each topic's turns across candidates ($TP_{AC}$)
p-values: t1, t3, t5, t6, t7, t8, t10, t11, t14 ($p < 0.005$); t12 ($p < 0.05$)

$j \in \{1, ..., N\}$. $TP_{AC}(t)$ measures what percentage of turns with topic $t$ within the debate was spoken by candidate $X$, while $TP_{AT}(t)$ measures what percentage of candidate $X$'s turns within the debate was about topic $t$. For example, $TP_{AC}(Energy)$ captures a candidate's contribution towards the topic *Energy* overall in the debate, where as $TP_{AT}(Energy)$ captures how much of the candidate's contribution in the debate was about the topic *Energy*. More formally, we define $TP_{AC}$ and $TP_{AT}$ as below.

$$TP_{AC}(t_j) = Turns(X, t_j) / \sum_{i=1}^{C_D} Turns(c_i, t_j)$$

$$TP_{AT}(t_j) = Turns(X, t_j) / \sum_{j=1}^{N} Turns(X, t_j)$$

where $Turns(X, t)$ denotes turns of candidate $X$ of topic $t$. For each candidate, we calculate the correlation of $TP_{AC}(t_j)$ and $TP_{AT}(t_j)$ for each topic $t_j$, with his or her power index. The correlation values obtained are listed in Figure 12.1 and Figure 12.2.

As Figure 12.1 shows, we obtained significant positive correlations for the values of topic percentages across candidates, for 10 out of 15 topics ($TP_{AC}(t_j)$). These numbers are, however, biased

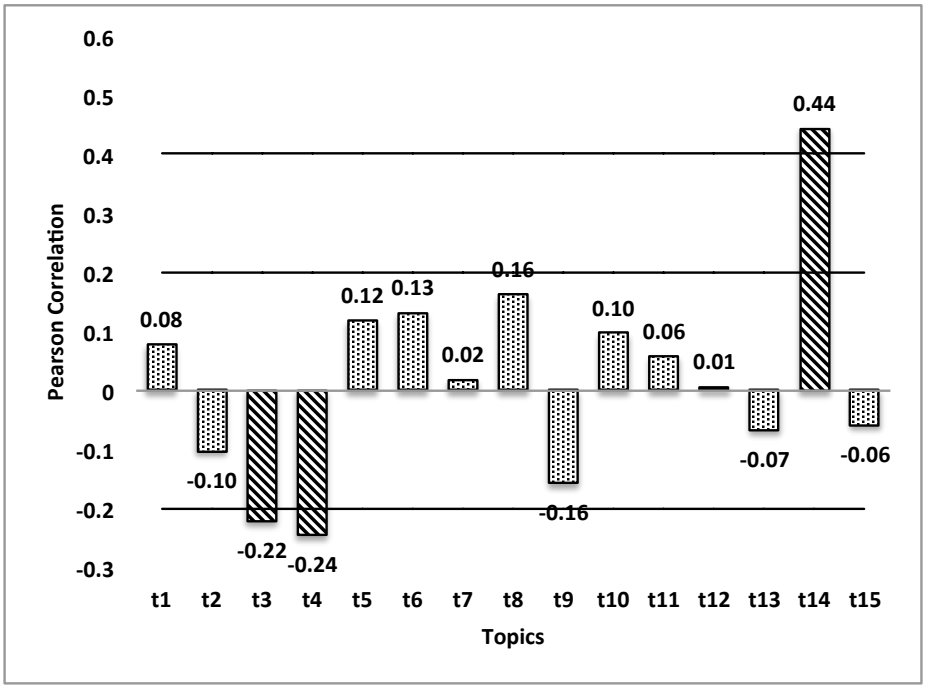



Figure 12.2: Distribution of each candidate's turns across topics ($TP_{AT}$)
p-values: t14 ($p < 0.005$); t3, t4 ($p < 0.05$)

by the fact that candidates with higher power indices gets to talk more than others (Chapter 11, Section 11.4, page 230). Hence, what is more interesting is to notice the topics that did not significantly correlate with power (i.e.,$t_2$ (*US Issues*), $t_4$ (*Election*), $t_9$ (*Judiciary*), $t_{13}$ (*Monetary Policy*), and $t_{15}$ (*Education*)).

When we consider the values across topics ($TP_{AT}(t_j)$), we find high to weak significant correlations for some of the topics (Figure 12.2). We find that the candidates with higher power indices talked significantly more about some topics ($t_{14}$ (*Economy*)) than others, while they talked significantly less about some others ($t_4$ (*Election*) and $t_5$ (*Immigration*)). It is important to note that these numbers are not biased by the disproportionate distribution of turns they got as might be the case with $TP_{AC}(t_j)$. It is also interesting to note that all topics that were not significantly correlated to power in Figure 12.1 had a negative correlation coefficient in Figure 12.2, although only 2 of them were significant negative correlations.

The correlations that we observed between power and some topics are artifacts of the dominant issues at the time of this particular election campaign. However, it is an important finding that power

correlates with the distribution of topics in this domain, which contrasts with (Bales 1970).

## 12.3 Modeling Topic Shifts in Interactions

As discussed in the end of Section 12.2, the correlations obtained between the distribution of topics and power of candidates are artifacts of the dominant issues during the 2012 presidential election cycle. We do not expect these correlations to carry over to other forms of interactions, or even to political debates set in another time. A topical dimension with broader relevance is how topics change during the course of an interaction (e.g., who introduces more topics, who attempts to shift topics etc.). For instance, Nguyen et al. (2013) found that topic shifts within an interaction are correlated with the role a participant plays in it (e.g., being a moderator). They also analyzed US presidential debates, but with the objective of validating a topic segmentation method they proposed earlier (Nguyen et al. 2012) that takes into consideration topic shifting tendencies of individuals. They do not study the topic shifting tendencies among the candidates in relation to their power differences.

In the rest of this chapter, we investigate whether the topic shifts by candidates in the debates correlate with their power indices. The biggest challenge in performing this analysis is to automatically detect topic shifts. We describe this issue in detail in Section 12.3.1, and then describe the different methods we adopt in order to handle this issue in Section 12.4 and Section 12.5.1.

### 12.3.1 Challenges

Let us start by considering an excerpt from the debate held on 06/13/2011 at Goffstown, New Hampshire shown in Table 12.2. The topic of discussion is the issue of gay marriage and the "Don't Ask Don't Tell" policy (prohibiting military personnel from discriminating against or harassing closeted homosexual or bisexual service members or applicants, while barring openly gay, lesbian, or bisexual persons from military service). In this debate KING (John King, CNN Anchor) is the moderator (marked with an M in parentheses). In turn 223, PAWLENTY is talking about his opinion on the definition of marriage. Afterwards, KING passes around questions related to this topic to the other candidates — PAUL, ROMNEY, GINGRICH, SANTORUM, and BACHMANN till turn 233 (the last four turns are omitted from the transcript shown in Table 12.2). In Turn 234, KING asks

| Turn # | Turn Speaker and Text | Substantive? |
|---|---|---|
| 223 | PAWLENTY: I support a constitutional amendment to define marriage between a man and woman. I was the co-author of the state – a law in Minnesota to define it and now we have courts jumping over this. | [**S**] |
| 224 | KING (M): OK. Let's just go through this. | [**NS**] |
| 225 | PAUL: The federal government shouldn't be involved. I wouldn't support an amendment. [...] I don't think government should give us a license to get married. It should be in the church. | [**S**] |
| 226 | KING (M): Governor Romney, constitutional amendment or state decision? | [**NS**] |
| 227 | ROMNEY: Constitutional. | [**NS**] |
| 228 | KING (M): Mr. Speaker? | [**NS**] |
| 229 | GINGRICH: Well, I helped author the Defense of Marriage Act which the Obama administration should be frankly protecting in court. I think if that fails, at that point, you have no choice except to constitutional amendment. | [**S**] |
| | [...] | |
| 234 | KING: All right, let me ask you another question. [...] would you leave that policy in place or would you try to change it, go back to "don't ask/don't tell," or something else? | [**S**] |
| 235 | CAIN: If I had my druthers, I never would have overturned "don't ask/don't tell" in the first place. Now that they have changed it, I wouldn't create a distraction trying to turn it over as president. Our men and women have too many other things to be concerned about rather than have to deal with that as a distraction. | [**S**] |
| | [...] | |
| 240 | KING (M): Leave it in place, what you inherit from the Obama administration or overturn it? | [**S**] |
| 241 | ROMNEY: Well, one, we ought to be talking about the economy and jobs. But given the fact you're insistent, the – the answer is, I believe that "don't ask/don't tell" should have been kept in place until conflict was over. | [**S**] |

Table 12.2: Debate excerpt about marriage equality and the "Don't Ask/Don't Tell" policy. Goffstown, NH. 06/13/11.

[**S**]/ [**NS**] denote substantiveness of turns

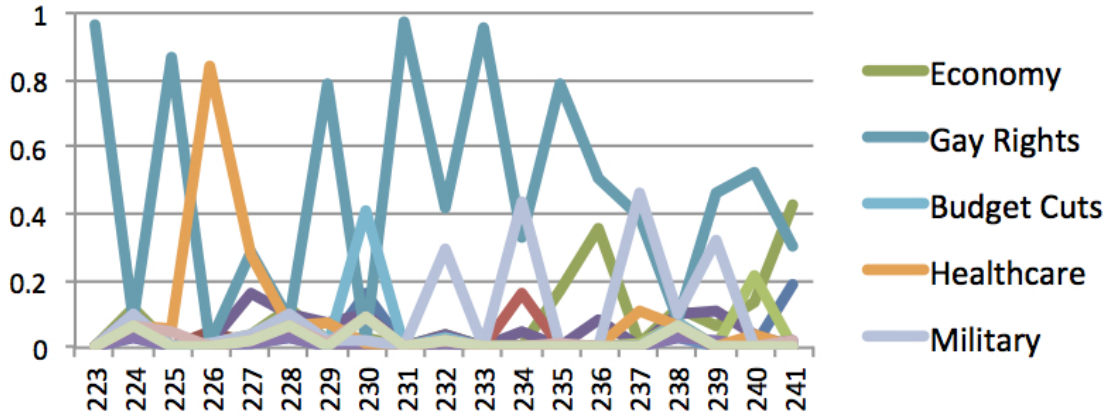



Figure 12.3: Topic probabilities assigned by LDA to debate turns.

about the issue of "Don't Ask Don't Tell" to CAIN to which he responds in turn 235. Turns 236 till 239 contains PAWLENTY's and PAUL's opinions on this issue. KING then directs the question to ROMNEY, to which ROMNEY responds in turn 241 starting with *we ought to be talking about the economy and jobs*. Ideally, we should be able to detect KING shifting the topic at turn 234 and ROMNEY shifting (or attempting to shift) the topic in turn 241.

A naive approach to detecting topic shifts in the debates is to detect instances where a turn's topic differed from the previous turn's topic. Suppose we use the topic assignments from LDA (Section 12.2) where each turn is assigned the topic for which the LDA returned the highest probability. Detecting topic shifts in this manner is problematic, because LDA assumes each turn to be independent and hence fails to take into account the sequential structure of turns that make the interaction. Not all turns by themselves contribute to the conversational topics in an interaction. A large number of turns, especially by the moderator, manage the conversation rather than contribute content to it. These include turns redirecting questions to specific candidates (e.g., turns 224, 226 and 228 in Table 12.2) as well as moderator interruptions (e.g., "Quickly.", "We have to save time"). Furthermore, some other turns address a topic only when considered together with preceding turns, but not when read in isolation. These include turns that are short one-word answers (e.g., turn 227) and turns that are uninterpretable without resolving anaphora (e.g., "That's right"). While these turns are substantive to human readers, topic modeling approaches such as LDA cannot assign them topics correctly because of their terseness.

This issue is shown pictorially in Figure 12.3, which shows the line graph of topic probabilities assigned by LDA to the sequence of turns in Table 12.2. As the graph shows, non-substantive turns are assigned spurious topic probabilities by LDA. For example, turn 224 by KING (*OK. Lets just go through this.*) was assigned small probabilities for all topics; the highest of which was *economy* (probability of 0.12). This error is problematic when modeling topic shifts, since this turn and the next one by PAUL would have been incorrectly identified as shifts in topic from their corresponding previous turns.

In this chapter, we look at two methods to handle this issue. First, we introduce the notion of turn substantivity and use that in conjunction with LDA to obtain more reliable identification of topic shifts. Second, we use the Speaker Identity for Topic Segmentation (SITS) system (Nguyen et al. 2012) to identify topic segments.

## 12.4 LDA With Substantivity Handling

In this section, we describe a method where we introduce a method to automatically detect non-substantive turns (i.e., turns that do not contribute topical content to the conversation) and use that information in detecting topic shifts. We start by defining the notion of topic substantivity.

### 12.4.1 The Notion of Turn Substantivity

We define the turns that do not, in isolation, contribute substantially to the conversational topics as **non-substantive** turns. In order to obtain a gold standard for non-substantivity, we obtained manual annotations for each turn in one entire debate (dated 06/13/11) as either *substantive (S)* or *non-substantive (NS)*. The annotators were instructed not to consider the identity of the speaker or the context of the turn (preceding/following turns) in making their assessment. The exact annotation instructions were as follows:

- The purpose of annotating this dataset is to tag each turn as substantive(S) or non-substantive (NS), depending on whether or not the text covered in it can be assigned a valid, coherent topic. Here "topic" is taken in the common language sense.

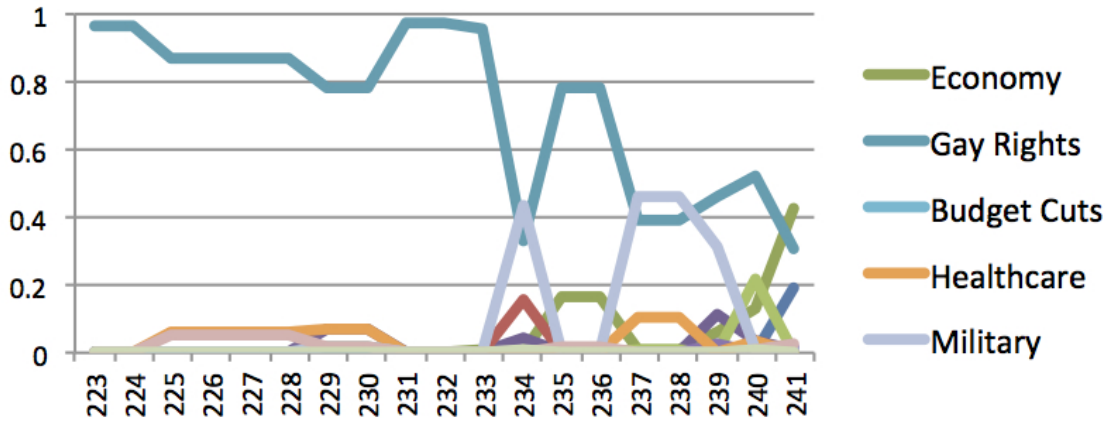



Figure 12.4: Topic probabilities assigned using LDA with non-substantivity handling.

- A turn could possibly be assigned more than one topic. Your task is to only decide whether or not such an assignment can be made or not.
- You must make your decision based solely on the text covered in the turn. The identity of the speaker must not affect your judgment. Furthermore, we take each turn in isolation. So if a turn addresses a topic when read with a preceding turn, but when read in isolation does not address a topic, it should be classified as "NS". This also covers cases in which the turn is not interpretable without resolving an anaphora, as in *That is correct*.

We obtained annotations from two annotators and obtained a high inter-annotator agreement (observed agreement = 89.3%; Kappa ($\kappa$) = .76). We took the assessments by one of the annotators as the gold standard, in which 108 (31.5%) of the 343 turns were identified as *non-substantive*.

We show the *S* vs. *NS* assessments for each turn in column 3 of Table 12.2. If we assume that the non-substantive turns follow the same topic probabilities as the most recent substantive turn, we obtain the line graph shown in Figure 12.4. This topic assignment captures the topic dynamics in the segment more accurately. It identifies *Gay Rights* as the predominant topic until turn 234 followed by a mix of *Gay Rights* and *Military* as topics while discussing the "Don't Ask/Don't Tell' policy. It also captures the attempt by ROMNEY in turn 242 to shift the topic to *Economy*.

| Method | Accuracy (%) | F-measure |
|---|---|---|
| WC_Thresh | 82.6 | 73.7 |
| SD_Thresh | 76.2 | 64.7 |
| WC_Thresh + SD_Thresh | 76.8 | 70.4 |

Table 12.3: Accuracy and F-measure of identifying non-substantive turns.

### 12.4.2 Automatically Identifying Non-substantive Turns

In order to automatically detect non-substantive turns, we investigate a few alternatives. A simple observation is that many of the *NS* turns such as redirections of questions or short responses have only a few words. We tried a word count threshold based method (**WC_Thresh**) where we assign a turn to be *NS* if the number of tokens (words) in the turn is less than a threshold. Another intuition is that for a non-substantive turn, it would be hard for the LDA to assign topics and hence all topics will get almost equal probabilities assigned. In order to capture this, we used a method based on a standard deviation threshold (**SD_Thresh**), where we assign a turn to be *NS* if the standard deviation of that turn's topic probabilities is below a threshold. We also used a combination system where we tag a turn to be *NS* if either system tags it to be. We tuned for the value of the thresholds and the best performances obtained for each case are shown in Table 12.3. We obtained the best results for the WC_Thresh method with a threshold of 28 words, while for SD_Thresh the optimal threshold is .13 (almost twice the mean).

### 12.4.3 Topic Assignments

We first ran the LDA at a turn-level for all debates, keeping the number of topics to be 15, and selected the best model after 2000 iterations. Then, we ran the WC_Thresh method described above to detect *NS* turns. For all *NS* turns, we replace the topic probabilities assigned by LDA with the last substantive turn's topic probabilities. Note that an *S* turn coming after one or more *NS* turns could still be of the same topic as the last *S* turn, i.e., non-substantivity of a turn is agnostic to whether the topic changes after that or not. A topic shift (or attempt) happens only when LDA assigns a different topic to a substantive turn.

### 12.4.4 Topic Dynamics Features

We now describe various features we use to capture the topical dynamics within each debate, with respect to each candidate. When we compute a feature value, we use the topic probabilities assigned to each turn as described in the previous section. For some features we only use the topic with the highest probability, while for some others, we use the probabilities assigned to all topics. We consider features along four dimensions which we describe in detail below.

#### 12.4.4.1 Topic Shift Patterns

We build various features to capture how often a candidate stays on the topic being discussed. We say a candidate attempted to shift the topic in a turn if the topic assigned to that turn differs from the topic of the previous (substantive) turn. We use a feature to count the number of times a candidate attempts to shift topics within a debate (**TS_Attempt#**) and a version of that feature normalized over the total number of turns (**TS_Attempt#$^{\mathbf{N}}$**). We also use a variation of these features which considers only the instances of topic shift attempts by the candidates when responding to a question from the moderator (**TS_AttemptAfterMod#** and **TS_AttemptAfterMod#$^{\mathbf{N}}$**). We also compute a softer notion of topic shift where we measure the average Euclidean distance between topic probabilities of each of the candidate turns and turns prior to them (**EuclideanDist**). This feature in essence captures whether the candidate stayed on topic, even if he/she did not completely switch topics in a turn.

#### 12.4.4.2 Topic Shift Sustenance Patterns

We use a feature to capture the average number of turns for which topic shifts by a candidate was sustained (**TS_SustTurns**). However, as discussed in Section 12.3, the turns vary greatly in terms of length. A more sensible measure is the time period for which a topic shift was sustained. We approximate the time by the number of word tokens and compute the average number of tokens in the turns that topic shifts by a candidate were sustained (**TS_SustTime**).

#### 12.4.4.3 Topic Shift Success Patterns

We define a topic shift to be successful if it was sustained for at least three turns. We compute three features — total number of successful topic shifts by a candidate (**TS_Success#**), that number normalized over the total number of turns by the candidate (**TS_Success#$^{N}$**), and the success rate of candidate's topic shifts (**TS_SuccessRate**)

#### 12.4.4.4 Topic Introduction Patterns

We also looked at cases where a candidate introduces a new topic, i.e., shifts to a topic which is entirely new for the debate. We use the number of topics introduced by a candidate as a feature (**TS_Intro#**). We also use features to capture how important those topics were, measured in terms of the number of turns about those topics in the en tire debate (**TS_IntroImpTurns**) and the time spent on those topics in the entire debate (**TS_IntroImpTime**).

### 12.4.5 Topic Dynamics and Power

We performed a correlation analysis on the features described in the previous section with respect to each candidate against the power he/she had at the time of the debate (based on recent poll scores). Figure 12.5 shows the Pearson's product correlation between each topical feature and candidate's power. Dark bars denote statistically significant ($p < 0.05$) features.

We obtained significant strong positive correlation for TS_Attempt# and TS_AttemptAfterMod#. However, the normalized measure TS_Attempt#$^{N}$ did not have any significant correlation, suggesting that the correlation obtained for TS_Attempt# is mostly due to the fact that candidates with more power have more turns, a finding that is already established by Chapter 11, Section 11.4, page 230. However, interestingly, we obtained a weak, but statistically significant, negative correlation for TS_AttemptAfterMod#$^{N}$ which suggests that more powerful candidates tend to stay on topic when responding to moderators. We did not obtain any correlation for EuclideanDist.

We did not obtain any significant correlations between candidate's power and their topic shift sustenance features. We obtained significant correlation for topic shift success (TS_Success#), modeled based on the sustenance of topic shifts, suggesting that powerful candidates have a higher number of successful topic shifts. However, TS_SuccessRate or TS_Success#$^{N}$ did not obtain any

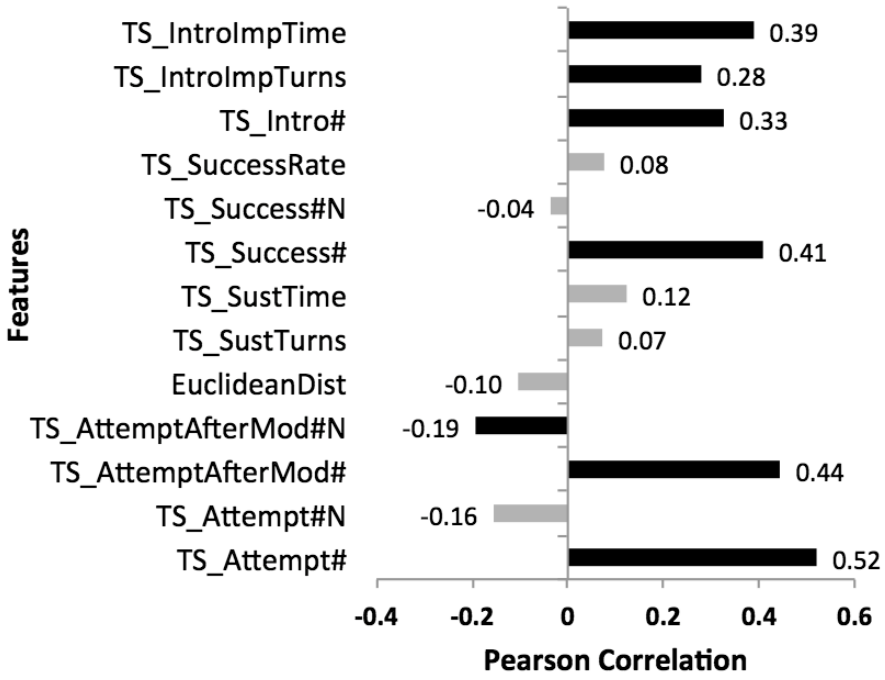



Figure 12.5: Correlations between power index and topic dynamics features.

significant correlation. We also found that powerful candidates are more likely to introduce new topics (TS_Intro#) and that the topics they introduce tend to be important (TS_IntroImpTurns and TS_IntroImpTime).

## 12.5 Segmentation Using SITS-based Approaches

Topic segmentation, the task of segmenting interactions into coherent topic segments, is an important step in analyzing interactions. In addition to its primary purpose, topic segmentation also identifies the speaker turn where the conversation changed from one topic to another, i.e., where the topic shifted, which may shed light on the characteristics of the speaker who changed the topic. We use the SITS approach proposed by (Nguyen et al. 2012) to detect topic shifts. We also propose a different way of using SITS to obtain an analysis of our corpus, which we call SITS$^{var}$. We discuss both in turn, and then provide a discussion.

### 12.5.1 Segmentation Using SITS

Most computational approaches towards automatic topic segmentation have focused mainly on the content of the contribution without taking into account the social aspects or speaker characteristics. Different discourse participants may have different tendencies to introduce or shift topics in interactions. In order to address this shortcoming, Nguyen et al. (2012) proposed a new topic segmentation model called Speaker Identity for Topic Segmentation (SITS), in which they explicitly model the individual's tendency to introduce new topics.

Like traditional topic modeling approaches, the SITS system also considers each turn to be a bag of words generated from a mixture of topics. These topics themselves are multinomial distributions over terms. In order to account for the topic shifts that happen during the course of an interaction, they introduce a binary latent variable $l_{d;t}$ called the topic shift to indicate whether the speaker changed the topic or not in conversation $d$ at turn $t$. To capture the individual speaker's topic shifting tendency, they introduced another latent variable called topic shift tendency ($\pi_x$) of speaker $x$. The $\pi_x$ value represents the propensity of speaker $x$ to perform a topic shift.

### 12.5.2 Segmentation Using SITS$^{\text{var}}$

Within the SITS formulation, the topic shifting tendency of an individual ($\pi_x$) is considered a constant across conversations. While an individual may have an inherent propensity to shift topics or not, we argue that the topic shifting tendency he or she displays can vary based on the social settings in which he or she interacts and his or her status within those settings. In other words, the same discourse participant may behave differently in different social situations and at different points in time. This is especially relevant in the context of our dataset, where the debates happen over a period of 10 months, and the power and status of each candidate in the election campaign vary greatly within that time period.

We propose a variant of SITS which takes this issue into account. We consider each candidate to have a different "persona" in each debate. To accomplish this, we create new identities for each candidate $x$ for each debate $d$, denoted by $x\_d$. For example, 'ROMNEY_08-11-2011' denotes the persona of the candidate ROMNEY in the debate held on 08-11-2011. Running the SITS system using this formulation, we obtain different $\pi_{x\_d}$ values for candidate $x$ for different debates, capturing different topic shift tendencies of $x$.

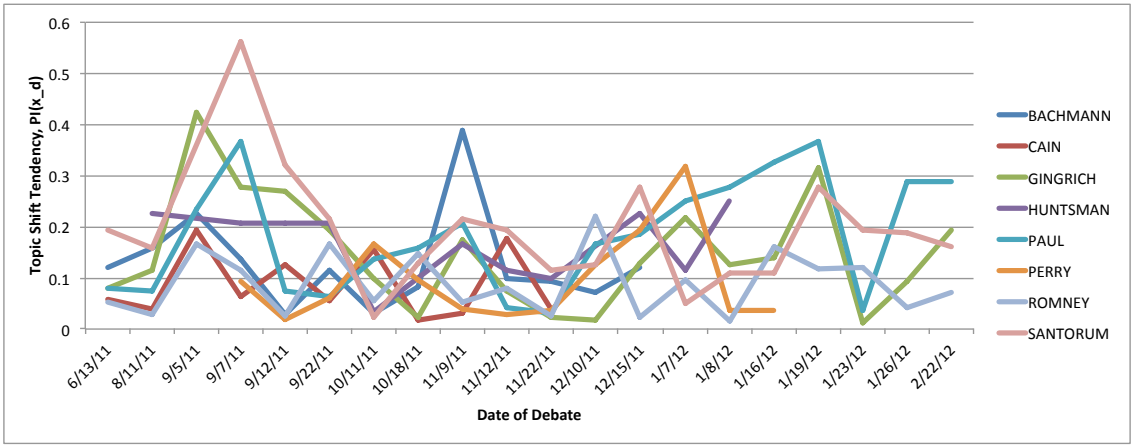



Figure 12.6: SITS$^{var}$ topic shift tendency values of candidates across debates.

### 12.5.3 Execution

We perform both the SITS and SITS$^{var}$ analyses on the 20 debates in our corpus. We used the non-parametric version of SITS for both runs, since it systemically estimates the number of topics in the data. We set the maximum number of iterations at 5000, sample lag at 100 and initial number of topics at 25. We refer the reader to (Nguyen et al. 2013) for details on these parameters.

For each candidate, we calculate the mean and standard deviation of the topic shift tendency ($\pi_{x\_d}$) of his or her personas across all debates he or she participated in. We then average these means and standard deviations, and obtain an average mean of 0.14 and an average standard deviation of 0.09. This shows that the topic shift tendencies of candidates vary by a considerable amount across debates. Figure 12.6 shows the $\pi_{x\_d}$ value fluctuating across different debates.

### 12.5.4 Topic Shifting Tendency and Power

Nguyen et al. (2013) used the SITS analysis as a means to model influence in multi party conversations. They propose two features to detect influencers: Total Topic Shifts (TTS) and Weighted Topic Shifts (WTS). TTS($x, d$) captures the expected number of topic shifts the individual $x$ makes in conversation $d$. This expectation is calculated through the empirical average of samples from the Gibbs sampler, after a burn-in period. We refer the reader to (Nguyen et al. 2013) for more details on how this value is computed. WTS($x, d$) is the value of TTS($x, d$) weighted by $1 - \pi_x$. The intuition here is that a topic shift by a speaker with low topic shift tendency must be weighted

| Feature Set | Feature | Correlation |
|---|---|---|
| TopSh | Total Topic Shifts (TTS) | 0.12 |
| | Weighted Topic Shifts (WTS) | 0.16 |
| TopSh$^{var}$ | Total Topic Shifts (TTS$^{var}$) | 0.12 |
| | Weighted Topic Shifts (WTS$^{var}$) | 0.15 |
| | Topic Shift Tendency (PI$^{var}$) | **-0.27** |

Table 12.4: Correlations between power index and SITS-based topic shift features

**boldface** denotes statistical significance ($p < 0.05$)

higher than that by a speaker with a high topic shift tendency. We use these two features as well, and denote the set of these two features as TopSh.

We also extract the TTS and WTS features using our SITS$^{var}$ variation of topic segmentation analysis and denote them as TTS$^{var}$ and WTS$^{var}$ respectively. In addition, we also use a feature PI$^{var}(x, d)$ which is the $\pi_{x\_d}$ value obtained by the SITS$^{var}$ for candidate $x$ in debate $d$. It captures the topic shifting tendency of candidate $x$ in debate $d$. (We do not include the SITS $\pi_x$ value in our correlation analysis since it is constant across debates.) We denote the set of these three features obtained from the SITS$^{var}$ run as TopSh$^{var}$.

Table 12.4 shows the Pearson's product correlation between each topical feature and candidate's power. We obtain a highly significant ($p = 0.002$) negative correlation between topic shift tendency of a candidate (PI) and his/her power. In other words, the variation in the topic shifting tendencies is significantly correlated with the candidates' recent poll standings. Candidates who are higher up in the polls tend to stay on topic while the candidates with less power attempt to shift topics more often. This is in line with our previous findings from Section 12.4.5 that candidates with higher power attempt to shift topics less often than others when responding to moderators. It is also in line with the findings from Chapter 11 (Section 11.4, page 230) that candidates with higher power tend not to interrupt others. On the other hand, we did not obtain any significant correlation for the features proposed by Nguyen et al. (2013).

## 12.6 Utility of Topic Shifts in Power Ranking

In this section, we investigate the the utility of topic shift features in the problem of automatically ranking the participants of debates based on their power (Chapter 11, Section 11.5). We repeat the formulation here: given a debate $d$ with a set of participants $C_d = \{x_1, x_2, ...x_n\}$ and corresponding power indices $P(x_i)$ for $1 < i < n$, find a ranking function $r : C_d \rightarrow \{1...n\}$ such that for all $1 < i, j < n$, $r(x_i) > r(x_j) \iff P(x_i) > P(x_j)$. We use the same machine learning framework presented there. Our baseline system (BASELINE) uses three features: *WordDev*, *QstnDev* and *MentionPercent* described in Section 11.5. Table 12.5 shows the results obtained using the baseline features (BASELINE) as well as combinations of TopSh and TopSh$^{var}$ features. The baseline system obtained a Kendall Tau of 0.55, NDCG of 0.962 and NDCG@3 of 0.932. The topic shift features by themselves performed much worse, with TopSh$^{var}$ posting marginally better results than TopSh. Combining the topic shift and baseline features increases performance considerably. TopSh$^{var}$ obtained better performance than TopSh across the board. BASELINE + TopSh$^{var}$ posted the overall best system obtaining a Tau of 0.60, NDCG of 0.970, and NDCG@3 of 0.937. These results demonstrates the utility of topic shift features in the power ranking problem, especially using the SITS$^{var}$ approach.

| | **Kendall's Tau** | **NDCG** | **NDCG@3** |
|---|---|---|---|
| BASELINE | 0.55 | 0.962 | 0.932 |
| TopSh | 0.36 | 0.907 | 0.830 |
| TopSh$^{var}$ | 0.39 | 0.919 | 0.847 |
| BASELINE + TopSh | 0.59 | 0.967 | 0.929 |
| BASELINE + TopSh$^{var}$ | **0.60** | **0.970** | **0.937** |
| BASELINE + TopSh + TopSh$^{var}$ | 0.59 | 0.968 | 0.934 |

Table 12.5: Power ranker results using topic shift features on 5-fold cross validation.
BASELINE: Baseline system using *WordDev*, *QstnDev* and *MentionPercent*
NDCG: Normalized Discounted Cumulative Gain

## 12.7 Conclusion

In this chapter, we studied how topic dynamics in the presidential debates correlated with participants' power of confidence, modeled after their recent standings in public polls. We first analyzed the distribution of topics across a candidates turns. We found that candidates with higher power talked more often about certain topics and less often about certain other topics than those with lower power. We then analyzed how topic shifting patterns correlated with power. We investigated two different ways of detecting topic shifts in conversations — LDA with substantivity handling, and a speaker identity based topic segmentation system. We found that overall, people with higher power tend to stay on topic more often that those with lower power. We also showed that topic shift based features can be used to significantly improve the predictive performance of the automatic ranking system we presented in in Chapter 11.

# Part IV

# CONCLUSIONS

# Chapter 13

# Conclusions and Future Work

In this thesis, we presented an extensive data-oriented study of how social power relations are manifested in different linguistic and structural aspects of interactions. We performed this study on two different genres: organizational emails, which contains task oriented written interactions, and political debates, which contains discursive spoken interactions. We showed that power is manifested in both the language and structure of social interactions, and that we can use these linguistic and structural manifestations to automatically infer the underlying power relations. We further investigated whether a person's gender and the gender makeup of an interaction affect the manifestations of his/her power (or lack of it) and found that gender of an interactant and of his/her interaction environment, both affect the manifestations of power. We also studied how different types of power manifest differently in interactions and showed that they exhibit different but predictable patterns.

In this chapter, we first summarize the major findings from the study of manifestations of power presented in this thesis in Section 13.1 and then describe the major contributions of this thesis in Section 13.2. In Section 13.3, we discuss the major limitations of the work presented in this thesis. In Section 13.4, we describe the future directions in which this research can be taken.

## 13.1 Summary of Findings

In our study on the genre of organizational email, we found that power is manifested in the language as well as in the dialog structure of interactions. We showed that superiors and subordinates have significantly different values for their thread structure features as well as dialog act based features.

Superiors send significantly more messages than subordinates, but their messages are significantly shorter. Superiors also have significantly more recipients in their emails. In terms of dialog act features, we found that superiors issue almost twice as many requests for action as subordinates, whereas subordinates' contribute significantly more information in the conversation. When we includes the gender of the participants into the analysis, we further understand these manifestations. We found that gender and gender environment affect the ways power is manifested in interactions in complex ways, resulting in patterns in the discourse that reveal the underlying factors. For example, although superiors use significantly more overt displays of power than subordinates, female superiors use the least overt displays of power. We also studied the level of beliefs expressed by interactants and found that superiors use significantly fewer non committed beliefs, and significantly more non-beliefs. We also found that different types of power are manifested differently in the interactions, with respect to the linguistic and structural features we used.

We presented an automatic system that can predict the direction of power between pairs of people based on single threads of email interactions, that we have made publicly available via a Browser plugin. We found that while lexical features have great predictive power for distinguishing between superiors and subordinates, adding structural features improves the performance significantly. In addition, models that are trained also using features capturing the gender of participants further improved the prediction performance. Similarly, adding the belief information to the lexical features also reported significant improvements on the accuracy of our system.

In the study we performed on the genre of political debates, we analyzed the 2012 Republican party presidential primary election debates. We modeled the power of confidence a candidate has coming in to a debate based on their most recent poll standings and analyzed how different structural aspects of the interaction in the debates correlated with the power of confidence each candidate had. We found that the power affected both how a candidate behaved within the debates as well as how others behaved towards them. Powerful people spoke more, were asked more questions and were interrupted more. In other words, power affected how candidates behaved within the debates as well as how others behaved towards them. We also presented an automatic ranking system that can rank the participants of a debate in terms of their relative power based on the language used in the debates and the structure of interactions.

Unlike the genre of organizational emails, we found that lexical features do not help in the

problem of inferring power relations in the genre of political debates. It shows that in a setting like organizational email where power differences might affect the discourse intentions of participants of an interaction, lexical features will help greatly to infer power relations, whereas in settings such as our presidential debates where the discourse intentions of all participants are the same, the lexical features perform poorly.

## 13.2 Summary of Created Resources

The contributions of this thesis go beyond the study of power and its manifestations that we summarized in Section 13.1. We also built different datasets and computational systems that have relevance to NLP problems beyond the research questions we asked in this particular thesis (for example, the Gender Identified Enron Corpus and the committed belief tagger). We summarize these contributions below.

- **Overt Display of Power Annotations:** We built a corpus of 1734 sentences from the Enron email corpus collection that are annotated with instances of overt display of power. The annotated corpora has been made publicly available.

- **Power Types Annotations:** We also built a corpus of 122 email threads annotated with different types of power relations between participants. The annotations capture instances of situational power, influence, power over communication, as well as perceived hierarchical power. These annotations are also made publicly available.

- **Gender Identified Enron Corpus:** We released an extension to the Enron email corpus in which we have assigned the gender of authors of 87% of the emails. The Gender Identified Enron Corpus has been made publicly available.

- **Topical Non-substantivity Annotations:** We obtained annotations for topical non-substantivity of speaker turns in one of the presidential debates.

- **Minority Preference Multi-class SVM:** We introduced two new methods for SVM multi-class classification that improves the performance on minority class prediction — Divide and Conquer (DAC) and Cascaded Minority Preference (CMP). These approaches have already

been applied to other problems by other researchers obtaining significant improvements (e.g., (Hou et al. 2013)).

- **An Improved Dialog Act Tagger:** We built a dialog act tagger with around 23% error reduction in minority class prediction performance and an overall 10% accuracy error reduction. We have made this dialog act tagger publicly available.
- **An Overt Display of Power Tagger:** We built an automatic tagger to detect instances of overt displays of power in interactions. This system is also made publicly available
- **A New Committed Belief Tagger:** We built a committed belief tagger as part of this thesis. This tagger has since generated great research interest in applying it to other NLP tasks such as knowledge base population and sentiment/opinion analysis.

## 13.3 Limitations

In this section, we will discuss the major limitations of the study presented in this thesis.

### 13.3.1 Scope of Overall Findings

The study presented in this thesis is performed on the Enron email corpus for the organizational email genre and the 2012 Republican Party presidential debates for the political debates genre. However, it remains an open question whether the findings from this study will carry over to other corpora in the same genre. For example, further research is needed to verify whether the conclusions drawn about the manifestations of power in organizational email in our study will hold true in workplaces that are very different from Enron, such as a non-profit organization, an academic institution, or a corporate workplace in a different country/culture. This is especially important since Enron email corpus represents a work environment known for its cutthroat competition (Carroll et al. 2012). Similarly, in the genre of political debates, further research is needed to verify if our findings will hold in the Democratic Party presidential debates, or in political debates held in other election cycles or in other countries.

### 13.3.2 Scope of the Study of Overt Display of Power

One of the major contributions of this thesis is the notion of overt display of power, along with its annotations and the automatic tagger trained using those annotations. We use this as one of the dimensions to study the manifestations of power in this thesis — automatically obtained labels in Chapters 7-9 and manual annotations in Chapter 10. However, it can be argued that the linguistic strategies of display of power might differ greatly even within workplace emails, depending on the cultural background of the participants. Note that these cultural differences of exercise of power exist within the same language; i.e., the social meaning of language differs depending on where it is used. For example, the linguistic expressions (in English) that denote overt displays of power in an American corporate environment (that we capture) may be different from those in a British or Asian corporate environment. It is also the case that the perception of what is an overt display of power is partly subjective. What appears to be an overt display of power to an American reader/observer might differ from that of someone who's familiar only with a British or Asian corporate environment.

While the findings from our study with regard to overt displays of power still hold true since our manual annotator was also an American individual who has had some corporate experience, it is not clear how well the notion as is captured by our annotations will transfer to workplace interactions in other cultures. Our study of overt display of power could benefit from obtaining more annotations on email interactions from other organizations, and by extending it to other genres of interactions such as online discussion forums.

## 13.4 Future Directions

There are many future directions in which to take further the research presented in this thesis. We summarize some of the major directions below.

### 13.4.1 Improving and Extending Analysis and Systems

Using the analysis framework we built, we can extend our study of organizational emails and political debates in many ways. We list some of these lines of future research below:

- Two of the prominent ways we model the dialog structure of email interactions are using our dialog act tagger and overt display of power tagger. Although we improved the minority class performance of these taggers significantly as part of the research presented in this thesis, there is still room for improvement. Currently, the dialog act tagger we use has an accuracy of 92.2%, however, with a request for action F-measure of only 54.2%. Similarly, the overt display of power tagger reports an F-measure of 54.4%. We plan to incorporate more deep syntactic features as well as other feature representation paradigms to improve the performance of these taggers.

- Our power prediction system presented in Chapter 7 and extended in Chapters 8 and 9 predicts the direction of power between pairs of participants, however it does not handle the cases where there exists no power relation. In our preliminary analysis, we found this to be a very hard problem, obtaining F-measures in the range of 20%-30% for making the 3-way distinction. We plan to perform further research in this direction.

- In this thesis, we chose the genre of political debates to investigate the manifestations of power in topic shifting patterns, and the genre of organizational emails to study the manifestations of power in expressions of beliefs. In future work, we will study the topic shift patterns in organizational emails and expressions of beliefs in political debates.

- In the work presented in this thesis, we analyzed the manifestations of gender only in connection with power. However, the dataset that we built can be used to perform more in-depth gender studies (for example, do men behave differently than women in female environments?).

- One very popular tool that has been used in much computational sociolinguistics work in recent years is the linguistic inquiry and word count (LIWC) tool by (Pennebaker et al. 2001). In our preliminary analysis of using LIWC-based features in the genre of political debates, we did not obtain any significant correlations. We plan to investigate this further, both in the organizational emails and political debates.

### 13.4.2 Applying New Methodologies

Another way to take the work presented in this thesis further is by applying new methods of analysis. Specifically, we have identified two research directions — deep learning techniques and social

network analysis. Deep learning methods have shown great promise in many language processing tasks recently. They provide ways to represent the dialog and social context directly in the learning framework rather than by engineering features. Another way to extend the power analysis framework presented in this thesis is through merging the micro analysis of language in the interactions with the macro analysis of social networks formed by those interactions. There are two ways this could be implemented. First, one could enrich the social network created by the interactions in a community by incorporating insights from the micro analysis of individual interactions that our power analysis framework provides, and then apply network algorithms. Second, our power analysis framework itself can gain from incorporating information produced by the social network analysis of the community.

### 13.4.3 Exploring New Corpora, Domains and Genres

In this thesis, we studied two genres of interactions — organizational emails and political debates — using two specific corpora of interactions. As discussed in Section 13.3.1, it is not clear whether the findings from our study carries over to other corpora, domains and genres. One way to extend this work further is to apply it to other corpora in the same genre. For the organizational email genre, one could apply the analysis to the newly released Avocado dataset (Oard 2015), which contain emails from a defunct information technology company. For the political debates genre, once could apply the analysis to debates from other election cycles, transcripts of which are also being maintained by The American Presidency Project.[1]

Another way to extend the work presented in this thesis is by applying the analysis framework to other genres to study how the manifestations of power differ across different genres. We have done preliminary work in investigating the applicability of this line of work on Wikipedia discussion forums. Like Enron, it is also a genre of task-oriented written interactions, however, in an online setting. It also has a hierarchy in which some editors are promoted to the administrator role after extensive review process by the community. We are also planning to apply our analysis framework to analyze Reddit discussion forums to study how status differences between interactants in terms of their "karma points" are manifested in the language and structure of interactions.

[1] americanpresidency.org

### 13.4.4 Investigating Practical Applications

A fourth direction of future research is in applying the analysis presented in this thesis to practical applications in specific domains. We have identified three domains — information retrieval, online education, and business marketing. One could investigate whether revealing the power dynamics between participants in stored interactions in online forums and communities will be helpful to determine relevance for a user with information needs. For example, do users in knowledge sharing forums wants to limit their search to posts by authors with higher power? In business marketing, we will explore how power analysis can benefit the reach of advertisements in an online community. There is also potential for gain in identifying opinion leaders from online forums as a way to improve market research insights.

# Part V

# BIBLIOGRAPHY

# Bibliography

Apoorv Agarwal, Adinoyi Omuya, Aaron Harnly, and Owen Rambow. 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 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P12-2032`.

Apoorv Agarwal, Adinoyi Omuya, Jingwei Zhang, and Owen Rambow. Enron corporation: You're the boss if people get mentioned to you. In *Proceedings of the 2014 International Conference on Social Computing*, SocialCom '14, pages 2:1–2:4, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2888-3. doi: 10.1145/2639968.2640065. URL `http://doi.acm.org/10.1145/2639968.2640065`.

Apoorv Agarwal, Jiehan Zheng, Shruti Kamath, Sriramkumar Balasubramanian, and Shirin Ann Dey. Key female characters in film have more to talk about besides men: Automating the bechdel test. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 830–840, Denver, Colorado, May–June 2015. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N15-1084`.

Jalal S Alowibdi, Ugo Buy, Paul Yu, et al. Language independent gender classification on twitter. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 739–743. IEEE, 2013.

John Langshaw Austin. *How to Do Things with Words*. Harvard University Press, Cambridge, Mass., 1975.

Kathryn Baker, Michael Bloodgood, Bonnie J. Dorr, Nathaniel W. Filardo, Lori S. Levin, and Christine D. Piatko. A modality lexicon and its use in automatic tagging. In *LREC*, 2010.

Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. ACM, 2011.

Robert F. Bales. *Personality and interpersonal behavior*. Holt, Rinehart, and Winston (New York), 1970.

Robert F Bales, Fred L Strodtbeck, Theodore M Mills, and Mary E Roseborough. Channels of communication in small groups. *American Sociological Review*, pages 16(4), 461–468, 1951.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender in twitter: Styles, stances, and social networks. *CoRR*, abs/1210.4567, 2012. URL `http://dblp.uni-trier.de/db/journals/corr/corr1210.html#abs-1210-4567`.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014. ISSN 1467-9841. doi: 10.1111/josl.12080. URL `http://dx.doi.org/10.1111/josl.12080`.

Srinivas Bangalore and Aravind Joshi. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–266, 1999.

Srinivas Bangalore, Pierre Boullier, Alexis Nasr, Owen Rambow, and Benoît Sagot. MICA: A probabilistic dependency parser based on tree insertion grammars. In *NAACL HLT 2009 (Short Papers)*, 2009.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

Paul N. Bennett and Jaime Carbonell. Detecting action-items in e-mail. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information*

*Retrieval*, SIGIR '05, pages 585–586, New York, NY, USA, 2005. ACM. ISBN 1-59593-034-5. doi: 10.1145/1076034.1076140. URL http://doi.acm.org/10.1145/1076034.1076140.

Paul N. Bennett and Jaime G. Carbonell. Combining probability-based rankers for action-item detection. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 324–331, Rochester, New York, April 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N/N07/N07-1041.

Robert Bierstedt. An Analysis of Social Power. *American Sociological Review*, 1950. URL http://dx.doi.org/10.2307/2089716.

Or Biran, Sara Rosenthal, Jacob Andreas, Kathleen McKeown, and Owen Rambow. Detecting influencers in written online conversations. In *Proceedings of the Second Workshop on Language in Social Media*, pages 37–45, Montréal, Canada, June 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W12-2105.

David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *The Journal of machine Learning research*, 3:993–1022, 2003.

David M Blei, Thomas L Griffiths, and Michael I Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):7, 2010.

S Kathryn Boe. Language as an expression of caring in women. *Anthropological linguistics*, pages 271–285, 1987.

Pierre Bourdieu and John B Thompson. *Language and symbolic power*. Harvard University Press, 1991.

Derek Bousfield. lmpoliteness in the struggle for power. *Impoliteness in language: Studies on its interplay with power in theory and practice*, 21:127, 2008.

Derek Bousfield and Miriam A Locher. *Impoliteness in language: Studies on its interplay with power in theory and practice*, volume 21. Walter de Gruyter, 2008.

David B. Bracewell, Marc Tomlinson, and Hui Wang. A motif approach for identifying pursuits of power in social discourse. In *ICSC*, pages 1–8. IEEE Computer Society, 2012.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 773–782, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1078`.

Michael Bratman. Intention, plans, and practical reason. 1987.

Mark E Brooke and Sik Hung Ng. Language and social influence in small conversational groups. *Journal of Language and Social Psychology*, pages 5(3), 201–210, 1986.

Penelope Brown and Stephen C. Levinson. *Politeness : Some Universals in Language Usage (Studies in Interactional Sociolinguistics)*. Cambridge University Press, February 1987. ISBN 0521313554. URL `http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0521313554`.

Roger Brown and Marguerite Ford. Address in american english. *The Journal of Abnormal and Social Psychology*, 62(2):375, 1961.

Roger Brown and Albert Gilman. The pronouns of power and solidarity. 1960.

Harry Bunt. Dialogue pragmatics and context specification. In Harry Bunt and William J. Black, editors, *Abduction, Belief and Context in Dialogue*, pages 81–150. 2000.

Elena Cabrio and Serena Villata. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P12-2041`.

Claire Cardie and John Wilkerson. Text annotation for political science research. 2008.

Archie B Carroll, Kenneth J Lipartito, James E Post, and Patricia H Werhane. *Corporate Responsibility: The American Experience*. Cambridge University Press, 2012.

Vitor Carvalho and William Cohen. Improving "email speech acts" analysis via n-gram selection. In *Proceedings of the Analyzing Conversations in Text and Speech*, pages 35–41, New York City, New York, June 2006. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W06/W06-3406`.

Vitor R. Carvalho and William W. Cohen. Learning to extract signature and reply lines from email. In *IN PROCEEDINGS OF THE CONFERENCE ON EMAIL AND ANTI-SPAM*, 2004.

Anurat Chapanond, Mukkai S Krishnamoorthy, and Bülent Yener. Graph theoretic and spectral analysis of Enron email data. *Computational & Mathematical Organization Theory*, 11(3):265–281, 2005.

Gal Chechik, Geremy Heitz, Gal Elidan, Pieter Abbeel, and Daphne Koller. Max-margin classification of data with absent features. *The Journal of Machine Learning Research*, 9:1–21, 2008.

Na Cheng, R. Chandramouli, and K. P. Subbalakshmi. Author gender identification from text. *Digit. Investig.*, 8(1):78–88, July 2011. ISSN 1742-2876. doi: 10.1016/j.diin.2011.04.002. URL `http://dx.doi.org/10.1016/j.diin.2011.04.002`.

Herbert H. Clark. *Using Language*. cup, Cambridge, England, 1996.

Jennifer Coates. *Language and Gender: A Reader*. Wiley-blackwell, 1998.

Jennifer Coates. *Women, Men and Everyday Talk*. Palgrave Macmillan, 2013. ISBN 9781137314949. URL `https://books.google.com/books?id=ed3QAQAAQBAJ`.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. Learning to Classify Email into "Speech Acts" . In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 309–316, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Mark G. Core and James F. Allen. Coding dialogs with the damsl annotation scheme. In *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Cambridge, MA, November 1997. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.50.7024.

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

Germán Creamer, Ryan Rowe, Shlomo Hershkop, and Salvatore J. Stolfo. Segmentation and automated social hierarchy detection through email network analysis. In Haizheng Zhang, Myra Spiliopoulou, Bamshad Mobasher, C. Lee Giles, Andrew Mccallum, Olfa Nasraoui, Jaideep Srivastava, and John Yen, editors, *Advances in Web Mining and Web Usage Analysis*, pages 40–58. Springer-Verlag, Berlin, Heidelberg, 2009. ISBN 978-3-642-00527-5. doi: 10.1007/978-3-642-00528-2_3. URL http://dx.doi.org/10.1007/978-3-642-00528-2_3.

Jonathan Culpeper. Towards an anatomy of impoliteness. *Journal of Pragmatics*, 25(3):349–367, 1996. ISSN 0378-2166. doi: 10.1016/0378-2166(95)00014-3.

Jonathan Culpeper. Reflections on impolitness, relational work and power. *Impoliteness in Language. Studies on its Interplay with Power in Theory and Practice [Language, Power and Social Process 21], Mouton de Gruyter, Berlin*, pages 17–44, 2008.

Robert A. Dahl. The concept of power. *Syst. Res.*, 2(3):201–215, 1957. doi: 10.1002/bs.3830020303. URL http://dx.doi.org/10.1002/bs.3830020303.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 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, 2012. ACM. ISBN 978-1-4503-1229-5. doi: 10.1145/2187836.2187931. URL http://doi.acm.org/10.1145/2187836.2187931.

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

Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. Committed Belief Annotation and Tagging. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 68–73, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W09-3012`.

Eleanor Dickey. Forms of address and terms of reference. *Journal of linguistics*, 33(02):255–274, 1997.

Jana Diesner and Kathleen M. Carley. Exploration of communication networks from the Enron email corpus. In *In Proc. of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005*, pages 21–23, 2005.

Penelope Eckert and Sally McConnell-Ginet. *Language and Gender*. Cambridge University Press, 2003.

Richard M. Emerson. Power-Dependence Relations. *American Sociological Review*, 27(1):31–41, 1962. ISSN 00031224. doi: 10.2307/2089716. URL `http://dx.doi.org/10.2307/2089716`.

Norman Fairclough. *Language and power*. Pearson Education, 2001.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. The conll-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 1–12, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W10-3001`.

David Ferrucci and Adam Lally. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348, 2004.

Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. Behind the article: Recognizing dialog acts in wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786, Avignon, France, April 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/E12-1079`.

John R. French and Bertram Raven. The Bases of Social Power. In Dorwin Cartwright, editor, *Studies in Social Power*, pages 150–167+. University of Michigan Press, 1959.

Pedro J García-Laencina, José-Luis Sancho-Gómez, and Aníbal R Figueiras-Vidal. Pattern classification with missing data: a review. *Neural Computing and Applications*, 19(2):263–282, 2010.

Sean Gerrish and David Blei. Predicting legislative roll calls from text. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning*, ICML '11, pages 489–496, New York, NY, USA, June 2011. ACM. ISBN 978-1-4503-0619-5.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W14-2106`.

Eric Gilbert. 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, 2012. ACM. ISBN 978-1-4503-1086-4. doi: 10.1145/2145204.2145359. URL `http://doi.acm.org/10.1145/2145204.2145359`.

Carol Gilligan. *In a Different Voice*. Harvard University Press, 1982.

Marco Guerini, Carlo Strapparava, and Oliviero Stock. Corps: A corpus of tagged political speeches for persuasive communication processing. *Journal of Information Technology & Politics*, 5(1): 19–32, 2008.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.*, 46:389–422, March 2002. ISSN

0885-6125. doi: http://dx.doi.org/10.1023/A:1012487302797. URL http://dx.doi.org/10.1023/A:1012487302797.

Charles B. Handy. *Understanding Organisations*. Institute of Purchasing & Supply, 1985. ISBN 9780785556923. URL http://books.google.com/books?id=JBSEAQAACAAJ.

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

Leo Hickey. Surprise, surprise, but do so politely. *Journal of pragmatics*, 15(4):367–372, 1991.

Janet Holmes. *An Introduction to Sociolinguistics*. Pearson Longman, 1992.

Janet Holmes. *Women, men and politeness*. Longman, 1995.

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

Yufang Hou, Katja Markert, and Michael Strube. Cascading collective classification for bridging anaphora recognition using a rich linguistic feature set. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 814–820, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D13-1077.

Dirk Hovy. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, Beijing, China, July 2015. Association for Computational Linguistics.

Jun Hu, Rebecca Passonneau, and Owen Rambow. Contrasting the Interaction Structure of an Email and a Telephone Corpus: A Machine Learning Approach to Annotation of Dialogue Function Units. In *Proceedings of the SIGDIAL 2009 Conference*, London, UK, September 2009. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W09-3953.

Meishan Hu, Aixin Sun, and Ee-Peng Lim. Comments-oriented blog summarization by sentence extraction. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 901–904. ACM, 2007.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P14-1105`.

D Janbek and P Prado. Rethinking the role of virtual communities in terrorist websites. *Combating Terrorism Exchange*, 2(4):23–12, 2012.

Nathalie Japkowicz. Learning from imbalanced data sets: A comparison of various strategies. pages 10–15. AAAI Press, 2000.

Thorsten Joachims. 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, 1999. MIT Press.

Thorsten Joachims. Training Linear SVMs in Linear Time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217–226. ACM, 2006.

Pamela W. Jordan and Barbara Di Eugenio. Control and initiative in collaborative problem solving dialogues. In *Working Notes of the AAAI Spring Symposium on Computational Models for Mixed Initiative*, pages 81–84, 1997.

Stefan Kaufmann, Cleo Condoravdi, and Valentina Harizanov. *Formal Approaches to Modality*, pages 72–106. Mouton de Gruyter, 2006.

Parambir S Keila and DB Skillicorn. Detecting unusual and deceptive communication in email. In *Centers for Advanced Studies Conference*, pages 17–20, 2005.

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

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

Jihie Kim, Grace Chern, Donghui Feng, Erin Shaw, and Eduard Hovy. Mining and Assessing Discussions on the Web Through Speech Act Analysis. In *Proceedings of the Workshop on Web Content Mining with Human Language Technologies at the 5th International Semantic Web Conference*, 2006.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. Classifying Dialogue Acts in One-on-one Live Chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 862–871. Association for Computational Linguistics, 2010a.

Su Nam Kim, Li Wang, and Timothy Baldwin. Tagging and Linking Web Forum Posts. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 192–202. Association for Computational Linguistics, 2010b.

Robert Allen King, Pradeep Racherla, and Victoria D Bush. What we know and don't know about online word-of-mouth: A review and synthesis of the literature. *Journal of Interactive Marketing*, 28(3):167–183, 2014.

Paul Kiparsky and Carol Kiparsky. Facts. In Manfred Bierwisch and Karl Erich Heidolph, editors, *Progress in Linguistics*, pages 143–173. Mouton, The Hague, Paris, 1970.

Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *Machine learning: ECML 2004*, pages 217–226. Springer, 2004.

Angelika Kratzer. Modality. In Arnim von Stechow and Dieter Wunderlich, editors, *Semantics: An International Handbook of Contemporary Research*. Walter de Gruyter, Berlin, 1991.

Taku Kudo and Yuji Matsumoto. Yamcha: Yet another multipurpose chunk annotator, 2005.

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

Robin Lakoff. Language and Woman's Place. *Language in society*, 2(01):45–79, 1973.

Andrew Lampert, Robert Dale, and Cecile Paris. Detecting emails containing requests for action. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 984–992. Association for Computational Linguistics, 2010.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

Hsuan-Tien Lin, Chih-Jen Lin, and Ruby C. Weng. A Note on Platt's Probabilistic Outputs for Support Vector Machines. *Mach. Learn.*, 68:267–276, October 2007. ISSN 0885-6125. doi: 10.1007/s10994-007-5018-6. URL http://dl.acm.org/citation.cfm?id=1286062.1286078.

Miriam A. Locher. *Power and politeness in action: disagreements in oral communication*. Language, power, and social process. M. de Gruyter, 2004. ISBN 9783110180077. URL http://books.google.com/books?id=Aa32A4gWb8sC.

Miriam A Locher and Derek Bousfield. Introduction: Impoliteness and power in language. *LANGUAGE POWER AND SOCIAL PROCESS*, 21:1, 2008.

Max Louwerse, King-Ip Lin, Amanda Drescher, and Gün Semin. Linguistic cues predict fraudulent events in a corporate social network. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 961–966, 2010.

Stephan Ludwig, Ko De Ruyter, Mike Friedman, Elisabeth C Brüggen, Martin Wetzels, and Gerard Pfann. More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing*, 77(1):87–103, 2013.

Yuval Marton, Nizar Habash, and Owen Rambow. Dependency parsing of modern standard arabic with lexical and inflectional features. *Computational Linguistics*, 39(1):161–194, 2013.

Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, pages 249–272, 2007.

Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

Marjorie McShane, Sergei Nirenburg, and Ron Zacharsky. Mood and modality: Out of the theory and into the fray. *Natural Language Engineering*, 19(1):57–89, 2004.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Sara Mills. *Gender and politeness*, volume 17. Cambridge University Press, 2003.

Saif Mohammad and Tony Yang. 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 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W11-1709`.

Seungwhan Moon, Saloni Potdar, and Lara Martin. Identifying student leaders from mooc discussion forums through language influence. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*, pages 15–20, Doha, Qatar, October 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W14-4103`.

Shahriar Tanvir Hasan Murshed, Joseph G. Davis, and Liaquat Hossain. Social network analysis and organizational disintegration: the case of Enron corporation. In *International Conference on Information Systems (ICIS2007)*, 2007.

Galileo Mark S. Namata, Jr., Lise Getoor, and Christopher P. Diehl. Inferring organizational titles in online communication. In *Proceedings of the 2006 conference on Statistical network analysis*, ICML'06, pages 179–181, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 978-3-540-73132-0. URL `http://dl.acm.org/citation.cfm?id=1768341.1768359`.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2):4, 2007.

Sik Hung Ng and James J Bradac. *Power in language: Verbal communication and social influence.* Sage Publications, Inc, 1993.

Sik Hung Ng, Dean Bell, and Mark Brooke. Gaining turns and achieving high in influence ranking in small conversational groups. *British Journal of Social Psychology*, pages 32, 265–275, 1993.

Sik Hung Ng, Mark Brooke, and Michael Dunne. Interruption and in influence in discussion groups. *Journal of Language and Social Psychology*, pages 14(4),369–381, 1995.

Dong Nguyen, Dolf Trieschnigg, A. Seza Doğruöz, Rilana Gravel, Mariet Theune, Theo Meder, and Franciska De Jong. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL http://www.aclweb.org/anthology/C14-1184.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. Sits: A hierarchical nonparametric model using speaker identity for topic segmentation in multiparty conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 78–87, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P12-1009.

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

et al. Oard, Douglas. Avocado research email collection ldc2015t03. dvd. 2015.

William M. O'Barr. *Linguistic evidence: language, power, and strategy in the courtroom*. Studies on law and social control. Academic Press, 1982. ISBN 9780125235211. URL http://books.google.com/books?id=bqO0PwAACAAJ.

Philip V. Ogren, Philipp G. Wetzler, and Steven Bethard. 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)*, 2008.

Adinoyi Omuya, Vinodkumar Prabhakaran, and Owen Rambow. 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 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N13-1099.

Rebecca J Passonneau and Diane J Litman. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139, 1997.

Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2011.

Kristiaan Pelckmans, Jos De Brabanter, Johan AK Suykens, and Bart De Moor. Handling missing values in support vector machine classifiers. *Neural Networks*, 18(5):684–692, 2005.

James W Pennebaker. The secret life of pronouns. *New Scientist*, 211(2828):42–45, 2011.

James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 2001.

C. Raymond Perrault and James F. Allen. A plan-based analysis of indirect speech acts. *Computational Linguistics*, 6(3–4):167–182, 1980.

Kelly Peterson, Matt Hohensee, and Fei Xia. 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 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W11-0711`.

Jeffrey Pfeffer. *Power in organizations*. Pitman, Marshfield, MA., 1981. ISBN 0273016385. URL `http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+020986300&sourceid=fbw_bibsonomy`.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999.

Vinodkumar Prabhakaran and Owen Rambow. 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 2013. Asian Federation of Natural Language Processing. URL `http://www.aclweb.org/anthology/I13-1025`.

Vinodkumar Prabhakaran and Owen Rambow. 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 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P14-2056.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. Automatic committed belief tagging. In *Coling 2010: Posters*, pages 1014–1022, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL http://www.aclweb.org/anthology/C10-2117.

Vinodkumar Prabhakaran, Michael Bloodgood, Mona Diab, Bonnie Dorr, Lori Levin, Christine D. Piatko, Owen Rambow, and Benjamin Van Durme. Statistical modality tagging from rule-based annotations and crowdsourcing. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 57–64, Jeju, Republic of Korea, July 2012a. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W12-3807.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. Predicting overt display of power in written dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 518–522, Montréal, Canada, June 2012b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N12-1057.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. Annotations for power relations on email threads. In *Proceedings of the Eighth conference on LREC*, Istanbul, Turkey, May 2012c. European Language Resources Association (ELRA).

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. Who's (Really) the Boss? Perception of Situational Power in Written Interactions. In *24th International Conference on Computational Linguistics (COLING)*, Mumbai, India, 2012d. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Ajita John, and Dorée D. Seligmann. 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 2013a. Asian Federation of Natural Language Processing. URL http://www.aclweb.org/anthology/I13-1042.

Vinodkumar Prabhakaran, Ajita John, and Dorée D. Seligmann. Power dynamics in spoken interactions: a case study on 2012 republican primary debates. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 99–100. International World Wide Web Conferences Steering Committee, 2013b.

Vinodkumar Prabhakaran, Ashima Arora, and Owen Rambow. Staying on topic: An indicator of power in political debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, October 2014a. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Ashima Arora, and Owen Rambow. Power of confidence: How poll scores impact topic dynamics in political debates. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 77–82, Baltimore, Maryland, June 2014b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W14/W14-2710.

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

Vinodkumar Prabhakaran, Tomas By, Julia Hirschberg, Owen Rambow, Samira Shaikh, Tomek Strzalkowski, Jennifer Tracey, Michael Arrigo, Rupayan Basu, Micah Clark, Adam Dalton, Mona Diab, Louise Guthrie, Anna Prokofieva, Stephanie Strassel, Gregory Werner, Janyce Wiebe, and Yorick Wilks. A New Dataset and Evaluation for Belief/Factuality. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, Denver, USA, June 2015. Association for Computational Linguistics.

Gerard Prendergast, David Ko, and V Yuen Siu Yin. Online word of mouth and consumer purchase intentions. *International Journal of Advertising*, 29(5):687–708, 2010.

Anna Prokofieva and Julia Hirschberg. Hedging and speaker commitment. In *5th International Workshop on Emotion, Social Signals, Sentiment and Linked Open Data. LREC*, 2014.

Owen Rambow, Lokesh Shrestha, John Chen, and Christy Laurdisen. Summarizing email threads. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Short Papers*, pages 105–108, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

Anand S. Rao and Michael P. Georgeff. Modeling rational agents within a BDI-architecture. In James Allen, Richard Fikes, and Erik Sandewall, editors, *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, pages 473–484. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA, 1991.

Scott A. Reid and Sik Hung Ng. Conversation as a resource for in influence: evidence for prototypical arguments and social identification processes. *European Journal of Social Psych.*, pages 30, 83–100, 2000.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

Andrew Rosenberg and Julia Hirschberg. Charisma perception from text and speech. *Speech Communication*, 51(7):640–655, 2009.

Ryan Rowe, German Creamer, Shlomo Hershkop, and Salvatore J. Stolfo. Automated social hierarchy detection through email network analysis. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web Mining and Social Network Anal.* ACM, 2007.

Juhani Rudanko. Aggravated impoliteness and two types of speaker intention in an episode in shakespeare's timon of athens. *Journal of Pragmatics*, 38(6):829–841, 2006.

Sacks, E Schegloff, and G Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735, 1974.

Roser Saurí and James Pustejovsky. FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268, 2009. ISSN 1574-020X. URL http://dx.doi.org/10.1007/s10579-009-9089-9. 10.1007/s10579-009-9089-9.

Klaus R Scherer. Voice and speech correlates of perceived social influence in simulated juries. In *H. Giles and R. St Clair (Eds), Language and social psychology*, pages 88–120. Oxford: Blackwell, 1979.

Karin Kipper Schuler. Verbnet: a broad-coverage, comprehensive verb lexicon. 2005.

John R Searle. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press, 1969.

John Rogers Searle. A Classification of Illocutionary Acts. *Language in society*, 5(01):1–23, 1976.

J. Bryan Sexton and Robert L. Helmreich. Analyzing cockpit communication: The links between language, performance, error, and workload. In *Proceedings of the Tenth International Symposium on Aviation Psychology*, pages 689–695, 1999.

Jitesh Shetty and Jafar Adibi. The Enron email dataset database schema and brief statistical report. 2004.

Jitesh Shetty and Jafar Adibi. Discovering important nodes through graph entropy the case of Enron email database. In *Proceedings of the 3rd international workshop on Link discovery*, LinkKDD '05, pages 74–81, New York, NY, USA, 2005. ACM. ISBN 1-59593-215-1. doi: http://doi.acm.org/10.1145/1134271.1134282. URL http://doi.acm.org/10.1145/1134271.1134282.

Yanchuan Sim, Brice D. L. Acree, Justin H. Gross, and Noah A. Smith. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on EMNLP*, pages 91–101, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D13-1010.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. Detecting arguing and sentiment in meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, 2007.

Swapna Somasundaran, Janyce Wiebe, and Josef Ruppenhofer. Discourse level opinion interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 801–808, Manchester, UK, August 2008. Coling 2008 Organizing Committee. URL http://www.aclweb.org/anthology/C08-1101.

Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational linguistics*, 26 (3):339–373, 2000.

Carlo Strapparava, Marco Guerini, and Oliviero Stock. Predicting persuasiveness in political discourses. In *LREC*. Citeseer, 2010.

Tomek Strzalkowski, George Aaron Broadwell, Jennifer Stromer-Galley, Samira Shaikh, Sarah Taylor, and Nick Webb. Modeling socio-cultural phenomena in discourse. In *Proceedings of the 23rd International Conference on COLING 2010*, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL http://www.aclweb.org/anthology/C10-1117.

Swabha Swayamdipta and Owen Rambow. The pursuit of power and its manifestation in written dialog. *2012 IEEE Sixth International Conference on Semantic Computing*, 0:22–29, 2012. doi: http://doi.ieeecomputersociety.org/10.1109/ICSC.2012.49.

Deborah Tannen. *You just don't understand: Women and men in conversation*. Virago London, 1991.

Deborah Tannen. *Gender and Conversational Interaction*. Oxford: Oxford University Press., 1993.

Deborah Tannen. *Talking from 9 to 5: How Women's and Men's Conversational Styles Affect who Gets Heard, who Gets Credit, and what Gets Done at Work*. W. Morrow, 1994. ISBN 9780688112431. URL https://books.google.com/books?id=uP7ehYicXQYC.

Sarah M. Taylor, Ting Liu, Samira Shaikh, Tomek Strzalkowski, George Aaron Broadwell, Jennifer Stromer-Galley, Umit Boz, Xiaoai Ren, Jingsi Wu, and Feifei Zhang. Chinese and American Leadership Characteristics: Discovery and Comparison in Multi-party On-Line Dialogues. In *ICSC*, pages 17–21. IEEE Computer Society, 2012. ISBN 978-1-4673-4433-3.

Marina Terkourafi. Toward a unified theory of politeness, impoliteness, and rudeness. *Impoliteness in Language: Studies on its Interplay with Power in Theory and Prac-tice, Derek Bousfield and Miriam A. Locher (eds)*, pages 45–74, 2008.

Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney, Australia, July 2006. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W06/W06-1639`.

Stephen Tomlinson. Learning Task Experiments in the TREC 2010 Legal Track. In E. M. Voorhees and Lori P. Buckland, editors, *The Nineteenth Text REtrieval Conference Proceedings (TREC 2010)*. National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA), November 2010.

Teun A Van Dijk. Structures of discourse and structures of power. 12, 1989.

Marilyn Walker and Steve Whittaker. Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, pages 70–78. Association for Computational Linguistics, 1990.

Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM, 2006.

Thomas E. Wartenberg. *The forms of power: from domination to transformation*. Temple University Press, 1990. ISBN 9780877226482. URL `http://books.google.sh/books?id=yK52QgAACAAJ`.

Gregory J. Werner, Vinodkumar Prabhakaran, Mona Diab, and Owen Rambow. Committed belief tagging on the factbank and lu corpora: A comparative study. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, Denver, USA, June 2015. Association for Computational Linguistics.

Candace West. Not just 'doctors' orders': directive-response sequences in patients' visits to women and men physicians. *Discourse & Society*, 1(1):85–112, 1990.

Candace West and Don H Zimmerman. Doing Gender. *Gender and society*, 1(2):125–151, 1987.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning subjective language. *Computational linguistics*, 30(3):277–308, 2004.

Theresa Wilson and Janyce Wiebe. Annotating attributions and private states. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 53–60, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W05/W05-0308`.

John T Woolley and Gerhard Peters. The american presidency project [online]. santa barbara, ca. *Available from World Wide Web: http://www.presidency.ucsb.edu.*, 11, 2011.

Alexander Yeh. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, COLING '00, pages 947–953, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/992730.992783. URL `http://dx.doi.org/10.3115/992730.992783`.

Jen-Yuan Yeh and Aaron Harnly. 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 2006.

Renxian Zhang, Dehong Gao, and Wenjie Li. Towards Scalable Speech Act Recognition in Twitter: Tackling Insufficient Training Data. *EACL 2012*, page 18, 2012.

Don H. Zimmerman and Candace West. Sex roles, interruptions and silences in conversation. 1975.

# Part VI

# APPENDICES

# Appendix A

# Example Email Threads

## A.1 Example Thread (ID: 65)

<thread>

<thread_id>65</thread_id>

<message>

<depth> 0</depth>

<parent_id> NULL</parent_id>

<message_id> 106860</message_id>

<date_time> 2001-04-18 10:28:50</date_time>

<subject> Charge Methodology</subject>

<from name="David Forster" id="28701" address="David.Forster@ENRON.com" />

<to name="Andy Zipper" id="5063" address="Andy.Zipper@ENRON.com" />

<to name="Andy Zipper" id="5063" address="azipper@enron.com" />

<content>

M1.1. Andy,

M1.2. Attached are some ideas for possible charge structures for EnronOnline.

M1.3. I am recommending something which will probably be surprising, given our conversation.

M1.4. Let's discuss when you have a moment.

M1.5. Dave

M1.6. Recommendation

M1.7. a) For new commodity areas, continue to charge a set up fee in accordance with our previously agreed schedule.

M1.8. (e.g. $350,000 for new Market Area)

M1.9. b) Charge a per-volume maintenance fee which is comparable to industry brokerage fees,

M1.10. with a minimum charge equivalent to $4,000 per Product * Total number of Products).

M1.11. This method results in a total charge of approx. $46.5 million pa.

M1.12. (providing coverage for existing charges, some growth and London's charges)

M1.13. This method is recommended because it balances a fee to reflect real expenditure of effort on behalf on Enron Online staff (the per Product minimum)

M1.14. with a structure which is recognizeable (and hopefully more easily sold) to the traders.

M1.15. This structure is primarily not cost-driven, but is value-driven;

M1.16. those who derive the greatest value pay the highest costs.

M1.17. Example Charges

M1.18. Here are some example charges if we use the recommended method:

M1.19. Commodity Charge US Nat Gas $24,282,875 US Power $ 5,163,563 Metals $ 5,798,144 Crude & Products $ 3,578,560 Norwegian Power $ 380,869 Global Credit $ 400,000 Coal $ 966,265 Bandwidth $ 312,000

M1.20. Or, by Group: ENA $30,647,772 EEL $ 8,774,844 EGM $ 6,741,955 EIM $ 106,250 EBS $ 312,000 Total: $46,582,821

M1.21. Sensitivity

M1.22. With the recommended structure, if transactions for 2001 are:

M1.23. a) The same as the last half of 2000 * 2, then we recover approx. $46 million.

M1.24. b) Double, then we recover approx. $90 million

M1.25. c) Half, then we recover approx.$24 million

M1.26. c) Zero, then we recover approx. $6.4 million

M1.27. Alternatives - Basic Structure

M1.28. Any of the following could be combined to create additional alternatives:

M1.29. Alternative 1: As per the recommended structure, but charge a flat per-transaction fee instead of a per volume fee. This would result in a charge of $xx per transaction.

M1.30. Alternative 2: Charge by Product Types (we currently have 358, so full charge would be approx. $112,000 per Product Type)

M1.31. Alternative 3: Charge by Products (we currently have 1500 per day, so full charge would be approx. $27,000 per Product)

M1.32. Alternative 4: Charge by Country/Commodity (we currently have 61, so full charge would be approx. $656,000 per Country/Commodity)

M1.33. Alternative 5: Use the same methodology as currently used for the cost allocation (55M1.34. Alternatives - Different Structures

M1.35. Alternative 5: Separate Marketing costs and charge directly to business units based on activity

M1.36. Alternative 6: Separate Development costs as a separate item not covered by the basic charge structure, but recovered solely through increases in EnronOnline business.

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 106860</parent_id>

<message_id> 2000020</message_id>

<date_time> 2001-04-18 12:56:13</date_time>

<subject> Charge Methodology</subject>

<from id="" name="" address="" />

<content>

M2.1. Okay. good start.

M2.2. I like the idea of a minimum for each product,

M2.3. and I like staying with existing structure for new products.

M2.4. Let's look at alternative 1, the flat fee per trade,

M2.5. and see what it would need to be to yield $35mm in revs based on average tradecount YTD,

M2.6. i.e. extraplolate that out for rest of year for a pro forma.

M2.7. Thoughts ?

</content>

</message>

<message>

<depth> 2</depth>

<parent_id> 2000020</parent_id>

<message_id> 106859</message_id>

<date_time> 2001-04-18 15:23:37</date_time>

<subject> Charge Methodology</subject>

<from name="David Forster" id="28701" address="David.Forster@ENRON.com" />

<to name="Andy Zipper" id="5063" address="Andy.Zipper@ENRON.com" />

<to name="Andy Zipper" id="5063" address="azipper@enron.com" />

<content>

M3.1. Sorry - xxx (below) was supposed to be replaced with $54.50 per transaction,

M3.2. which is based on 2 * the July00-Dec00 transaction count.

M3.3. Note this results in $40 million of recovery, which includes Amita's costs.

M3.4. If you just look at $35 million recovery, the per-transaction fee is $47.68.

M3.5. This structure has the advantage that it is a little closer to the current cost allocation methodology,

M3.6. but this methodology is not well known by the business units.

M3.7. I actually started drafting this email with Alternative 1 as the recommendation,

M3.8. but decided that if we have to sell this internally, the brokerage lookalike structure would be easer to sell and defend.

M3.9. Dave

</content>

</message>
<message>
<depth> 3</depth>
<parent_id> 106859</parent_id>
<message_id> 2000021</message_id>
<date_time> 2001-04-18 15:29:33</date_time>
<subject> Charge Methodology</subject>
<from id="" name="" address="" />
<content>
M4.1. Could you send me the commission numbers used for each of the major products in your volume based approach .
</content>
</message>
<message>
<depth> 4</depth>
<parent_id> 2000021</parent_id>
<message_id> 106858</message_id>
<date_time> 2001-04-18 15:35:30</date_time>
<subject> Charge Methodology</subject>
<from name="David Forster" id="28701" address="David.Forster@ENRON.com" />
<to name="Andy Zipper" id="5063" address="Andy.Zipper@ENRON.com" />
<to name="Andy Zipper" id="5063" address="azipper@enron.com" />
<content>
M5.1. Sure.
M5.2. These are actually taken directly from the spreadsheet which Mike Bridges prepared several months ago.
M5.3. If this is a route you want to pursue, we should first verify each of the numbers as current.
M5.4. Dave
</content>
</message>
</thread>

## A.2 Example Thread (ID: 72)

<thread>
<thread_id>72</thread_id>
<message>
<depth> 0</depth>

<parent_id> NULL</parent_id>

<message_id> 275758</message_id>

<date_time> 2001-05-22 10:52:36</date_time>

<subject> EGM Good Job!</subject>

<from name="Shona Wilson" id="4708" address="Shona.Wilson@ENRON.com" />

<to name="Brent A Price" id="16090" address="Brent.A.Price@ENRON.com" />

<to name="Scott Earnest" id="36088" address="Scott.Earnest@ENRON.com" />

<to name="Michelle Bruce" id="78130" address="Michelle.Bruce@ENRON.com" />

<to name="D Todd Hall" id="26856" address="D.Todd.Hall@ENRON.com" />

<to name="Brent A Price" id="16090" address="bprice@enron.com" />

<to name="Scott Earnest" id="36088" address="Searnes@ENRON.com" />

<to name="Michelle Bruce" id="78130" address="Mbruce@ENRON.com" />

<to name="D Todd Hall" id="26856" address="Thall@ENRON.com" />

<cc name="Sally Beck" id="46153" address="Sally.Beck@ENRON.com" />

<cc name="Beth Apollo" id="11750" address="Beth.Apollo@ENRON.com" />

<cc name="Sally Beck" id="46153" address="Sbeck@ENRON.com" />

<cc name="Beth Apollo" id="11750" address="bapollo@enron.com" />

<content>

M1.1. During the morning meeting today Rick Buy commented on what a good job EGM has done in the officialization process.

M1.2. He took your names because he wanted to know who to thank for this.

M1.3. His comments are based on the summary graph -

M1.4. his words - EGM has by far the most books but never shows up on the log -

M1.5. who can we thank for that?

M1.6. Keep up the good work!

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 275758</parent_id>

<message_id> 275759</message_id>

<date_time> 2001-05-22 13:16:14</date_time>

<subject> EGM Good Job!</subject>

<from name="Michelle Bruce" id="78130" address="Michelle.Bruce@ENRON.com" />

<to name="Shona Wilson" id="4708" address="Shona.Wilson@ENRON.com" />

<to name="Brent A Price" id="16090" address="Brent.A.Price@ENRON.com" />

<to name="Scott Earnest" id="36088" address="Scott.Earnest@ENRON.com" />

<to name="D Todd Hall" id="26856" address="D.Todd.Hall@ENRON.com" />

<to name="Shona Wilson" id="4708" address="A2A6890E-4BBECE71-862568E1-76B49B@ENRON.com" />

<to name="Brent A Price" id="16090" address="bprice@enron.com" />

<to name="Scott Earnest" id="36088" address="Searnes@ENRON.com" />

<to name="D Todd Hall" id="26856" address="Thall@ENRON.com" />

<cc name="Sally Beck" id="46153" address="Sally.Beck@ENRON.com" />

<cc name="Beth Apollo" id="11750" address="Beth.Apollo@ENRON.com" />

<cc name="Sally Beck" id="46153" address="Sbeck@ENRON.com" />

<cc name="Beth Apollo" id="11750" address="bapollo@enron.com" />

<content>

M2.1. Shona-

M2.2. The officialization process that EGM - Global Products uses was developed when we were implementing our Global Risk Management Worldwide Close process.

M2.3. We (at that point) had an excel speadsheet that would search ERMS by book codes and pull in the officialized post id for the day -

M2.4. if there was a blank, the book was not official at which point the book administrator then goes back into the ERMS database, officializes and then we run VAR and our daily reports.

M2.5. We are now doing this via our Global Reporting database for Global Products.

M2.6. In developing the process, it was myself, Scott Earnest, Mark Fondren and Simon Thurbin (London office).

M2.7. We now have Bill Kazemervisz and Vera Ilyina who are running the database daily and making certain that all listed books have been officialized.

M2.8. John Swinney is our Risk Manager of the Global Products team

M2.9. and he is involved making certain that the risk managers are updating our new books list so that we are certain to mark them official nightly as well as integrating any new processes relating to VAR, exotic file uploads, etc.

M2.10. As you can see, it is very much a team effort that has really paid off on "next morning issues" for our group.

M2.11. If you have any additional questions or comments, please call me.

M2.12. Thank you,

M2.13. Michelle xt. 57532

</content>

</message>

</thread>

## A.3 Example Thread (ID: 81)

<thread>

<thread_id>81</thread_id>

<message>

<depth> 0</depth>

<parent_id> NULL</parent_id>

<message_id> 284996</message_id>

<date_time> 2001-05-23 12:04:31</date_time>

<subject> Existing Trading Track Rotations</subject>

<from name="Karen Buckley" id="17792" address="Karen.Buckley@ENRON.com" />

<to name="Doug Gilbert-smith" id="805" address="Doug.Gilbert-Smith@ENRON.com" />

<to name="Lloyd Will" id="2796" address="Lloyd.Will@ENRON.com" />

<to name="Stacey W White" id="2310" address="Stacey.W.White@ENRON.com" />

<to name="Don Baughman Jr" id="6010" address="Don.Baughman@ENRON.com" />

<to name="Harry Arora" id="2445" address="Harry.Arora@ENRON.com" />

<to name="Mark Dana Davis" id="5075" address="Dana.Davis@ENRON.com" />

<to name="Louise Kitchen" id="14758" address="Rogers.Herndon@ENRON.com" />

<to name="David Gossett" id="20863" address="David.Gossett@ENRON.com" />

<to name="Chris Gaskill" id="12620" address="Chris.Gaskill@ENRON.com" />

<to name="Robert Superty" id="3385" address="Robert.Superty@ENRON.com" />

<to name="Fred Lagrasta" id="4902" address="Fred.Lagrasta@ENRON.com" />

<to name="Ed McMichael Jr" id="1599" address="Ed.McMichael@ENRON.com" />

<to name="Scott Neal" id="85818" address="Scott.Neal@ENRON.com" />

<to name="Doug Gilbert-smith" id="805" address="252841e-ff4f6af2-86256881-52c8f9@ENRON.com" />

<to name="Lloyd Will" id="2796" address="lwill@enron.com" />

<to name="Stacey W White" id="2310" address="64b26e5d-217acff6-8625665d-76b99e@ENRON.com" />

<to name="Don Baughman Jr" id="6010" address="AEFF4E89-1CD73C39-86256659-5290F3@ENRON.com" />

<to name="Harry Arora" id="2445" address="harora@enron.com" />

<to name="Mark Dana Davis" id="5075" address="ddavis@enron.com" />

<to name="Rogers Herndon" id="2438" address="6AA81A13-79303B50-862566EB-59BF78@ENRON.com" />

<to name="David Gossett" id="20863" address="cf2aee90-40a8d4f7-882568d9-4b4af9@ENRON.com" />

<to name="Chris Gaskill" id="12620" address="Cgaskill@ENRON.com" />

<to name="Robert Superty" id="3385" address="Rsupert@ENRON.com" />

<to name="Fred Lagrasta" id="4902" address="Flagras@ENRON.com" />

<to name="Ed McMichael Jr" id="1599" address="Emcmich@ENRON.com" />

<to name="Scott Neal" id="85818" address="sneal@enron.com" />

<content>

M1.1. Attached is the current list of rotations for the Trading Track participants and future assigned rotations (as decided at the time of hiring).

M1.2. Can you please re-confirm you have these people in your group currently,

M1.3. as I appear to have conflicting information.

M1.4. I will re-send, if any changes occur.

M1.5. If there is any movement of these people between groups can you plese keep me in the loop.

M1.6. Kind regards,

M1.7. Karen.

M1.8. x54667

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 284996</parent_id>

<message_id> 30994</message_id>

<date_time> 2001-05-24 13:54:28</date_time>

<subject> Existing Trading Track Rotations</subject>

<from name="Lloyd Will" id="2796" address="Lloyd.Will@ENRON.com" />

<to name="Karen Buckley" id="17792" address="Karen.Buckley@ENRON.com" />

<to name="Doug Gilbert-smith" id="805" address="Doug.Gilbert-Smith@ENRON.com" />

<to name="Stacey W White" id="2310" address="Stacey.W.White@ENRON.com" />

<to name="Don Baughman Jr" id="6010" address="Don.Baughman@ENRON.com" />

<to name="Harry Arora" id="2445" address="Harry.Arora@ENRON.com" />

<to name="Mark Dana Davis" id="5075" address="Dana.Davis@ENRON.com" />

<to name="Louise Kitchen" id="14758" address="Rogers.Herndon@ENRON.com" />

<to name="David Gossett" id="20863" address="David.Gossett@ENRON.com" />

<to name="Chris Gaskill" id="12620" address="Chris.Gaskill@ENRON.com" />

<to name="Robert Superty" id="3385" address="Robert.Superty@ENRON.com" />

<to name="Fred Lagrasta" id="4902" address="Fred.Lagrasta@ENRON.com" />

<to name="Ed McMichael Jr" id="1599" address="Ed.McMichael@ENRON.com" />

<to name="Scott Neal" id="85818" address="Scott.Neal@ENRON.com" />

<to name="Karen Buckley" id="17792" address="Kbuckley@ENRON.com" />

<to name="Doug Gilbert-smith" id="805" address="252841e-ff4f6af2-86256881-52c8f9@ENRON.com" />

<to name="Stacey W White" id="2310" address="64b26e5d-217acff6-8625665d-76b99e@ENRON.com" />

<to name="Don Baughman Jr" id="6010" address="Dbaughm@ENRON.com" />

<to name="Harry Arora" id="2445" address="harora@enron.com" />

<to name="Mark Dana Davis" id="5075" address="ddavis@enron.com" />
<to name="Rogers Herndon" id="2438" address="6AA81A13-79303B50-862566EB-59BF78@ENRON.com" />
<to name="David Gossett" id="20863" address="cf2aee90-40a8d4f7-882568d9-4b4af9@ENRON.com" />
<to name="Chris Gaskill" id="12620" address="Cgaskill@ENRON.com" />
<to name="Robert Superty" id="3385" address="Rsupert@ENRON.com" />
<to name="Fred Lagrasta" id="4902" address="Flagras@ENRON.com" />
<to name="Ed McMichael Jr" id="1599" address="Emcmich@ENRON.com" />
<to name="Scott Neal" id="85818" address="sneal@enron.com" />
<content>
M2.1. Karen attached is the latest status of the power trading track folks.
M2.2. I will use your template going forward to track changes.
M2.3. Thanks.
</content>
</message>
<message>
<depth> 1</depth>
<parent_id> 284996</parent_id>
<message_id> 1033837</message_id>
<date_time> 2001-05-23 12:12:32</date_time>
<subject> Existing Trading Track Rotations</subject>
<from name="Karen Buckley" id="17792" address="Karen.Buckley@ENRON.com" />
<to name="John Lavorato" id="2273" address="John.J.Lavorato@ENRON.com" />
<to name="Kimberly Hillis" id="66518" address="Kimberly.Hillis@ENRON.com" />
<to name="John Lavorato" id="2273" address="JLAVORA@ENRON.com" />
<to name="Kimberly Hillis" id="66518" address="Khillis@ENRON.com" />
<content>
M3.1. John
M3.2. Following the interviews next week,
M3.3. suggest your management team get together to review the rotations for the Trading Track,
M3.4. to see if they continue to meet business/Employe needs to ensure we have accurate data from a tracking perspective.
M3.5. Karen
</content>
</message>
<message>
<depth> 2</depth>
<parent_id> 1033837</parent_id>

<message_id> 2000025</message_id>

<date_time> 2001-05-24 10:38:36</date_time>

<subject> Existing Trading Track Rotations</subject>

<from id="" name="" address="" />

<content>

M4.1. Great.

M4.2. Do we have any outside people.

</content>

</message>

<message>

<depth> 3</depth>

<parent_id> 2000025</parent_id>

<message_id> 1033836</message_id>

<date_time> 2001-05-25 09:04:40</date_time>

<subject> Existing Trading Track Rotations</subject>

<from name="Karen Buckley" id="17792" address="Karen.Buckley@ENRON.com" />

<to name="John Lavorato" id="2273" address="John.J.Lavorato@ENRON.com" />

<to name="John Lavorato" id="2273" address="JLAVORA@ENRON.com" />

<content>

M5.1. Yes, we have 5 externals coming in next Wednesday.

M5.2. From the 32 resumes shortlisted by you

M5.3. 14 of which were interviewed by your direct reports (remainder were not interested in Houston/job)

M5.4. 8 of which were recommeded by your traders

M5.5. 6 of which accepted an 2nd round interview

M5.6. (five of which are confirmed for next Wednesday, 1 could not make that date)

M5.7. We therefore have a confirmd number of 15 to interview next Wednesday,

M5.8. however I was given additional internal names last night to follow up on this am.

M5.9. I will confirm exact number of interviews later this morning but could be anywhere in the region from 15-20.

M5.10. I have asked the current group in the Trading Track to faciliate an office tour of the trading floors/gas control room.

M5.11. Following which there will be an informal lunch for all of the candidates in one of the conference rooms to include the current Trading Track folks.

M5.12. Interviews will follow at 2.00 pm at the Allen Center.

M5.13. Schedules are being formalised today and will be forwarded to you.

M5.14. Thanks,

</content>

</message>

</thread>

## A.4 Example Thread (ID: 86)

<thread>
<thread_id>86</thread_id>
<message>
<depth> 0</depth>
<parent_id> NULL</parent_id>
<message_id> 1033844</message_id>
<date_time> 2001-05-24 12:09:26</date_time>
<subject> Fines and EOL</subject>
<from name="Hunter S Shively" id="43716" address="Hunter.S.Shively@ENRON.com" />
<to name="John Lavorato" id="2273" address="John.J.Lavorato@ENRON.com" />
<to name="John Lavorato" id="2273" address="JLAVORA@ENRON.com" />
<content>
M1.1. Fines:
M1.2. I know you have talked with Phillip but I wanted to give my two cents.
M1.3. I think the fines are a great idea.
M1.4. We must be accountable for VAR.
M1.5. However the fines do not allow the desks to push the envelope.
M1.6. The desks need to stay a couple million under VAR to protect against volatility and factor changes.
M1.7. We do not have the tools to predict our VAR with any strong degree of accuracy
M1.8. but we are penalized for going over a few hundred thousand dollars.
M1.9. I propose a one day grace period of 10M1.10. This would allow the desks to max their VAR and protect against unexpected changes.
M1.11. EOL:
M1.12. We continue to have trouble with brokers not working our EOL numbers.
M1.13. We believe Dynegy has written a program to mirror our cash markets on their system.
M1.14. When we suspend, they suspend
M1.15. and as our markets move, so do theirs.
</content>
</message>
<message>
<depth> 1</depth>
<parent_id> 1033844</parent_id>

<message_id> 1033845</message_id>
<date_time> 2001-05-24 12:30:55</date_time>
<subject> Fines and EOL - correction</subject>
<from name="Hunter S Shively" id="43716" address="Hunter.S.Shively@ENRON.com" />
<to name="John Lavorato" id="2273" address="John.J.Lavorato@ENRON.com" />
<to name="John Lavorato" id="2273" address="JLAVORA@ENRON.com" />
<content></content>
</message>
<message>
<depth> 1</depth>
<parent_id> 1033844</parent_id>
<message_id> 286575</message_id>
<date_time> 2001-05-24 15:48:51</date_time>
<subject> Fines and EOL</subject>
<from name="John Lavorato" id="2273" address="John.J.Lavorato@ENRON.com" />
<to name="Hunter S Shively" id="43716" address="Hunter.S.Shively@ENRON.com" />
<to name="Hunter S Shively" id="43716" address="hshivel@enron.com" />
<content>
M3.1. Call me about the dynegy thing I have an idea.
</content>
</message>
</thread>

## A.5 Example Thread (ID: 2542)

<thread>
<thread_id>2542</thread_id>
<message>
<depth> 0</depth>
<parent_id></parent_id>
<message_id> 1300350</message_id>
<date_time> 2001-11-21 12:17:03</date_time>
<subject></subject>
<from name="Michelle Nelson" id="81118" address="Michelle.Nelson@ENRON.com" />
<to name="Mike Maggi" id="3774" address="Mike.Maggi@ENRON.com" />
<to name="Mike Maggi" id="3774" address="mmaggi@enron.com" />

<content>

M1.1. did you book my flight to pakistan?

M1.2. i am thinking that i need to be leaving soon. like tomorrow.

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 1300350</parent_id>

<message_id> 2001058</message_id>

<date_time> 2001-11-21 12:18:29</date_time>

<subject></subject>

<from id="" name="" address="" />

<content>

M2.1. nope ny

</content>

</message>

<message>

<depth> 2</depth>

<parent_id> 2001058</parent_id>

<message_id> 1300352</message_id>

<date_time> 2001-11-21 12:19:56</date_time>

<subject></subject>

<from name="Michelle Nelson" id="81118" address="Michelle.Nelson@ENRON.com" />

<to name="Mike Maggi" id="3774" address="Mike.Maggi@ENRON.com" />

<to name="Mike Maggi" id="3774" address="mmaggi@enron.com" />

<content>

M3.1. very funny.

M3.2. so when are we leaving?

</content>

</message>

<message>

<depth> 3</depth>

<parent_id> 1300352</parent_id>

<message_id> 2001059</message_id>

<date_time> 2001-11-21 12:20:15</date_time>

<subject></subject>

<from id="" name="" address="" />

<content>
M4.1. next thursday
</content>
</message>
<message>
<depth> 4</depth>
<parent_id> 2001059</parent_id>
<message_id> 1300353</message_id>
<date_time> 2001-11-21 12:20:35</date_time>
<subject></subject>
<from name="Michelle Nelson" id="81118" address="Michelle.Nelson@ENRON.com" />
<to name="Mike Maggi" id="3774" address="Mike.Maggi@ENRON.com" />
<to name="Mike Maggi" id="3774" address="mmaggi@enron.com" />
<content>
M5.1. for how long?
</content>
</message>
<message>
<depth> 5</depth>
<parent_id> 1300353</parent_id>
<message_id> 2001060</message_id>
<date_time> 2001-11-21 12:22:14</date_time>
<subject></subject>
<from id="" name="" address="" />
<content>
M6.1. til sunday
</content>
</message>
<message>
<depth> 6</depth>
<parent_id> 2001060</parent_id>
<message_id> 1300357</message_id>
<date_time> 2001-11-21 12:23:54</date_time>
<subject></subject>
<from name="Michelle Nelson" id="81118" address="Michelle.Nelson@ENRON.com" />
<to name="Mike Maggi" id="3774" address="Mike.Maggi@ENRON.com" />
<to name="Mike Maggi" id="3774" address="mmaggi@enron.com" />

<content>
M7.1. you're a trouble maker. aren't you?
</content>
</message>
<message>
<depth> 6</depth>
<parent_id> 2001060</parent_id>
<message_id> 1300354</message_id>
<date_time> 2001-11-21 12:23:24</date_time>
<subject></subject>
<from name="Michelle Nelson" id="81118" address="Michelle.Nelson@ENRON.com" />
<to name="Mike Maggi" id="3774" address="Mike.Maggi@ENRON.com" />
<to name="Mike Maggi" id="3774" address="mmaggi@enron.com" />
<content>
M8.1. and how do you suggest that i manage that?
</content>
</message>
<message>
<depth> 7</depth>
<parent_id> 1300354</parent_id>
<message_id> 2001061</message_id>
<date_time> 2001-11-21 12:23:51</date_time>
<subject></subject>
<from id="" name="" address="" />
<content>
M9.1. you have to go for work
</content>
</message>
<message>
<depth> 8</depth>
<parent_id> 2001061</parent_id>
<message_id> 1300355</message_id>
<date_time> 2001-11-21 12:24:18</date_time>
<subject></subject>
<from name="Michelle Nelson" id="81118" address="Michelle.Nelson@ENRON.com" />
<to name="Mike Maggi" id="3774" address="Mike.Maggi@ENRON.com" />
<to name="Mike Maggi" id="3774" address="mmaggi@enron.com" />

<content>

M10.1. i see.

M10.2. and where do i tell fred and ny that i am going?

</content>

</message>

<message>

<depth> 10</depth>

<parent_id></parent_id>

<message_id> 1300356</message_id>

<date_time> 2001-11-21 12:33:25</date_time>

<subject></subject>

<from name="Michelle Nelson" id="81118" address="Michelle.Nelson@ENRON.com" />

<to name="Mike Maggi" id="3774" address="Mike.Maggi@ENRON.com" />

<to name="Mike Maggi" id="3774" address="mmaggi@enron.com" />

<content>

M11.1. you have all the answers don't you?

M11.2. don't you think that someone will figure out your little plan?

M11.3. oh and thanks for coming down here and giving me some news.....fink

</content>

</message>

<message>

<depth> 11</depth>

<parent_id> 1300356</parent_id>

<message_id> 2001062</message_id>

<date_time> 2001-11-21 12:34:12</date_time>

<subject></subject>

<from id="" name="" address="" />

<content>

M12.1. i know im all talk, ill be down in a little while

</content>

</message>

<message>

<depth> 12</depth>

<parent_id> 2001062</parent_id>

<message_id> 1300359</message_id>

<date_time> 2001-11-21 12:35:00</date_time>

<subject></subject>

<from name="Michelle Nelson" id="81118" address="Michelle.Nelson@ENRON.com" />

<to name="Mike Maggi" id="3774" address="Mike.Maggi@ENRON.com" />

<to name="Mike Maggi" id="3774" address="mmaggi@enron.com" />

<content>

M13.1. you are all talk. and you are a fink.

</content>

</message>

<message>

<depth> 13</depth>

<parent_id> 1300359</parent_id>

<message_id> 2001063</message_id>

<date_time> 2001-11-21 12:38:18</date_time>

<subject></subject>

<from id="" name="" address="" />

<content>

M14.1. why am i a fink

</content>

</message>

<message>

<depth> 14</depth>

<parent_id> 2001063</parent_id>

<message_id> 1300360</message_id>

<date_time> 2001-11-21 12:41:37</date_time>

<subject></subject>

<from name="Michelle Nelson" id="81118" address="Michelle.Nelson@ENRON.com" />

<to name="Mike Maggi" id="3774" address="Mike.Maggi@ENRON.com" />

<to name="Mike Maggi" id="3774" address="mmaggi@enron.com" />

<content>

M15.1. you actually know what that is?

M15.2. i'm impressed!

</content>

</message>

<message>

<depth> 15</depth>

<parent_id> 1300360</parent_id>

<message_id> 2001064</message_id>

<date_time> 2001-11-21 12:44:22</date_time>

<subject></subject>
<from id="" name="" address="" />
<content>
M16.1. thats my goal
</content>
</message>
<message>
<depth> 16</depth>
<parent_id> 2001064</parent_id>
<message_id> 1300361</message_id>
<date_time> 2001-11-21 12:47:08</date_time>
<subject></subject>
<from name="Michelle Nelson" id="81118" address="Michelle.Nelson@ENRON.com" />
<to name="Mike Maggi" id="3774" address="Mike.Maggi@ENRON.com" />
<to name="Mike Maggi" id="3774" address="mmaggi@enron.com" />
<content>
M17.1. whatever.
M17.2. so i was right about your puff break wasn't i?
</content>
</message>
<message>
<depth> 18</depth>
<parent_id></parent_id>
<message_id> 1300362</message_id>
<date_time> 2001-11-21 12:58:32</date_time>
<subject></subject>
<from name="Michelle Nelson" id="81118" address="Michelle.Nelson@ENRON.com" />
<to name="Mike Maggi" id="3774" address="Mike.Maggi@ENRON.com" />
<to name="Mike Maggi" id="3774" address="mmaggi@enron.com" />
<content>
M18.1. hmm.
M18.2. i think that you are lying.
</content>
</message>
<message>
<depth> 19</depth>
<parent_id> 1300362</parent_id>

<message_id> 2001065</message_id>

<date_time> 2001-11-21 13:00:23</date_time>

<subject></subject>

<from id="" name="" address="" />

<content>

M19.1. i would never

</content>

</message>

<message>

<depth> 20</depth>

<parent_id> 2001065</parent_id>

<message_id> 1300368</message_id>

<date_time> 2001-11-21 13:02:14</date_time>

<subject></subject>

<from name="Michelle Nelson" id="81118" address="Michelle.Nelson@ENRON.com" />

<to name="Mike Maggi" id="3774" address="Mike.Maggi@ENRON.com" />

<to name="Mike Maggi" id="3774" address="mmaggi@enron.com" />

<content>

M20.1. are you sure about that?

M20.2. i am bored.

</content>

</message>

</thread>

## A.6 Example Thread (ID: 6837)

<thread>

<thread_id>6837</thread_id>

<message>

<depth> 0</depth>

<parent_id></parent_id>

<message_id> 225550</message_id>

<date_time> 1999-11-09 05:48:00</date_time>

<subject> JAPRO Gruppen Aktiebolag</subject>

<from name="Sara Shackleton" id="64528" address="Sara.Shackleton@ENRON.com" />

<to name="Mark Taylor" id="2378" address="Mark.Taylor@ENRON.com" />

<content>

M1.1. Martin sent this message to London and Michael advised that Sullivan & Cromwell be retained.

M1.2. With respect to interest by the CFTC in the proposed transaction, would you agree that the same law firm advise on that issue as well?

M1.3. Makes sense to me.

M1.4. Also, what is Energydesk.com LImited?

M1.5. Sara

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 225550</parent_id>

<message_id> 225592</message_id>

<date_time> 1999-11-16 07:21:00</date_time>

<subject> JAPRO Gruppen Aktiebolag</subject>

<from name="Sara Shackleton" id="64528" address="Sara.Shackleton@ENRON.com" />

<to name="Mark Taylor" id="2378" address="Mark.Taylor@ENRON.com" />

<content>

M2.1. Per my voice mail.

M2.2. Let me know what you think.

M2.3. SS

</content>

</message>

<message>

<depth> 2</depth>

<parent_id> 225592</parent_id>

<message_id> 1187843</message_id>

<date_time> 1999-11-16 11:01:00</date_time>

<subject> JAPRO Gruppen Aktiebolag</subject>

<from name="Mark Taylor" id="2378" address="Mark.Taylor@ENRON.com" />

<to name="Sara Shackleton" id="64528" address="Sara.Shackleton@ENRON.com" />

<content>

M3.1. I think S&C is fine - they are helping us with CFTC issues related to online trading

M3.2. and Energy Desk.com seems somwhat related.

</content>

</message>

<message>

<depth> 3</depth>

<parent_id> 1187843</parent_id>

<message_id> 225598</message_id>

<date_time> 1999-11-17 08:24:00</date_time>

<subject> JAPRO Gruppen Aktiebolag</subject>

<from name="Sara Shackleton" id="64528" address="Sara.Shackleton@ENRON.com" />

<to name="Martin Rosell" id="76849" address="martin.rosell@enron.com" />

<content>

M4.1. Martin:

M4.2. Sorry for the log jam but I always thought that the law firm was the best idea.

M4.3. Call if you need assistance.

M4.4. Sara

</content>

</message>

</thread>

## A.7 Example Thread (ID: 127941)

<thread>

<thread_id>127941</thread_id>

<message>

<depth> 0</depth>

<parent_id></parent_id>

<message_id> 763333</message_id>

<date_time> 2001-10-05 14:59:07</date_time>

<subject> City of Glendale</subject>

<from name="Kim S Ward" id="18442" address="kward@enron.com" />

<to name="Sara Shackleton" id="64528" address="sshackl@enron.com" />

<content>

M1.1. Sara,

M1.2. Believe it or not, we are very close getting our signed ISDA from the City of Glendale.

M1.3. Steve Lins, the City attorney had a couple of questions which I will attempt to relay without having a copy of the documents.

M1.4. 1) I am assuming that he obtained a for legal opinion letter or document of some sort.

M1.5. This document references a confirmation and we are not sure what this references.

M1.6. Typically, it references a transaction, which in this case, there are no transactions yet.

M1.7. He feels this reference should be deleted.

M1.8. What is your opinion regarding this?

M1.9. 2) We sent him a couple of form documents to facilitate the documents required under the ISDA.

M1.10. One form ws a form resolution.

M1.11. They have already received City Council approval to enter into financial tranactions and to enter into an ISDA with us.

M1.12. Steve is going to get a certified copy of this Resolution.

M1.13. Will this suffice?

M1.14. When you return, I may try to do one last conference call to alleviate any unanswered questions.

M1.15. I think we will have an executed ISDA with the City of Glendale by the end of next week.

M1.16. I am going to be out there meeting with them on Wednesday.

M1.17. Thanks for your help,

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 763333</parent_id>

<message_id> 874438</message_id>

<date_time> 2001-10-08 09:02:56</date_time>

<subject> City of Glendale</subject>

<from name="Sara Shackleton" id="64528" address="Sara.Shackleton@ENRON.com" />

<to name="Kim S Ward" id="18442" address="Kim.Ward@ENRON.com" />

<to name="Kim S Ward" id="18442" address="kward@enron.com" />

<cc name="Marie Heard" id="40104" address="Marie.Heard@ENRON.com" />

<cc name="Marie Heard" id="40104" address="Mheard@ENRON.com" />

<content>

M2.1. Kim:

M2.2. Can you obtain the name of Glendale's bond counsel (lawyer's name, phone number, email, etc.)?

M2.3. Thanks.

M2.4. SS

</content>

</message>

<message>

<depth> 2</depth>

<parent_id> 874438</parent_id>

<message_id> 763334</message_id>

<date_time> 2001-10-08 09:26:50</date_time>

<subject> City of Glendale</subject>

<from name="Kim S Ward" id="18442" address="kward@enron.com" />

<to name="Sara Shackleton" id="64528" address="sshackl@enron.com" />

<content>

M3.1. Glendale's City Attorney is Steve Lins.

M3.2. His phone number is 818-548-2080 and his email is slins@ci.glendale.ca.us.

M3.3. Please let me know if you need anything else.

M3.4. I will be in their offices on Wednesday.

</content>

</message>

<message>

<depth> 2</depth>

<parent_id> 874438</parent_id>

<message_id> 874028</message_id>

<date_time> 2001-10-08 10:15:27</date_time>

<subject> City of Glendale</subject>

<from name="Marie Heard" id="40104" address="Mheard@ENRON.com" />

<to name="Sara Shackleton" id="64528" address="sshackl@enron.com" />

<content>

M4.1. Sara:

M4.2. I do not see a copy of an opinion in the file nor have we received one since I sent the execution copies of the ISDA to Steve Lins.

</content>

</message>

<message>

<depth> 2</depth>

<parent_id> 874438</parent_id>

<message_id> 763337</message_id>

<date_time> 2001-10-08 16:18:22</date_time>

<subject> City of Glendale</subject>

<from name="Kim S Ward" id="18442" address="kward@enron.com" />

<to name="slins@ci.glendale.ca.us" id="106187" address="slins@ci.glendale.ca.us" />

<content>

M5.1. Steve,

M5.2. could you provide the name, phone number, etc. of your bond council for our attorney, Sara Shackleton?

M5.3. Thanks,

</content>

</message>

<message>

<depth> 3</depth>

<parent_id> 763334</parent_id>

<message_id> 1117626</message_id>

<date_time> 2001-10-08 09:27:29</date_time>

<subject> City of Glendale</subject>

<from name="Sara Shackleton" id="64528" address="sshackl@enron.com" />

<to name="Kim S Ward" id="18442" address="kward@enron.com" />

<content>

M6.1. I need the city's bond counsel (outside counsel).

</content>

</message>

<message>

<depth> 4</depth>

<parent_id> 1117626</parent_id>

<message_id> 763335</message_id>

<date_time> 2001-10-08 10:03:53</date_time>

<subject> City of Glendale</subject>

<from name="Kim S Ward" id="18442" address="kward@enron.com" />

<to name="Sara Shackleton" id="64528" address="sshackl@enron.com" />

<content>

M7.1. Is this to obtain outside opinion?

M7.2. I thought we were going to do that at our own expense.

</content>

</message>

<message>

<depth> 5</depth>

<parent_id> 763335</parent_id>

<message_id> 763443</message_id>

<date_time> 2001-10-08 10:38:46</date_time>

<subject> City of Glendale</subject>

<from name="Sara Shackleton" id="64528" address="Sara.Shackleton@ENRON.com" />

<to name="Kim S Ward" id="18442" address="Kim.Ward@ENRON.com" />

<to name="Kim S Ward" id="18442" address="kward@enron.com" />

<content>

M8.1. We are going to do this at our own expense.

M8.2. But we would like to hire Glendale's bond counsel.

M8.3. I don't know the name of Glendale's bond counsel or how to get in touch with them.

</content>

</message>

<message>

<depth> 6</depth>

<parent_id> 763443</parent_id>

<message_id> 763336</message_id>

<date_time> 2001-10-08 11:43:20</date_time>

<subject> City of Glendale</subject>

<from name="Kim S Ward" id="18442" address="kward@enron.com" />

<to name="Sara Shackleton" id="64528" address="sshackl@enron.com" />

<content>

M9.1. I will work on this for you - and will be in touch.

M9.2. Thanks!

</content>

</message>

</thread>

## A.8 Example Thread (ID: 129424)

<thread>

<thread_id>129424</thread_id>

<message>

<depth> 0</depth>

<parent_id></parent_id>

<message_id> 763816</message_id>

<date_time> 2001-10-10 10:54:04</date_time>

<subject> Oasis Dairy Farms Judgement</subject>

<from name="Kevin Hyatt" id="66073" address="Kevin.Hyatt@ENRON.com" />

<to name="Debbie Moseley" id="30133" address="Debbie.Moseley@ENRON.com" />

<to name="Kimberly Watson" id="66827" address="Kimberly.Watson@ENRON.com" />

<to name="Paul Cherry" id="11446" address="Paul.Cherry@ENRON.com" />

<to name="Debbie Moseley" id="30133" address="Dmosele2@ENRON.com" />

<to name="Kimberly Watson" id="66827" address="kwatson@enron.com" />

<to name="Paul Cherry" id="11446" address="b9fbefe6-3a4fb3b6-862566d0-70ef78@ENRON.com" />

<cc name="Eric Gadd" id="2474" address="Eric.Gadd@ENRON.com" />
<cc name="Steven Harris" id="2333" address="Steven.Harris@ENRON.com" />
<cc name="Eric Gadd" id="2474" address="Egadd@ENRON.com" />
<cc name="Steven Harris" id="2333" address="sharris1@enron.com" />
<content>
M1.1. Earlier this year I requested that the Enron litigation unit file suit against Oasis Dairy for collection of unpaid transport bills on TW.
M1.2. The suit was filed in the fifth judicial district court in Chaves County, NM.
M1.3. On October 1, 2001, TW was granted summary judgement in the case by the court in the amount of $29,250.56 inclusive of back interest and attorney fees.
M1.4. This amount will continue to accrue interest at 8.75M1.5. If you would like a copy of the judgement, please let me know.
M1.6. kh
</content>
</message>
<message>
<depth> 1</depth>
<parent_id> 763816</parent_id>
<message_id> 764354</message_id>
<date_time> 2001-10-11 17:41:37</date_time>
<subject> Oasis Dairy Farms Judgement</subject>
<from name="Kimberly Watson" id="66827" address="kwatson@enron.com" />
<to name="Kevin Hyatt" id="66073" address="khyatt@enron.com" />
<content>
M2.1. Kevin,
M2.2. Thanks for keeping up with this.
M2.3. Kim :>)
</content>
</message>
<message>
<depth> 1</depth>
<parent_id> 763816</parent_id>
<message_id> 764355</message_id>
<date_time> 2001-10-11 17:41:55</date_time>
<subject> Oasis Dairy Farms Judgement</subject>
<from name="Kimberly Watson" id="66827" address="kwatson@enron.com" />
<to name="Lynn Blair" id="2243" address="lblair@enron.com" />

<content>
M3.1. Lynn,
M3.2. FYI,
M3.3. Kim.
</content>
</message>
<message>
<depth> 1</depth>
<parent_id> 763816</parent_id>
<message_id> 764417</message_id>
<date_time> 2001-10-21 16:32:48</date_time>
<subject> Oasis Dairy Farms Judgement</subject>
<from name="Kimberly Watson" id="66827" address="kwatson@enron.com" />
<to name="Jan Moore" id="52078" address="Jmoore3@ENRON.com" />
<to name="Tracy Geaccone" id="112633" address="tgeacco@enron.com" />
<content>
M4.1. FYI,
M4.2. Kim.
</content>
</message>
<message>
<depth> 2</depth>
<parent_id> 764355</parent_id>
<message_id> 772369</message_id>
<date_time> 2001-10-11 18:41:47</date_time>
<subject> Oasis Dairy Farms Judgement</subject>
<from name="Lynn Blair" id="2243" address="lblair@enron.com" />
<to name="Terry Kowalke" id="3344" address="tkowalk@enron.com" />
<to name="Richard Hanagriff" id="1560" address="Rhanagr@ENRON.com" />
<content>
M5.1. Terry and Richard,
M5.2. let's discuss.
M5.3. Thanks.
M5.4. Lynn
</content>
</message>
<message>

<depth> 2</depth>
<parent_id> 764417</parent_id>
<message_id> 764704</message_id>
<date_time> 2001-10-22 11:34:12</date_time>
<subject> Oasis Dairy Farms Judgement</subject>
<from name="Tracy Geaccone" id="112633" address="Tracy.Geaccone@ENRON.com" />
<to name="Kimberly Watson" id="66827" address="Kimberly.Watson@ENRON.com" />
<to name="Kimberly Watson" id="66827" address="kwatson@enron.com" />
<content>
M6.1. Any idea when or if we will receive the money?
</content>
</message>
<message>
<depth> 2</depth>
<parent_id></parent_id>
<message_id> 764423</message_id>
<date_time> 2001-10-22 11:53:48</date_time>
<subject> Oasis Dairy Farms Judgement</subject>
<from name="Kimberly Watson" id="66827" address="kwatson@enron.com" />
<to name="Kevin Hyatt" id="66073" address="khyatt@enron.com" />
<content>
M7.1. Kevin,
M7.2. Who in legal have you been working with?
M7.3. Tracy was wondering when we thought we would receive the money.
M7.4. Thanks, Kim.
</content>
</message>
<message>
<depth> 3</depth>
<parent_id></parent_id>
<message_id> 772373</message_id>
<date_time> 2001-10-12 10:54:27</date_time>
<subject> Oasis Dairy Farms Judgement</subject>
<from name="Lynn Blair" id="2243" address="lblair@enron.com" />
<to name="Richard Hanagriff" id="1560" address="Rhanagr@ENRON.com" />
<to name="Terry Kowalke" id="3344" address="tkowalk@enron.com" />
<content>

M8.1. So is there nothing for us to do?

M8.2. Thanks.

M8.3. Lynn

</content>

</message>

<message>

<depth> 3</depth>

<parent_id> 764423</parent_id>

<message_id> 763857</message_id>

<date_time> 2001-10-22 13:33:48</date_time>

<subject> Oasis Dairy Farms Judgement</subject>

<from name="Kevin Hyatt" id="66073" address="Kevin.Hyatt@ENRON.com" />

<to name="Kimberly Watson" id="66827" address="Kimberly.Watson@ENRON.com" />

<to name="Kimberly Watson" id="66827" address="kwatson@enron.com" />

<cc name="Tracy Geaccone" id="112633" address="Tracy.Geaccone@ENRON.com" />

<cc name="Tracy Geaccone" id="112633" address="tgeacco@enron.com" />

<content>

M9.1. Bonnie White is the attorney.

M9.2. I have already called her to find out what our collection process is (if any).

M9.3. I know some guys named Guido and Mario who'll be happy to do it for 50M9.4. I'll let you know what I find out.

M9.5. kh

</content>

</message>

</thread>

## A.9 Example Thread (ID: 142489)

<thread>

<thread_id>142489</thread_id>

<message>

<depth> 0</depth>

<parent_id></parent_id>

<message_id> 1118651</message_id>

<date_time> 2001-10-28 21:32:41</date_time>

<subject> Out of Office - Monday morning</subject>

<from name="Rick Dietz" id="4682" address="rick.dietz@kingwoodcable.com" />

<to name="Shelley Corman" id="102777" address="Shelley.Corman@ENRON.com" />

<to name="Shelley Corman" id="102777" address="scorman@enron.com" />

<cc name="Alma Carrillo" id="5389" address="Alma.Carrillo@ENRON.com" />

<cc name="Alma Carrillo" id="5389" address="Acarril@ENRON.com" />

<content>

M1.1. Shelley,

M1.2. I will be out of the office tomorrow morning participating in a charity golf tournament sponsored by EDS at South Shore Harbor.

M1.3. Mark Giglotti, Jeannie Licciardo, Don Stacy and I are playing together.

M1.4. I know the timing may be bad because of the financial project we have been preparing but the charitable contribution to play was quite generous and I do not want to back out of my commitment to the other team members.

M1.5. HOWEVER, IF YOU NEED ME FOR ANY REASON, PLEASE PAGE ME AT 1(800) 609-6967 OR CALL MY CELL PHONE AT (713) 569-4140.

M1.6. ALMA WILL ALSO KNOW HOW TO GET A HOLD OF ME.

M1.7. I will only be 45 minutes away and will be able to come straight into the office, if needed.

M1.8. I will check in during the day with Linda Trevino, as she will be sitting in for me at your staff meeting.

M1.9. Rick

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 1118651</parent_id>

<message_id> 1120078</message_id>

<date_time> 2001-10-28 22:18:04</date_time>

<subject> Out of Office - Monday morning</subject>

<from name="Shelley Corman" id="102777" address="scorman@enron.com" />

<to name="Rick Dietz" id="4682" address="rick.dietz@kingwoodcable.com" />

<content>

M2.1. Rick

M2.2. I have a number of contracts that the bankers want early tomorrow.

M2.3. I am assuming Elizabeth will be in to help me?

</content>

</message>

<message>

<depth> 2</depth>

<parent_id> 1120078</parent_id>

<message_id> 1118647</message_id>

<date_time> 2001-10-28 23:27:30</date_time>

<subject> Out of Office - Monday morning</subject>

<from name="Rick Dietz" id="4682" address="rick.dietz@kingwoodcable.com" />

<to name="Shelley Corman" id="102777" address="Shelley.Corman@ENRON.com" />

<to name="Shelley Corman" id="102777" address="scorman@enron.com" />

<content>

M3.1. Elizabeth will be in.

M3.2. Also, Linda Trevino will be in the office.

M3.3. All contracts are in Envision as well as in the file room on 39.

M3.4. If you need me to be there, just let me know.

</content>

</message>

</thread>

## A.10 Example Thread (ID: 44299)

<thread>

<thread_id>44299</thread_id>

<message>

<depth> 0</depth>

<parent_id></parent_id>

<message_id> 164151</message_id>

<date_time> 2000-11-13 15:08:00</date_time>

<subject> Revised Agenda for next TAR&L meeting</subject>

<from name="Wayne Gardner" id="118705" address="Wayne.Gardner@ENRON.com" />

<to name="Donald Lassere" id="32891" address="Donald.Lassere@ENRON.com" />

<to name="Sue Nord" id="2366" address="Sue.Nord@ENRON.com" />

<to name="Michelle Hicks" id="5999" address="Michelle.Hicks@ENRON.com" />

<to name="Cynthia Harkness" id="21327" address="Cynthia.Harkness@ENRON.com" />

<to name="Lara Leibman" id="1855" address="Lara.Leibman@ENRON.com" />

<to name="Jan Haizmann" id="52065" address="Jan.Haizmann@enron.com" />

<to name="Rajen Shah" id="93723" address="Rajen.Shah@ENRON.com" />

<to name="James Ginty" id="36177" address="James.Ginty@ENRON.com" />

<to name="Derenda Plunkett" id="2072" address="Derenda.Plunkett@ENRON.com" />

<to name="David Merrill" id="28949" address="David.Merrill@ENRON.com" />

<to name="Robbi Rossi" id="48848" address="Robbi.Rossi@ENRON.com" />

<to name="Alisa Christensen" id="7360" address="Alisa.Christensen@ENRON.com" />
<to name="Jeff Dasovich" id="27095" address="Jeff.Dasovich@ENRON.com" />
<content>
M1.1. 1. Based on discussions from the 11/8 meeting, step through a specific US/Japan trading example for bandwidth purchased and resold by the US trading desk, specifically determining what the US trading desk can and cannot do with respect to each piece of the international capacity segment
M1.2. (see attached file).
M1.3. 2. Time permitting, step through a US/Japan trading example for bandwidth purchased and resold by the Singapore trading desk, specifically determining what a hypothetical Singapore trading desk could and could not do with respect to each piece of the international capacity segment.
M1.4. 3. Develop a prioritized plan of action specifying steps, accountabilities, format, and timeline for each area of responsibility to work together to provide necessary input for traders for each of the following jurisdictions:
M1.5. Europe Japan Hong Kong Australia Singapore Taiwan Korea Brazil Mexico Argentina Chile Venezuela Colombia
M1.6. 4. Review and discuss Dave Merrill's note on Korea.
M1.7. 5. Time permitting, step through a specific US/Korea trading example as in point one above.
</content>
</message>
<message>
<depth> 1</depth>
<parent_id> 164151</parent_id>
<message_id> 164168</message_id>
<date_time> 2000-11-14 07:43:00</date_time>
<subject> Revised Agenda for next TAR&L meeting</subject>
<from name="Jeff Dasovich" id="27095" address="Jeff.Dasovich@ENRON.com" />
<to name="Wayne Gardner" id="118705" address="Wayne.Gardner@ENRON.com" />
<content>
M2.1. Greetings:
M2.2. Sorry to bother you with this, but I'm travelling, and if you could leave the call-in number for tomorrow's meeting on my voice mail, I'll be forever indebted.
M2.3. Thanks a bunch.
</content>
</message>
</thread>

## A.11 Example Thread (ID: 14653)

<thread>

<thread_id>14653</thread_id>

<message>

<depth> 0</depth>

<parent_id></parent_id>

<message_id> 975420</message_id>

<date_time> 2000-04-11 08:22:00</date_time>

<subject> Hi</subject>

<from name="Vince J Kaminski" id="63574" address="Vince.J.Kaminski@ENRON.com" />

<to name="Stinson Gibner" id="2437" address="Stinson.Gibner@ENRON.com" />

<content>

M1.1. Stinson,

M1.2. This is the person from UT I mentioned.

M1.3. He is not interested in the summer internship.

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 975420</parent_id>

<message_id> 975418</message_id>

<date_time> 2000-04-11 08:25:00</date_time>

<subject> Hi</subject>

<from name="Vince J Kaminski" id="63574" address="Vince.J.Kaminski@ENRON.com" />

<to name="" id="121516" address="zkhokher@mail.utexas.edu" />

<cc name="Vince J Kaminski" id="63574" address="Vince.J.Kaminski@ENRON.com" />

<cc name="Stinson Gibner" id="2437" address="Stinson.Gibner@ENRON.com" />

<content>

M2.1. Zeigham,

M2.2. We discussed two options (not necessarily mutually exclusive):

M2.3. 1. summer internship

M2.4. 2. full employment.

M2.5. Are you interested exclusively in full employment?

M2.6. I need the answer ASAP, as we are going to discuss the additional summer intern positions this afternoon.

M2.7. Vince

</content>

</message>

<message>

<depth> 2</depth>

<parent_id> 975418</parent_id>

<message_id> 975413</message_id>

<date_time> 2000-04-11 10:06:00</date_time>

<subject> Hi</subject>

<from name="" id="121516" address="zkhokher@mail.utexas.edu" />

<to name="Vince J Kaminski" id="63574" address="Vince.J.Kaminski@ENRON.com" />

<cc name="Stinson Gibner" id="2437" address="Stinson.Gibner@ENRON.com" />

<content>

M3.1. Vince:

M3.2. I think full time employment starting in about six months seems to be the best option.

M3.3. It is probably best to get my dissertation wrapped up before taking on additional commitments.

M3.4. Regards

M3.5. Zeigham

</content>

</message>

<message>

<depth> 3</depth>

<parent_id> 975413</parent_id>

<message_id> 975412</message_id>

<date_time> 2000-04-11 10:22:00</date_time>

<subject> Hi</subject>

<from name="Vince J Kaminski" id="63574" address="Vince.J.Kaminski@ENRON.com" />

<to name="Stinson Gibner" id="2437" address="Stinson.Gibner@ENRON.com" />

<content>

M4.1. FYI

M4.2. Vince

</content>

</message>

</thread>

## A.12 Example Thread (ID: 16449)

<thread>
<thread_id>16449</thread_id>
<message>
<depth> 0</depth>
<parent_id></parent_id>
<message_id> 1189383</message_id>
<date_time> 2000-05-02 08:20:00</date_time>
<subject> Energy Language</subject>
<from name="Kenneth M Raisler" id="93703" address="Raislerk@sullcrom.com" />
<to name="" id="45275" address="goetscrj@bp.com" />
<to name="" id="67753" address="kneenjm@bp.com" />
<to name="" id="78226" address="McAdammj@bp.com" />
<to name="Elaine Walsh" id="37418" address="Elaine@citizenspower.com" />
<to name="Cynthia Sandherr" id="1557" address="csandhe@enron.com" />
<to name="Jeffrey Keeler" id="20209" address="jkeeler@enron.com" />
<to name="Mark E Haedicke" id="19235" address="Mark.E.Haedicke@ENRON.com" />
<to name="Mark Taylor" id="2378" address="Mark.Taylor@ENRON.com" />
<to name="Laurie Ferber" id="70061" address="laurie.ferber@gs.com" />
<to name="" id="47012" address="hall2r@kochind.com" />
<to name="" id="69417" address="lanced@kochind.com" />
<to name="" id="119774" address="william.mccoy@msdw.com" />
<to name="Steven Kline" id="109060" address="Steven.Kline@pge-corp.com" />
<to name="" id="102567" address="Schindlg@phibro.com" />
<to name="" id="79902" address="mgoldstein@sempratrading.com" />
<content>
M1.1. Please give me your views ASAP.
M1.2. Thanks for the language.
M1.3. We will make sure that weather derivatives are included.
M1.4. We met with the CFTC yesterday and they oppose excluding energies from the Act.
M1.5. Our bill would require that these excluded products, if traded on an electronic trading system, to be subject to the antimanipulation authority of the CFTC and not allow these products to require delivery.
M1.6. One issue they raised was price discovery and that these markets are not very transparent.
M1.7. What would be your reaction if we added a third provision (again this is only for energies traded on a electronic trading facility–bilateral transactions would not be subject to these provisions) that required some sort of price disclosure for these markets.

M1.8. That might help placate the CFTC, but I'm not sure what the industry reaction might be.

M1.9. Let me know.

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 1189383</parent_id>

<message_id> 1189394</message_id>

<date_time> 2000-05-03 05:46:00</date_time>

<subject> Energy Language</subject>

<from name="Kenneth M Raisler" id="93703" address="Raislerk@sullcrom.com" />

<to name="" id="45275" address="goetscrj@bp.com" />

<to name="" id="67753" address="kneenjm@bp.com" />

<to name="" id="78226" address="McAdammj@bp.com" />

<to name="Elaine Walsh" id="37418" address="Elaine@citizenspower.com" />

<to name="Cynthia Sandherr" id="1557" address="csandhe@enron.com" />

<to name="Jeffrey Keeler" id="20209" address="jkeeler@enron.com" />

<to name="Mark E Haedicke" id="19235" address="Mark.E.Haedicke@ENRON.com" />

<to name="Mark Taylor" id="2378" address="Mark.Taylor@ENRON.com" />

<to name="Laurie Ferber" id="70061" address="laurie.ferber@gs.com" />

<to name="" id="47012" address="hall2r@kochind.com" />

<to name="" id="69417" address="lanced@kochind.com" />

<to name="" id="119774" address="william.mccoy@msdw.com" />

<to name="Steven Kline" id="109060" address="Steven.Kline@pge-corp.com" />

<to name="" id="102567" address="Schindlg@phibro.com" />

<to name="" id="79902" address="mgoldstein@sempratrading.com" />

<content>

M2.1. I have conferred with industry representatives on the CFTC's suggestion.

M2.2. We have a problem with it from a couple of perspectives:

M2.3. 1. Although NYMEX is the benchmark for pricing of a few energy commodities, most of the pricing of transactions is done based on price reporting services such as Platt's, Megawatt Daily and Reuters.

M2.4. These services collect information on transactions from industry representatives and report usually on a daily basis benchmark prices for a large number of energy commodities.

M2.5. This activity has never been regulated by anyone.

M2.6. Obviously, these price reporting services have provided valuable price information to the industry.

M2.7. We do not see a need for regulation of price reporting whether it from a price reporting service or an electronic trading system.

M2.8. 2. We are concerned that the manner in which the CFTC would regulate/oversee price reporting would be very awkward and difficult.
M2.9. The drafting of such a provision would be complex and the discretion it would give the CFTC to potentially regulate through price reporting would be troubling.
M2.10. We are very interested in achieving the exclusion that the draft legislation currently provides.
M2.11. We would be prepared to discuss the matter further in the hope of maintaining the exclusion.
M2.12. Please let me know what you think.
</content>
</message>
</thread>

## A.13 Example Thread (ID: 188078)

<thread>
<thread_id>188078</thread_id>
<message>
<depth> 0</depth>
<parent_id></parent_id>
<message_id> 43417</message_id>
<date_time> 2002-04-02 10:06:52</date_time>
<subject> Enron Compressor Services</subject>
<from name="Chris Germany" id="3471" address="cgerman@enron.com" />
<to name="Kay Mann" id="239" address="kmann@enron.com" />
<to name="Mark Knippa" id="76124" address="Mknippa@ENRON.com" />
<cc name="Jack Wise" id="837" address="Jwise@ENRON.com" />
<cc name="Ed McMichael Jr" id="1599" address="Emcmich@ENRON.com" />
<cc name="Ruth Concannon" id="24606" address="rconcan@enron.com" />
<cc name="Sabra L Dinari" id="101073" address="Sdinari@ENRON.com" />
<cc name="Scott Mills" id="1704" address="Smills@ENRON.com" />
<cc name="Torrey Moorer" id="1990" address="Gcouch@ENRON.com" />
<content>
M1.1. I sold 3,253 dth per day of Florida Zone 3 gas at $3.60 to Bob Crites (713-420-2499) at El Paso effective 4/3/02 - 4/30/02.
M1.2. I'm leaving this deal out of Sitara because its an Enron Compressor Services deal, not ENA.
M1.3. Kay,
M1.4. I need to send a GISB agreement as Enron Compressor Services to El Paso.

M1.5. Would you work on that?

M1.6. Mark,

M1.7. I'm not sure who at Enron to notify about this deal.

M1.8. You and I can chat about that later.

M1.9. I only asked for 2 bids for the April gas.

M1.10. Reliant showed me a bid of $3.58 and El Paso shoed me a bid of $3.60.

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 43417</parent_id>

<message_id> 42000</message_id>

<date_time> 2002-04-02 10:17:57</date_time>

<subject> Enron Compressor Services</subject>

<from name="Chris Germany" id="3471" address="cgerman@enron.com" />

<to name="Chris Germany" id="3471" address="cgerman@enron.com" />

<to name="Kay Mann" id="239" address="kmann@enron.com" />

<to name="Mark Knippa" id="76124" address="Mknippa@ENRON.com" />

<to name="Jim Coffey Jr" id="53570" address="Jcoffey@ENRON.com" />

<cc name="Jack Wise" id="837" address="Jwise@ENRON.com" />

<cc name="Ed McMichael Jr" id="1599" address="Emcmich@ENRON.com" />

<cc name="Ruth Concannon" id="24606" address="rconcan@enron.com" />

<cc name="Sabra L Dinari" id="101073" address="Sdinari@ENRON.com" />

<cc name="Scott Mills" id="1704" address="Smills@ENRON.com" />

<cc name="Torrey Moorer" id="1990" address="Gcouch@ENRON.com" />

<cc name="Troy Denetsosie" id="111779" address="Tdenets@ENRON.com" />

<content>

M2.1. Mark,

M2.2. per your request I've copied Jim Coffey and Troy Denetsosie.

M2.3. I will work with legal and El Paso to set up a GISB agreement.

</content>

</message>

<message>

<depth> 2</depth>

<parent_id> 42000</parent_id>

<message_id> 41992</message_id>

<date_time> 2002-04-03 10:39:02</date_time>

<subject> Enron Compressor Services</subject>

<from name="Sabra L Dinari" id="101073" address="Sabra.L.Dinari@ENRON.com" />

<to name="Chris Germany" id="3471" address="Chris.Germany@ENRON.com" />

<to name="Chris Germany" id="3471" address="cgerman@enron.com" />

<content>

M3.1. Do I need to do something about this?

M3.2. Who is nominating this or where did you get the gas, is it a buy/sell?

M3.3. Just curious...I was out yesterday.

</content>

</message>

</thread>

## A.14 Example Thread (ID: 42685)

<thread>

<thread_id>42685</thread_id>

<message>

<depth> 0</depth>

<parent_id></parent_id>

<message_id> 163901</message_id>

<date_time> 2000-11-07 09:28:00</date_time>

<subject> Meet</subject>

<from name="Jacqueline Kelly" id="51295" address="JKelly@FairIsaac.com" />

<to name="Jeff Dasovich" id="27095" address="Jeff.Dasovich@ENRON.com" />

<to name="Jacqueline Kelly" id="51295" address="JKelly@FairIsaac.com" />

<to name="Kimberly Kupiecki" id="67317" address="kkupiecki@arpartners.com" />

<to name="Ted Chin" id="112109" address="tedchin@hotmail.com" />

<content>

M1.1. I agree that we need to get together.

M1.2. I am going to grind out a spreadsheet tomorrow night.

M1.3. Unfortunately, my friends birthday dinner is Thursday night.

M1.4. I know Ted is going to be gone for the weekend.

M1.5. Can we get together tomorrow or over the weekend without Ted?

M1.6. Ted is going to do some research on the discount rate, which can be dropped in to any analysis that we come up with.

M1.7. What do you think?

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 163901</parent_id>

<message_id> 163898</message_id>

<date_time> 2000-11-07 11:20:00</date_time>

<subject> Meet</subject>

<from name="Jeff Dasovich" id="27095" address="Jeff.Dasovich@ENRON.com" />

<to name="Jacqueline Kelly" id="51295" address="JKelly@FairIsaac.com" />

<to name="Kimberly Kupiecki" id="67317" address="kkupiecki@arpartners.com" />

<to name="Ted Chin" id="112109" address="tedchin@hotmail.com" />

<content>

M2.1. Hey, I know it's a pain, but I think there would be value in getting together (if folks are available) on Thursday evening from 7-10 with (lots of) beer and pizza and grind through the finance case.

M2.2. We can do it at my apartment, or anywhere else you folks would like to do it.

M2.3. Thoughts?

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 163901</parent_id>

<message_id> 163906</message_id>

<date_time> 2000-11-07 12:43:00</date_time>

<subject> Meet</subject>

<from name="Jeff Dasovich" id="27095" address="Jeff.Dasovich@ENRON.com" />

<to name="Jacqueline Kelly" id="51295" address="JKelly@FairIsaac.com" />

<cc name="Jacqueline Kelly" id="51295" address="JKelly@FairIsaac.com" />

<cc name="Kimberly Kupiecki" id="67317" address="kkupiecki@arpartners.com" />

<cc name="Ted Chin" id="112109" address="tedchin@hotmail.com" />

<content>

M3.1. as far as i'm concerned ted' off the team, period. (we kid!)

M3.2. Tomorrow's fine with me.

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 163901</parent_id>

<message_id> 163965</message_id>

<date_time> 2000-11-08 09:59:00</date_time>

<subject> Meet</subject>

<from name="Jeff Dasovich" id="27095" address="Jeff.Dasovich@ENRON.com" />

<to name="Jacqueline Kelly" id="51295" address="JKelly@FairIsaac.com" />

<cc name="Kimberly Kupiecki" id="67317" address="kkupiecki@arpartners.com" />

<cc name="Ted Chin" id="112109" address="tedchin@hotmail.com" />

<content>

M4.1. Hey: we meeting tonite?

M4.2. can you BELIEVE this bloody election?

</content>

</message>

</thread>

## A.15 Example Thread (ID: 2876)

<thread>

<thread_id>2876</thread_id>

<message>

<depth> 0</depth>

<parent_id></parent_id>

<message_id> 68102</message_id>

<date_time> 2002-01-17 13:06:59</date_time>

<subject></subject>

<from name="Michelle Nelson" id="81118" address="Michelle.Nelson@ENRON.com" />

<to name="Mike Maggi" id="3774" address="Mike.Maggi@ENRON.com" />

<to name="Mike Maggi" id="3774" address="mmaggi@enron.com" />

<content>

M1.1. come down here and visit me

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 68102</parent_id>

<message_id> 2001194</message_id>

<date_time> 2002-01-17 13:10:17</date_time>

<subject></subject>

<from id="" name="" address="" />

<content>

M2.1. i have to go back into lavo's office in a few minutes, then i have to go to lunch with john and one of our research guys from chicago,

M2.2. i will be back around 1:30

</content>

</message>

<message>

<depth> 2</depth>

<parent_id> 2001194</parent_id>

<message_id> 68100</message_id>

<date_time> 2002-01-17 13:13:35</date_time>

<subject></subject>

<from name="Michelle Nelson" id="81118" address="Michelle.Nelson@ENRON.com" />

<to name="Mike Maggi" id="3774" address="Mike.Maggi@ENRON.com" />

<to name="Mike Maggi" id="3774" address="mmaggi@enron.com" />

<content>

M3.1. that's not acceptable.

M3.2. come play now.

M3.3. ok so where can laura and i go?

</content>

</message>

<message>

<depth> 3</depth>

<parent_id> 68100</parent_id>

<message_id> 2001195</message_id>

<date_time> 2002-01-17 13:14:09</date_time>

<subject></subject>

<from id="" name="" address="" />

<content>

M4.1. dinking, i will meet you around 2:30

</content>

</message>

<message>

<depth> 4</depth>

<parent_id> 2001195</parent_id>

<message_id> 68098</message_id>

<date_time> 2002-01-17 13:14:44</date_time>

<subject></subject>

<from name="Michelle Nelson" id="81118" address="Michelle.Nelson@ENRON.com" />

<to name="Mike Maggi" id="3774" address="Mike.Maggi@ENRON.com" />

<to name="Mike Maggi" id="3774" address="mmaggi@enron.com" />

<content>

M5.1. dinking? do you mean drinking?

M5.2. i know, but where

</content>

</message>

<message>

<depth> 4</depth>

<parent_id> 2001195</parent_id>

<message_id> 68097</message_id>

<date_time> 2002-01-17 13:17:33</date_time>

<subject></subject>

<from name="Michelle Nelson" id="81118" address="Michelle.Nelson@ENRON.com" />

<to name="Mike Maggi" id="3774" address="Mike.Maggi@ENRON.com" />

<to name="Mike Maggi" id="3774" address="mmaggi@enron.com" />

<content>

M6.1. i think that we are going to go workout and then we can all go somewhere.

M6.2. is that ok?

</content>

</message>

<message>

<depth> 5</depth>

<parent_id> 68097</parent_id>

<message_id> 2001196</message_id>

<date_time> 2002-01-17 13:19:33</date_time>

<subject></subject>

<from id="" name="" address="" />

<content>

M7.1. i guess so

</content>

</message>

<message>
<depth> 5</depth>
<parent_id> 68097</parent_id>
<message_id> 2001197</message_id>
<date_time> 2002-01-17 13:19:05</date_time>
<subject></subject>
<from id="" name="" address="" />
<content>
M8.1. i guess so
</content>
</message>
<message>
<depth> 6</depth>
<parent_id> 2001196</parent_id>
<message_id> 68094</message_id>
<date_time> 2002-01-17 13:21:34</date_time>
<subject></subject>
<from name="Michelle Nelson" id="81118" address="Michelle.Nelson@ENRON.com" />
<to name="Mike Maggi" id="3774" address="Mike.Maggi@ENRON.com" />
<to name="Mike Maggi" id="3774" address="mmaggi@enron.com" />
<content>
M9.1. ok call me on my cell later.
M9.2. i think that we are going to go shopping and then we will meet you somewhere to drink.
</content>
</message>
<message>
<depth> 6</depth>
<parent_id> 2001197</parent_id>
<message_id> 68095</message_id>
<date_time> 2002-01-17 13:20:37</date_time>
<subject></subject>
<from name="Michelle Nelson" id="81118" address="Michelle.Nelson@ENRON.com" />
<to name="Mike Maggi" id="3774" address="Mike.Maggi@ENRON.com" />
<to name="Mike Maggi" id="3774" address="mmaggi@enron.com" />
<content>
M10.1. you are annoying me.
M10.2. what do you want to do?

</content>

</message>

</thread>

## A.16 Example Thread (ID: 88)

<thread>

<thread_id>88</thread_id>

<message>

<depth> 0</depth>

<parent_id> NULL</parent_id>

<message_id> 292368</message_id>

<date_time> 2001-05-24 04:57:13</date_time>

<subject> Recommended Offer by ICE for IPE Holdings</subject>

<from name="Justin Boyd" id="890" address="justin.boyd@enron.com" />

<to name="Greg Whalley" id="27104" address="Greg.Whalley@ENRON.com" />

<to name="John Sherriff" id="4671" address="john.sherriff@enron.com" />

<to name="Michael Brown" id="2202" address="michael.r.brown@enron.com" />

<to name="Greg Piper" id="1665" address="Greg.Piper@ENRON.com" />

<to name="Andy Zipper" id="5063" address="Andy.Zipper@ENRON.com" />

<to name="Greg Whalley" id="27104" address="DA82494B-FC99BFE6-862565DA-5FA576@ENRON.com" />

<to name="John Sherriff" id="4671" address="A177817E-7D75C390-8625653F-6307BB@ENRON.com" />

<to name="Michael Brown" id="2202" address="mbrown3@ENRON.com" />

<to name="Greg Piper" id="1665" address="Gpiper@ENRON.com" />

<to name="Andy Zipper" id="5063" address="azipper@enron.com" />

<content>

M1.1. This is to point out to you the approaching deadline for accepting ICE's Offer for IPE Holdings.

M1.2. If we wish to accept this Offer, we must do so no later than 3 pm on Tuesday 29 May.

M1.3. In the meantime, Michael has received calls from Richard Ward (IPE's CEO) inquiring as to our position,

M1.4. and it would be my view that we should accept the Offer.

M1.5. (Note that if ICE receives acceptances of 90M1.6. And assuming that such level of acceptances will be received, there is no reason for not accepting the Offer).

M1.7. Please would you let me know whether you wish to accept the Offer.

M1.8. Ideally, if we are to do so, I would plan to send the acceptance by close of business on Friday 25 May.

M1.9. Thanks

M1.10. Justin

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 292368</parent_id>

<message_id> 292145</message_id>

<date_time> 2001-05-24 10:27:54</date_time>

<subject> Recommended Offer by ICE for IPE Holdings</subject>

<from name="Greg Piper" id="1665" address="Greg.Piper@ENRON.com" />

<to name="Andy Zipper" id="5063" address="Andy.Zipper@ENRON.com" />

<to name="Andy Zipper" id="5063" address="azipper@enron.com" />

<cc name="Justin Boyd" id="890" address="justin.boyd@enron.com" />

<cc name="Michael Brown" id="2202" address="michael.r.brown@enron.com" />

<cc name="Justin Boyd" id="890" address="jboyd@ENRON.com" />

<cc name="Michael Brown" id="2202" address="mbrown3@ENRON.com" />

<content>

M2.1. Your thoughts on this?

M2.2. Also, assuming we accept, then what?

M2.3. Are we an owner in ICE?

M2.4. GP

</content>

</message>

<message>

<depth> 2</depth>

<parent_id> 292145</parent_id>

<message_id> 292367</message_id>

<date_time> 2001-05-24 11:12:36</date_time>

<subject> Recommended Offer by ICE for IPE Holdings</subject>

<from name="Justin Boyd" id="890" address="justin.boyd@enron.com" />

<to name="Greg Piper" id="1665" address="Greg.Piper@ENRON.com" />

<to name="Andy Zipper" id="5063" address="Andy.Zipper@ENRON.com" />

<to name="Greg Piper" id="1665" address="Gpiper@ENRON.com" />

<to name="Andy Zipper" id="5063" address="azipper@enron.com" />

<cc name="Michael Brown" id="2202" address="michael.r.brown@enron.com" />

<cc name="Michael Brown" id="2202" address="mbrown3@ENRON.com" />

<content>

M3.1. Greg, Andy

M3.2. If we accept the Offer, and assuming the Offer receives sufficient acceptances from the other IPE shareholders, then we would hold equity in ICE

M3.3. (in the form of A and B Shares)

M3.4. Justin

</content>

</message>

<message>

<depth> 3</depth>

<parent_id> NULL</parent_id>

<message_id> 292207</message_id>

<date_time> 2001-05-24 11:02:08</date_time>

<subject> Recommended Offer by ICE for IPE Holdings</subject>

<from name="Greg Piper" id="1665" address="Greg.Piper@ENRON.com" />

<to name="Andy Zipper" id="5063" address="Andy.Zipper@ENRON.com" />

<to name="Andy Zipper" id="5063" address="azipper@enron.com" />

<content>

M4.1. OK, so how much do we own

M4.2. and what rights will we have

M4.3. and do we do anything with it?

M4.4. GP

</content>

</message>

<message>

<depth> 3</depth>

<parent_id> 292367</parent_id>

<message_id> 2006628</message_id>

<date_time> 2001-05-24 12:27:00</date_time>

<subject> Recommended Offer by ICE for IPE Holdings</subject>

<from id="" name="" address="" />

<content>

M5.1. How much would we own and how much control would we have, if any?

M5.2. Thanks.

M5.3. GP

</content>

</message>

<message>

<depth> 4</depth>

<parent_id> 2006628</parent_id>

<message_id> 2006543</message_id>

<date_time> 2001-05-24 13:41:24</date_time>

<subject> Recommended Offer by ICE for IPE Holdings</subject>

<from id="" name="" address="" />

<content>

M6.1. Greg

M6.2. We would hold less than 1M6.3. so no control at all

M6.4. Justin

</content>

</message>

<message>

<depth> 5</depth>

<parent_id> 2006543</parent_id>

<message_id> 292205</message_id>

<date_time> 2001-05-24 14:55:49</date_time>

<subject> Recommended Offer by ICE for IPE Holdings</subject>

<from name="Greg Piper" id="1665" address="Greg.Piper@ENRON.com" />

<to name="Andy Zipper" id="5063" address="Andy.Zipper@ENRON.com" />

<to name="Andy Zipper" id="5063" address="azipper@enron.com" />

<content>

M7.1. Does our less than 1M7.2. GP

</content>

</message>

</thread>

## A.17 Example Thread (ID: 102)

<thread>

<thread_id>102</thread_id>

<message>

<depth> 0</depth>

<parent_id> NULL</parent_id>

<message_id> 277074</message_id>

<date_time> 2001-04-24 11:53:12</date_time>

<subject> March 2001 Invoice</subject>

<from name="Julie Meyers" id="4832" address="Julie.Meyers@ENRON.com" />

<to name="Tess Ray" id="49473" address="tess.ray@ENRON.com" />

<to name="Tess Ray" id="49473" address="TRAY2@ENRON.com" />

<cc name="Daren J Farmer" id="28042" address="Daren.J.Farmer@ENRON.com" />

<cc name="Liz Bellamy" id="70263" address="liz.Bellamy@ENRON.com" />

<cc name="farmer" id="40461" address="Farmer@ENRON.com" />

<cc name="" id="3985" address="??SDaren.J.Farmer@ENRON.com" />

<cc name="Liz Bellamy" id="70263" address="lbellamy@enron.com" />

<content>

M1.1. Tess there is a deal out there for the blue dolphin S#745589.

M1.2. And there are three deal out there for CSGT #639612, #639615, #745589.

M1.3. The problem is that none of these deals have actuals.

M1.4. It looks as though they have not been nom'd.

M1.5. Daren are these deals real or what?

M1.6. Why have they not been nom'd?

M1.7. Could they be under another Dow company?

M1.8. Tess, I'm leaving the office for a little bit today.

M1.9. But I'll be back this afternoon.

M1.10. Julie

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 277074</parent_id>

<message_id> 277071</message_id>

<date_time> 2001-04-26 10:42:36</date_time>

<subject> March 2001 Invoice</subject>

<from name="Tess Ray" id="49473" address="tess.ray@ENRON.com" />

<to name="Daren J Farmer" id="28042" address="Daren.J.Farmer@ENRON.com" />

<to name="Daren J Farmer" id="28042" address="dfarmer@enron.com" />

<cc name="Julie Meyers" id="4832" address="Julie.Meyers@ENRON.com" />

<cc name="Julie Meyers" id="4832" address="A8131D50-2229AC72-862564B4-7573B7@ENRON.com" />

<content>

M2.1. Daren -

M2.2. Dow Hydrocarbons and Resources, Inc., stated on 4/24/01, re: ENA deal # SA 639615, that their 03/01 price for ENA sales on CSGT @ B368-Brazos # 368 (i.e. cowtrap), is IF - $.06.

M2.3. We are invoicing them at HSC GDP DA.

M2.4. Need to know which price is correct and copy of confirmation.

</content>

</message>

<message>

<depth> 2</depth>

<parent_id> 277071</parent_id>

<message_id> 2000032</message_id>

<date_time> 2001-04-26 13:34:28</date_time>

<subject> March 2001 Invoice</subject>

<from id="" name="" address="" />

<content>

M3.1. Tess,

M3.2. We probably should be invoicing selling to Dow at IF.

M3.3. However, we should also have purchased the supply from Spinnaker (#144271 & 144264) at IF.

M3.4. Did Spinnaker bill us at Index or Gas Daily?

M3.5. What did we pay?

M3.6. Let me know this and we can proceed from there.

M3.7. Nelson Ferris did these deals.

M3.8. I will talk to him after I hear from you.

M3.9. D

</content>

</message>

<message>

<depth> 3</depth>

<parent_id> 2000032</parent_id>

<message_id> 277072</message_id>

<date_time> 2001-04-26 16:26:20</date_time>

<subject> March 2001 Invoice</subject>

<from name="Tess Ray" id="49473" address="tess.ray@ENRON.com" />

<to name="Charlene Richmond" id="3793" address="Charlene.Richmond@ENRON.com" />

<to name="Cynthia Hakemack" id="26663" address="Cynthia.Hakemack@ENRON.com" />

<to name="Charles Howard" id="922" address="Charles.Howard2@ENRON.com" />

<to name="Charlene Richmond" id="3793" address="9eb4b4f6-a07b83a3-862564a5-6b5450@ENRON.com" />

<to name="Cynthia Hakemack" id="26663" address="F83DF2BD-9B262624-86256500-6C4F0E@ENRON.com" />

<to name="Charles Howard" id="922" address="2ae91b5b-41bd5f60-8625698a-6f4c3c@ENRON.com" />

<cc name="Daren J Farmer" id="28042" address="Daren.J.Farmer@ENRON.com" />

<cc name="Daren J Farmer" id="28042" address="dfarmer@enron.com" />

<content>

M4.1. Charlene/Cindy/Charles:

M4.2. Can either of you answer Darren's question to me below?

M4.3. My sales deal is under ENA.

M4.4. Thanks,

M4.5. Tess

</content>

</message>

<message>

<depth> 4</depth>

<parent_id> 277072</parent_id>

<message_id> 277073</message_id>

<date_time> 2001-04-26 16:48:46</date_time>

<subject> March 2001 Invoice</subject>

<from name="Tess Ray" id="49473" address="tess.ray@ENRON.com" />

<to name="Mary Ellenberger" id="37728" address="Mary.Ellenberger@ENRON.com" />

<to name="Mary Ellenberger" id="595" address="1d1ab0b7-ea9cbefb-8625687e-777977@ENRON.com" />

<cc name="Daren J Farmer" id="28042" address="Daren.J.Farmer@ENRON.com" />

<cc name="Daren J Farmer" id="28042" address="dfarmer@enron.com" />

<content>

M5.1. Mary -

M5.2. Do you pay Spinnaker, for gas purchases?

M5.3. (See Daren's question below, re: 03/01 purchase price from Spinnaker (#144271 & 144264),

M5.4. Need to know if price is at IF or Gas Daily?

M5.5. Thanks,

M5.6. Tess

</content>

</message>

</thread>

## A.18 Example Thread (ID: 101)

<thread>

<thread_id>101</thread_id>

<message>

<depth> 0</depth>

<parent_id> NULL</parent_id>

<message_id> 44503</message_id>

<date_time> 2001-05-25 16:56:21</date_time>

<subject> LT Ercot Schedule C</subject>

<from name="Stacey W White" id="2310" address="Stacey.W.White@ENRON.com" />

<to name="Doug Gilbert-smith" id="805" address="Doug.Gilbert-Smith@ENRON.com" />

<to name="Doug Gilbert-smith" id="805" address="dsmith3@enron.com" />

<content>

M1.1. Did you get this reserve cleared through Lavorato?

M1.2. If you have not, can you do so immediately.

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 44503</parent_id>

<message_id> 2000031</message_id>

<date_time> 2001-05-27 15:26:38</date_time>

<subject> LT Ercot Schedule C</subject>

<from id="" name="" address="" />

<content>

M2.1. Kevin,

M2.2. This is for the full requirements uncertainty associated with green mountain and includes a reserve against teh liabilitiy exposure for the QSE agreements.

M2.3. Please let me know if you have any other questions,

M2.4. Doug

</content>

</message>

<message>

<depth> 2</depth>

<parent_id> 2000031</parent_id>

<message_id> 1033889</message_id>

<date_time> 2001-05-29 13:56:55</date_time>

<subject> LT Ercot Schedule C</subject>

<from name="Kevin M Presto" id="19920" address="Kevin.M.Presto@ENRON.com" />

<to name="John Lavorato" id="2273" address="John.J.Lavorato@ENRON.com" />

<to name="John Lavorato" id="2273" address="JLAVORA@ENRON.com" />

<content>

M3.1. I need approval for this ERCOT Schedule C.

M3.2. The e-mail from Stacey outlines the details.

M3.3. Thanks.

</content>

</message>

</thread>

## A.19 Example Thread (ID: 108)

<thread>

<thread_id>108</thread_id>

<message>

<depth> 0</depth>

<parent_id> NULL</parent_id>

<message_id> 277075</message_id>

<date_time> 2001-04-26 17:44:31</date_time>

<subject> March 2001 Invoice</subject>

<from name="Mary Ellenberger" id="37728" address="Mary.Ellenberger@ENRON.com" />

<to name="Tess Ray" id="49473" address="tess.ray@ENRON.com" />

<to name="Tess Ray" id="49473" address="TRAY2@ENRON.com" />

<cc name="Daren J Farmer" id="28042" address="Daren.J.Farmer@ENRON.com" />

<cc name="Daren J Farmer" id="28042" address="dfarmer@enron.com" />

<content>

M1.1. The question is not that "cut and dry".

M1.2. However, for the month of March Enron paid Spinnaker @ IF HSC -$0.085.

M1.3. Currently the volume is posted under deal ticket #144271 with is the deal tick for gas daily production.

M1.4. This production should be moved to Deal ticket #144264 with is the IF.

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 277075</parent_id>

<message_id> 277076</message_id>

<date_time> 2001-04-27 08:47:38</date_time>

<subject> March 2001 Invoice</subject>

<from name="Tess Ray" id="49473" address="tess.ray@ENRON.com" />

<to name="Mary Ellenberger" id="37728" address="Mary.Ellenberger@ENRON.com" />

<to name="Mary Ellenberger" id="595" address="1d1ab0b7-ea9cbefb-8625687e-777977@ENRON.com" />

<cc name="Daren J Farmer" id="28042" address="Daren.J.Farmer@ENRON.com" />

<cc name="Julie Meyers" id="4832" address="Julie.Meyers@ENRON.com" />

<cc name="Daren J Farmer" id="28042" address="dfarmer@enron.com" />

<cc name="Julie Meyers" id="4832" address="A8131D50-2229AC72-862564B4-7573B7@ENRON.com" />

<content>

M2.1. Thanks Mary!

M2.2. Daren, is the info that you need?

</content>

</message>

<message>

<depth> 2</depth>

<parent_id> 277076</parent_id>

<message_id> 2000033</message_id>

<date_time> 2001-04-27 11:12:29</date_time>

<subject> March 2001 Invoice</subject>

<from id="" name="" address="" />

<content>

M3.1. Ok, We should be paying Dow based on the index price.

M3.2. But, it should be IF HSC - .07 instead of -.06.

M3.3. Sales volumes should be allocated to deal 639612.

M3.4. D

</content>

</message>

<message>

<depth> 3</depth>

<parent_id> 2000033</parent_id>

<message_id> 277077</message_id>

<date_time> 2001-04-27 13:37:20</date_time>

<subject> March 2001 Invoice</subject>

<from name="Tess Ray" id="49473" address="tess.ray@ENRON.com" />

<to name="Joyce Viltz" id="60904" address="joyce.viltz@ENRON.com" />

<to name="Julie Meyers" id="4832" address="Julie.Meyers@ENRON.com" />

<to name="Joyce Viltz" id="60904" address="JVILTZ@ENRON.com" />

<to name="Julie Meyers" id="4832" address="A8131D50-2229AC72-862564B4-7573B7@ENRON.com" />

<cc name="Daren J Farmer" id="28042" address="Daren.J.Farmer@ENRON.com" />

<cc name="Daren J Farmer" id="28042" address="dfarmer@enron.com" />

<content>

M4.1. Julie -

M4.2. Per Daren's repsonse below, can you correct price on this deal for 03/01.

M4.3. Currently under 639615.

M4.4. Joyce-

M4.5. Per Daren's message below, the sales volumes are currently under deal # SA639615.

M4.6. Let me know when they've been reallocated to 639612.

</content>

</message>

</thread>

## A.20 Example Thread (ID: 109)

<thread>

<thread_id>109</thread_id>

<message>

<depth> 0</depth>

<parent_id> NULL</parent_id>

<message_id> 1033851</message_id>

<date_time> 2001-05-29 08:40:57</date_time>

<subject> Formosa - 1.25 Million</subject>

<from name="Jeffrey C Gossett" id="45450" address="Jeffrey.C.Gossett@ENRON.com" />

<to name="John Lavorato" id="2273" address="John.J.Lavorato@ENRON.com" />

<to name="John Lavorato" id="2273" address="JLAVORA@ENRON.com" />

<cc name="Jean Mrha" id="1994" address="Jean.Mrha@ENRON.com" />

<cc name="Jean Mrha" id="1994" address="Jmrha@ENRON.com" />

<content>

M1.1. John -

M1.2. Global Markets is maintaining that there are "accounting issues" with the $1.25 million.

M1.3. Should we get Faith Killen or Wes to get this wrapped up?

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 1033851</parent_id>

<message_id> 1033848</message_id>

<date_time> 2001-05-29 10:18:39</date_time>

<subject> Formosa - 1.25 Million</subject>

<from name="Jean Mrha" id="1994" address="Jean.Mrha@ENRON.com" />

<to name="John Lavorato" id="2273" address="John.J.Lavorato@ENRON.com" />

<to name="John Lavorato" id="2273" address="JLAVORA@ENRON.com" />

<cc name="Jeffrey C Gossett" id="45450" address="Jeffrey.C.Gossett@ENRON.com" />

<cc name="Jeffrey C Gossett" id="45450" address="Jgosset@ENRON.com" />

<content>

M2.1. There are no accounting issues.

M2.2. I am assuming that Nowlan is retrading the deal.

M2.3. I will set up a meeting with Wes.

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 1033851</parent_id>

<message_id> 2000034</message_id>

<date_time> 2001-05-30 00:47:19</date_time>

<subject> Formosa - 1.25 Million</subject>

<from id="" name="" address="" />

<content>

M3.1. This is bizzare.

M3.2. Global Markets promised Jean this money and won't seem to write her a check.

M3.3. It was also a low number relative to what was created for GM.

M3.4. John

</content>

</message>

<message>

<depth> 2</depth>

<parent_id> 2000034</parent_id>

<message_id> 1033849</message_id>

<date_time> 2001-05-30 16:53:41</date_time>

<subject> Formosa - 1.25 Million</subject>

<from name="Wes Colwell" id="5416" address="Wes.Colwell@ENRON.com" />

<to name="John Lavorato" id="2273" address="John.J.Lavorato@ENRON.com" />

<to name="John Lavorato" id="2273" address="JLAVORA@ENRON.com" />

<content>
M4.1. I am told that this money should come this week.
</content>
</message>
<message>
<depth> 2</depth>
<parent_id> 1033848</parent_id>
<message_id> 1033850</message_id>
<date_time> 2001-05-29 11:35:41</date_time>
<subject> Update - Formosa - 1.25 Million</subject>
<from name="Jeffrey C Gossett" id="45450" address="Jeffrey.C.Gossett@ENRON.com" />
<to name="Jean Mrha" id="1994" address="Jean.Mrha@ENRON.com" />
<to name="John Lavorato" id="2273" address="John.J.Lavorato@ENRON.com" />
<to name="Jean Mrha" id="1994" address="Jmrha@ENRON.com" />
<to name="John Lavorato" id="2273" address="JLAVORA@ENRON.com" />
<content>
M5.1. Jean/ John -
M5.2. It looks like they have fixed their "problems" and we should be getting our money through accounting tonight.
M5.3. Jean -
M5.4. Can you make sure that Carol is talking to Greg Whiting in Gas Accounting?
M5.5. Thanks
</content>
</message>
</thread>

## A.21 Example Thread (ID: 89)

<thread>
<thread_id>89</thread_id>
<message>
<depth> 0</depth>
<parent_id> NULL</parent_id>
<message_id> 289628</message_id>
<date_time> 2001-05-24 17:13:18</date_time>
<subject> Trading Track Interviews</subject>

<from name="Karen Buckley" id="17792" address="Karen.Buckley@ENRON.com" />

<to name="Chuck Ames" id="18735" address="Chuck.Ames@ENRON.com" />

<to name="Bilal Bajwa" id="2119" address="Bilal.Bajwa@ENRON.com" />

<to name="Russell Ballato" id="27055" address="Russell.Ballato@ENRON.com" />

<to name="Steve Gim" id="3242" address="Steve.Gim@ENRON.com" />

<to name="Mog Heu" id="11352" address="Mog.Heu@ENRON.com" />

<to name="Juan Padron" id="40214" address="Juan.Padron@ENRON.com" />

<to name="Vladi Pimenov" id="4769" address="Vladi.Pimenov@ENRON.com" />

<to name="Denver Plachy" id="31147" address="Denver.Plachy@ENRON.com" />

<to name="Paul Schiavone" id="89562" address="Schiavone.Paul@ENRON.com" />

<to name="Elizabeth Shim" id="2456" address="Elizabeth.Shim@ENRON.com" />

<to name="Vince J Kaminski" id="63574" address="Matt.Smith@ENRON.com" />

<to name="Joseph Wagner" id="2153" address="Joseph.Wagner@ENRON.com" />

<to name="Jason Wolfe" id="617" address="Jason.Wolfe@ENRON.com" />

<to name="Virawan Yawapongsiri" id="117805" address="Virawan.Yawapongsiri@ENRON.com" />

<to name="Chuck Ames" id="18735" address="Cames@ENRON.com" />

<to name="Bilal Bajwa" id="2119" address="Bbajwa@ENRON.com" />

<to name="Russell Ballato" id="27055" address="D8433394-425A1999-86256919-7AC9AF@ENRON.com" />

<to name="Steve Gim" id="3242" address="86e09235-60cf1e4e-8625692f-69b1b7@ENRON.com" />

<to name="Mog Heu" id="11352" address="b5b64a78-ae842218-86256923-734572@ENRON.com" />

<to name="Juan Padron" id="40214" address="Jpadron@ENRON.com" />

<to name="Vladi Pimenov" id="4769" address="vpimenov@enron.com" />

<to name="Denver Plachy" id="31147" address="Dplachy@ENRON.com" />

<to name="Paul Schiavone" id="89562" address="Pschiavo@ENRON.com" />

<to name="Elizabeth Shim" id="2456" address="eshim@enron.com" />

<to name="Vince J Kaminski" id="63574" address="msmith18@enron.com" />

<to name="Joseph Wagner" id="2153" address="5F4B2CE5-B365809A-86256921-7F49AD@ENRON.com" />

<to name="Jason Wolfe" id="617" address="Jwolfe@ENRON.com" />

<to name="Virawan Yawapongsiri" id="117805" address="Vyawapon@ENRON.com" />

<cc name="Adrianne Engler" id="5918" address="adrianne.engler@ENRON.com" />

<cc name="Adrianne Engler" id="5918" address="Aengler@ENRON.com" />

<content>

M1.1. All,

M1.2. We are scheduling Interviews for new candidates to the Trading Track,

M1.3. next Wedesday 30th May.

M1.4. I am looking for volunteers to particpate in this event.

M1.5. Agenda:

M1.6. Tuesday: 29th: Dinner with five external candidates Tuesday night.

M1.7. Wednesday 30th:

M1.8. 11.00 - 12.30 Office tour - Trading Floor/Gas Control Room - five external candidates

M1.9. 12.30 - 1.30 Lunch with all 15 candidates (comination of internal and external).

M1.10. I need 3 of you to attend dinner on the 29th,

M1.11. 2 of your to faciliate the office tour and give insight to the operations

M1.12. and would like for all of you to attend the lunch.

M1.13. Pls advise on both at your earliest conveneince.

M1.14. Rgds, Karen.

</content>

</message>

<message>

<depth> 1</depth>

<parent_id> 289628</parent_id>

<message_id> 291424</message_id>

<date_time> 2001-05-24 17:52:04</date_time>

<subject> Trading Track Interviews</subject>

<from name="Elizabeth Shim" id="2456" address="Elizabeth.Shim@ENRON.com" />

<to name="Karen Buckley" id="17792" address="Karen.Buckley@ENRON.com" />

<to name="Chuck Ames" id="18735" address="Chuck.Ames@ENRON.com" />

<to name="Bilal Bajwa" id="2119" address="Bilal.Bajwa@ENRON.com" />

<to name="Russell Ballato" id="27055" address="Russell.RWB.Ballato@ENRON.com" />

<to name="Steve Gim" id="3242" address="Steve.Gim@ENRON.com" />

<to name="Mog Heu" id="11352" address="Mog.Heu@ENRON.com" />

<to name="Juan Padron" id="40214" address="Juan.Padron@ENRON.com" />

<to name="Vladi Pimenov" id="4769" address="Vladi.Pimenov@ENRON.com" />

<to name="Denver Plachy" id="31147" address="Denver.Plachy@ENRON.com" />

<to name="Paul Schiavone" id="89562" address="Schiavone.Paul@ENRON.com" />

<to name="Vince J Kaminski" id="63574" address="Matt.Smith@ENRON.com" />

<to name="Joseph Wagner" id="2153" address="Joseph.JHW.Wagner@ENRON.com" />

<to name="Jason Wolfe" id="617" address="Jason.Wolfe@ENRON.com" />

<to name="Virawan Yawapongsiri" id="117805" address="Virawan.Yawapongsiri@ENRON.com" />

<to name="Karen Buckley" id="17792" address="Kbuckley@ENRON.com" />

<to name="Chuck Ames" id="18735" address="Cames@ENRON.com" />

<to name="Bilal Bajwa" id="2119" address="Bbajwa@ENRON.com" />

<to name="Russell Ballato" id="27055" address="Rballato@ENRON.com" />

<to name="Steve Gim" id="3242" address="86e09235-60cf1e4e-8625692f-69b1b7@ENRON.com" />

<to name="Mog Heu" id="11352" address="b5b64a78-ae842218-86256923-734572@ENRON.com" />
<to name="Juan Padron" id="40214" address="Jpadron@ENRON.com" />
<to name="Vladi Pimenov" id="4769" address="vpimenov@enron.com" />
<to name="Denver Plachy" id="31147" address="Dplachy@ENRON.com" />
<to name="Paul Schiavone" id="89562" address="Pschiavo@ENRON.com" />
<to name="Vince J Kaminski" id="63574" address="msmith18@enron.com" />
<to name="Joseph Wagner" id="2153" address="Jwagner2@ENRON.com" />
<to name="Jason Wolfe" id="617" address="Jwolfe@ENRON.com" />
<to name="Virawan Yawapongsiri" id="117805" address="Vyawapon@ENRON.com" />
<cc name="Adrianne Engler" id="5918" address="adrianne.engler@ENRON.com" />
<cc name="Adrianne Engler" id="5918" address="Aengler@ENRON.com" />
<content>
M2.1. Hi Karen,
M2.2. I'll be happy to volunteer to host the candidates for dinner.
M2.3. Since the upcoming week is bid week for gas trading,
M2.4. I will not be able to host the candidates during the day.
</content>
</message>
<message>
<depth> 1</depth>
<parent_id> 289628</parent_id>
<message_id> 2000026</message_id>
<date_time> 2001-05-27 02:19:31</date_time>
<subject> Trading Track Interviews</subject>
<from id="" name="" address="" />
<content>
M3.1. Karen,
M3.2. I was out last week.
M3.3. Do you still need any help?
M3.4. Mat
</content>
</message>
<message>
<depth> 2</depth>
<parent_id> 2000026</parent_id>
<message_id> 289637</message_id>
<date_time> 2001-05-29 11:25:45</date_time>

<subject> Trading Track Interviews</subject>

<from name=”Karen Buckley” id=”17792” address=”Karen.Buckley@ENRON.com” />

<to name=”Vince J Kaminski” id=”63574” address=”Matt.Smith@ENRON.com” />

<to name=”Vince J Kaminski” id=”63574” address=”msmith18@enron.com” />

<content>

M4.1. yes, can you attend the lunch tomorrow.

M4.2. Thanks, Karen.

</content>

</message>

<message>

<depth> 3</depth>

<parent_id> 289637</parent_id>

<message_id> 2000027</message_id>

<date_time> 2001-05-29 11:28:57</date_time>

<subject> Trading Track Interviews</subject>

<from id=”” name=”” address=”” />

<content>

M5.1. Sure thing.

</content>

</message>

<message>

<depth> 4</depth>

<parent_id> 2000027</parent_id>

<message_id> 289638</message_id>

<date_time> 2001-05-29 11:32:10</date_time>

<subject> Trading Track Interviews</subject>

<from name=”Karen Buckley” id=”17792” address=”Karen.Buckley@ENRON.com” />

<to name=”Vince J Kaminski” id=”63574” address=”Matt.Smith@ENRON.com” />

<to name=”Vince J Kaminski” id=”63574” address=”msmith18@enron.com” />

<content>

M6.1. great. thanks,,

M6.2. will send out the location details later.

M6.3. thanks,

</content>

</message>

</thread>

## A.22 Example Thread (ID: 99)

<thread>

<thread_id>99</thread_id>

<message>

<depth> 0</depth>

<parent_id> NULL</parent_id>

<message_id> 277651</message_id>

<date_time> 2001-05-25 15:00:57</date_time>

<subject> GISB contracts for intrastate and interstate gas</subject>

<from name="Keith Ford" id="66315" address="kford1@txu.com" />

<to name="Daren J Farmer" id="28042" address="dfarmer@enron.com" />

<content>

M1.1. At the request of Lynn Handlin, please find attached a copy of the following:

M1.2. GISB contract form

M1.3. Special Provisions for intrastate gas

M1.4. Special Provision for interstate gas

M1.5. Please review,

M1.6. and if interested in putting these contracts in place, forward to me all required information on the first page of the GISB contract form.

M1.7. I will then have the contracts prepared and forwarded to you for execution.

M1.8. Please keep in mind that TXU Fuel Company is an intrastate pipeline

M1.9. and can purchase interstate gas only under certain limited conditions.

M1.10. These conditions are covered under item number 8 in the Special Provisions.

M1.11. Thanks.

M1.12. Keith Ford

M1.13. TXU Fuel Company

M1.14. Contract Administration Supervisor

</content>

</message>

<message>

<depth> 2</depth>

<parent_id> NULL</parent_id>

<message_id> 727061</message_id>

<date_time> 2001-07-11 14:05:43</date_time>

<subject> GISB contracts for intrastate and interstate gas</subject>

<from name="Anthony Campos" id="5378" address="Anthony.Campos@ENRON.com" />

<to name="Daren J Farmer" id="28042" address="Daren.J.Farmer@ENRON.com" />

<to name="Daren J Farmer" id="28042" address="dfarmer@enron.com" />

<content>

M2.1. Mr. Farmer,

M2.2. I have forwarded your request with comments to Debra Perlingiere (Legal Specialist - x3-7658) and Stacey Dickson (Sr. Counsel - x3-5705)

M2.3. who handle negotiations for new Master Agreements.

M2.4. Please let me know if I may be of further assistance.

M2.5. Thank You,

M2.6. Anthony Campos

M2.7. Enron Corp._Global Contracts

M2.8. 713.853.7911 (office)

M2.9. 713.646.2495 (fax)

M2.10. 713.709.0373 (pager)

M2.11. Anthony.Campos@enron.com

</content>

</message>

</thread>

# Appendix B

# Power, Gender, and Gender Environment: Statistical Test Results

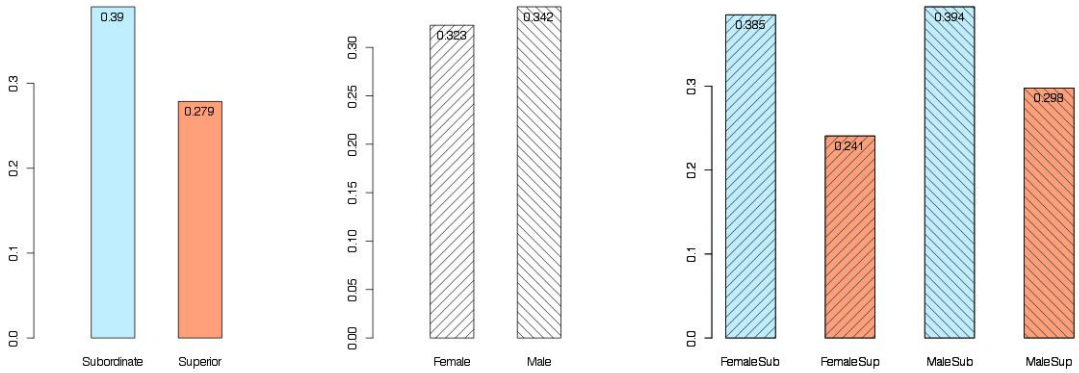



Figure B.1: Mean value differences along Gender and Power: Initiator

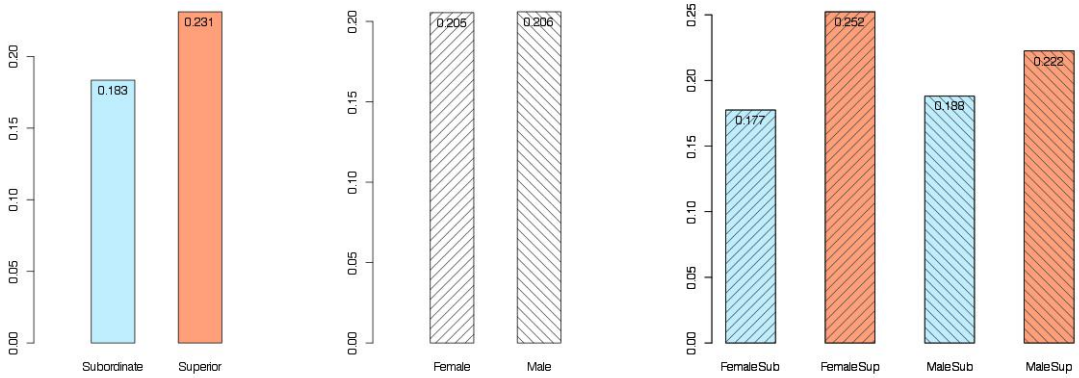



Figure B.2: Mean value differences along Gender and Power: FirstMsgPos

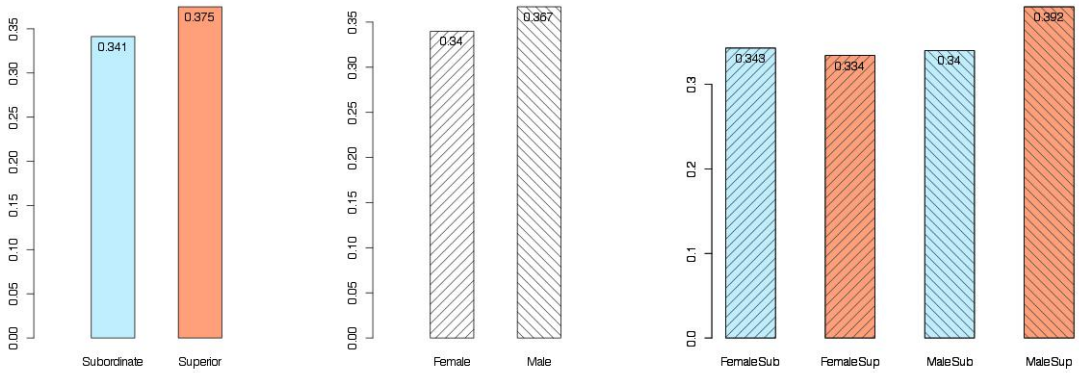



Figure B.3: Mean value differences along Gender and Power: LastMsgPos

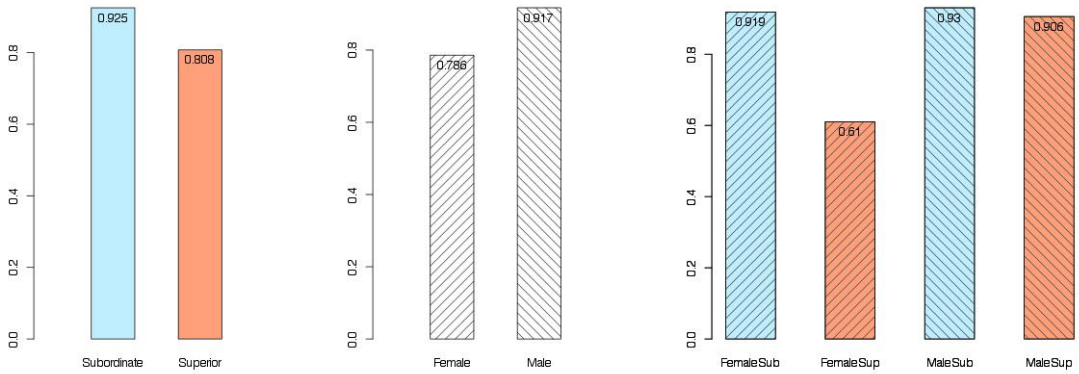



Figure B.4: Mean value differences along Gender and Power: MsgCount

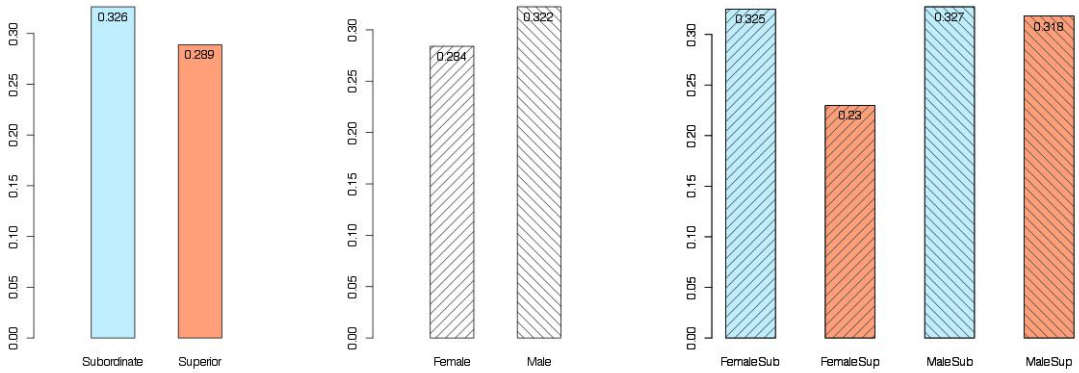



Figure B.5: Mean value differences along Gender and Power: MsgRatio

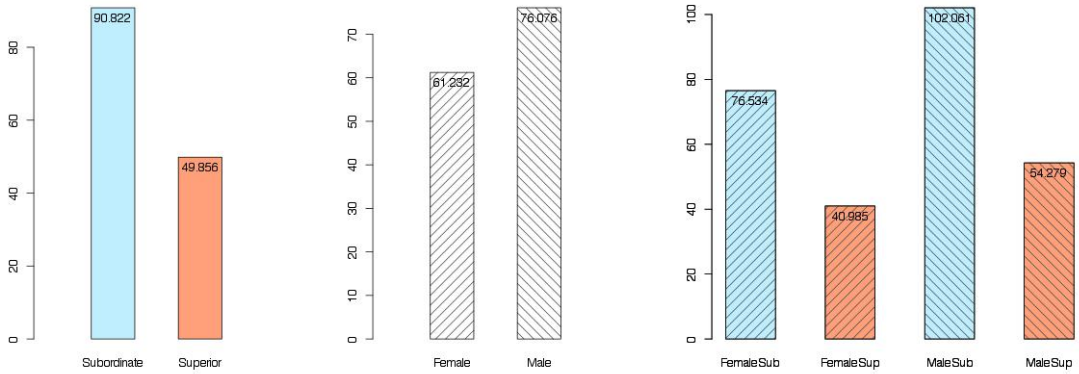



Figure B.6: Mean value differences along Gender and Power: TokenCount

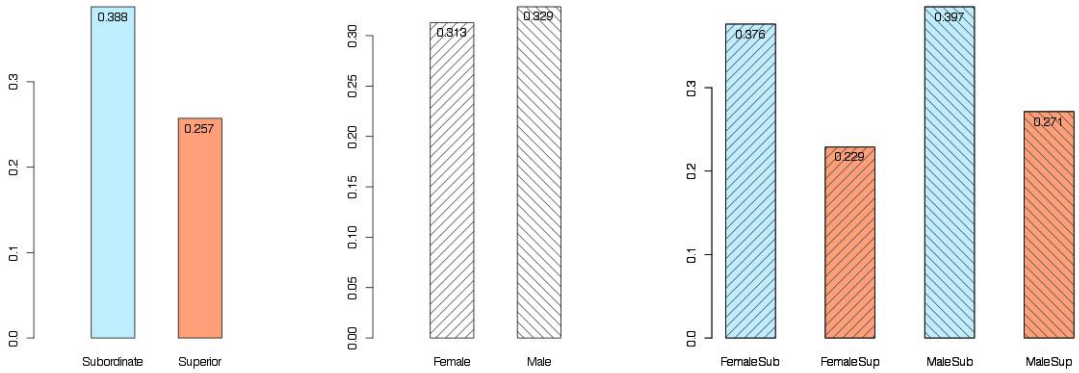



Figure B.7: Mean value differences along Gender and Power: TokenRatio

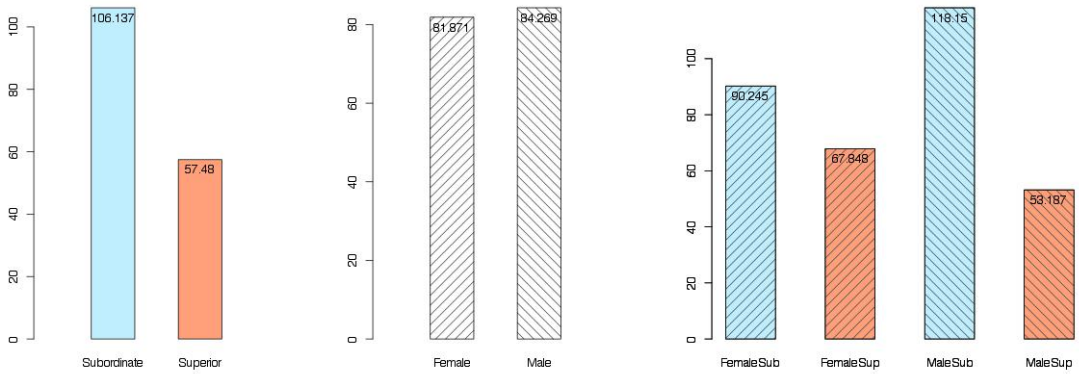



Figure B.8: Mean value differences along Gender and Power: TokenPerMsg

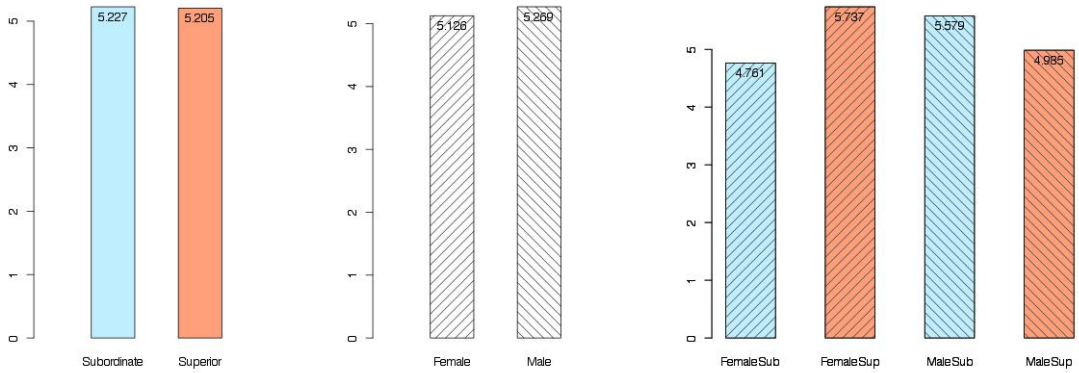



Figure B.9: Mean value differences along Gender and Power: AvgRecipients

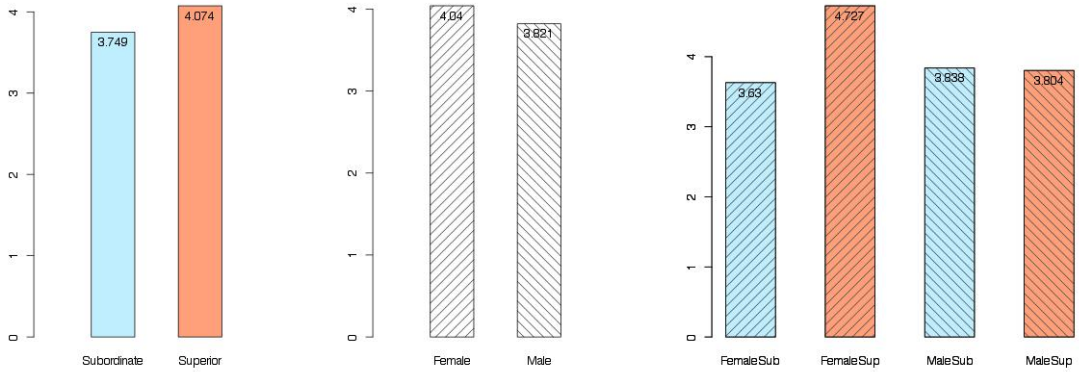



Figure B.10: Mean value differences along Gender and Power: AvgToRecipients

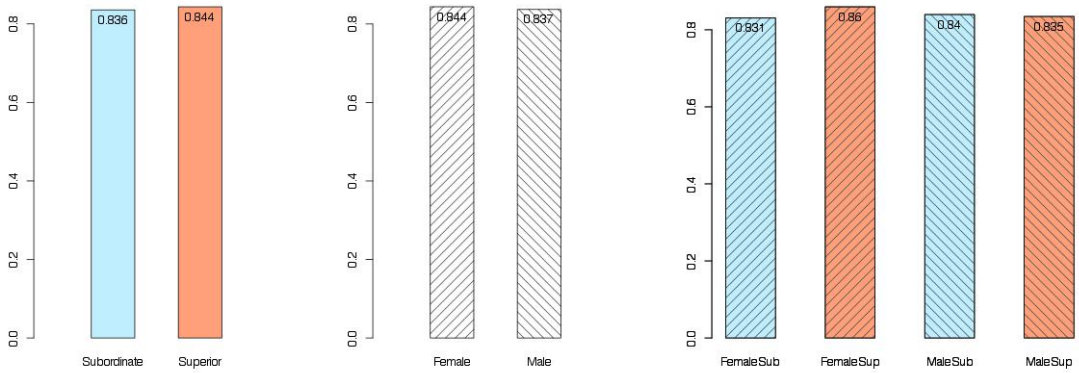



Figure B.11: Mean value differences along Gender and Power: InToList%

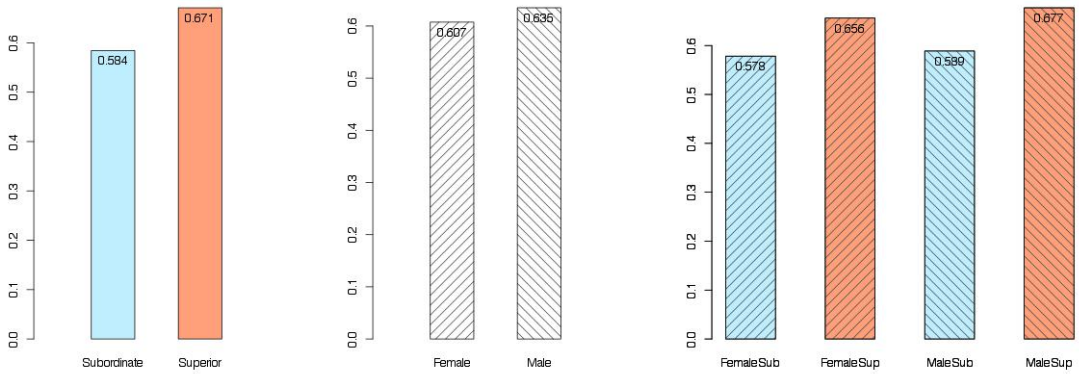



Figure B.12: Mean value differences along Gender and Power: AddPerson

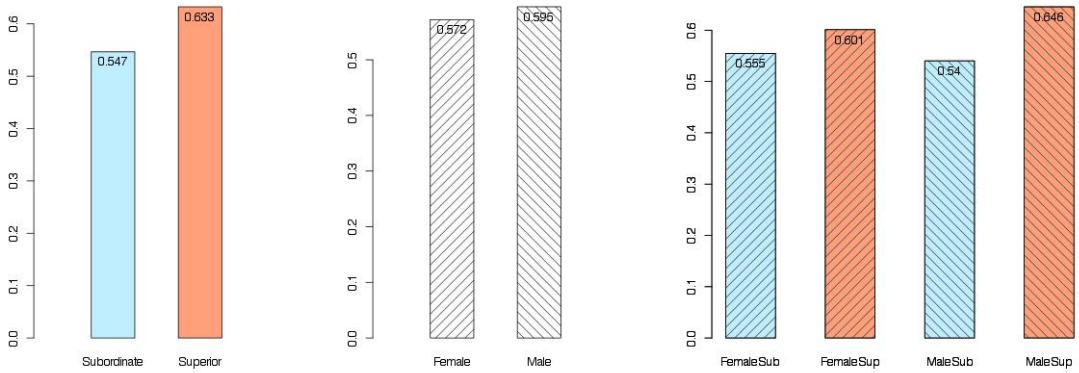



Figure B.13: Mean value differences along Gender and Power: RemovePerson

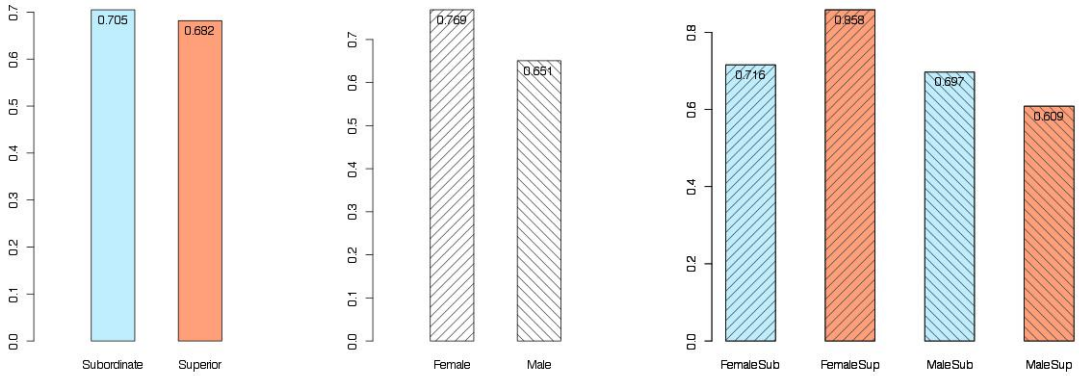



Figure B.14: Mean value differences along Gender and Power: ReplyRate

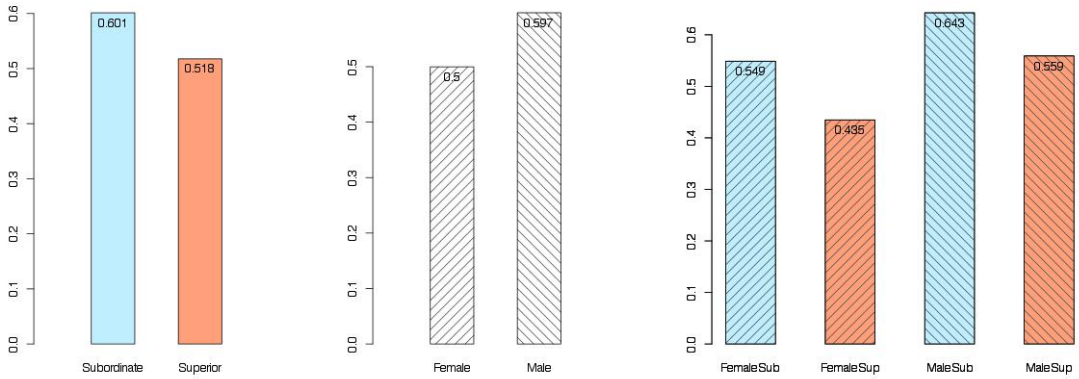



Figure B.15: Mean value differences along Gender and Power: ConventionalCount

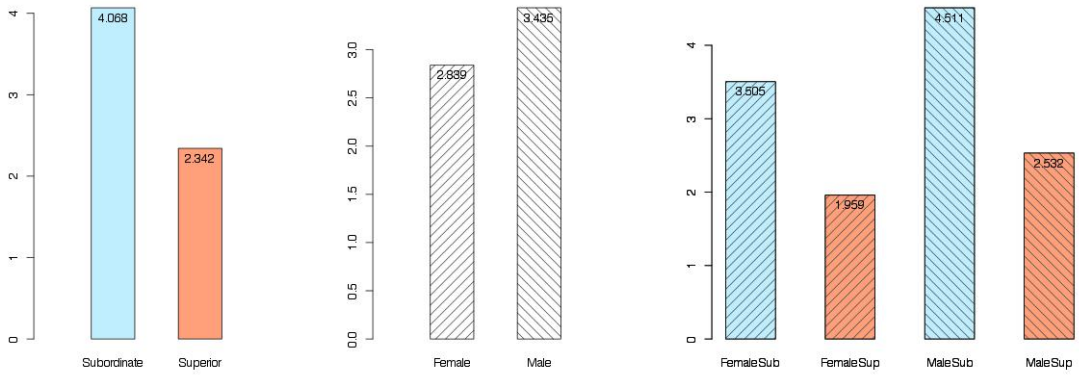



Figure B.16: Mean value differences along Gender and Power: InformCount

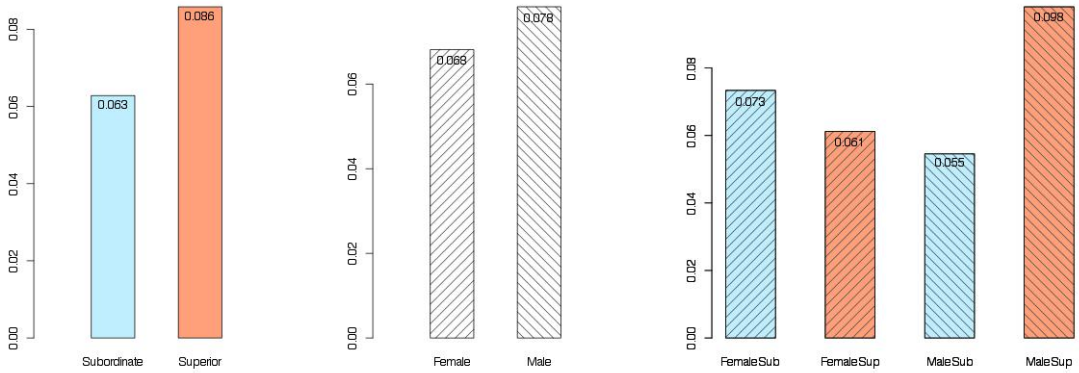



Figure B.17: Mean value differences along Gender and Power: ReqActionCount

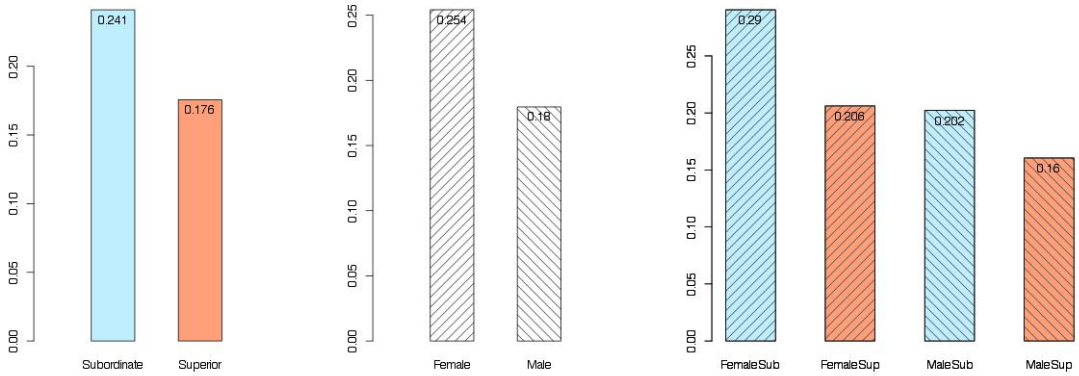



Figure B.18: Mean value differences along Gender and Power: ReqInformCount

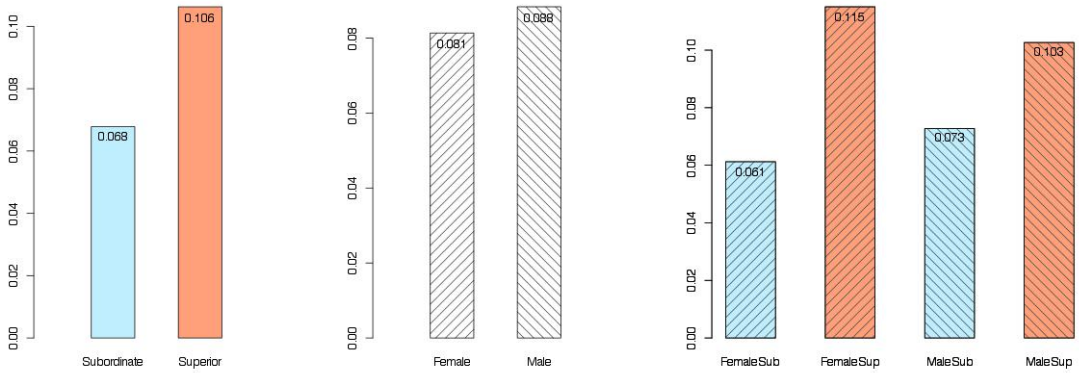



Figure B.19: Mean value differences along Gender and Power: DanglingReq%

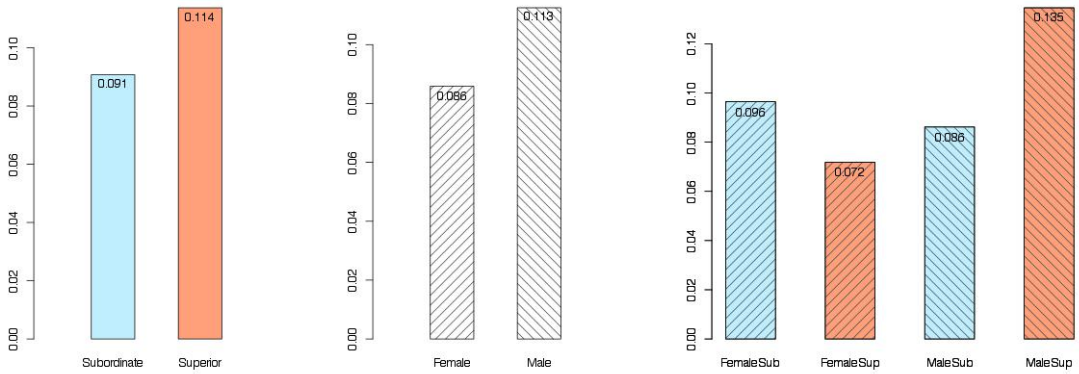



Figure B.20: Mean value differences along Gender and Power: ODPCount

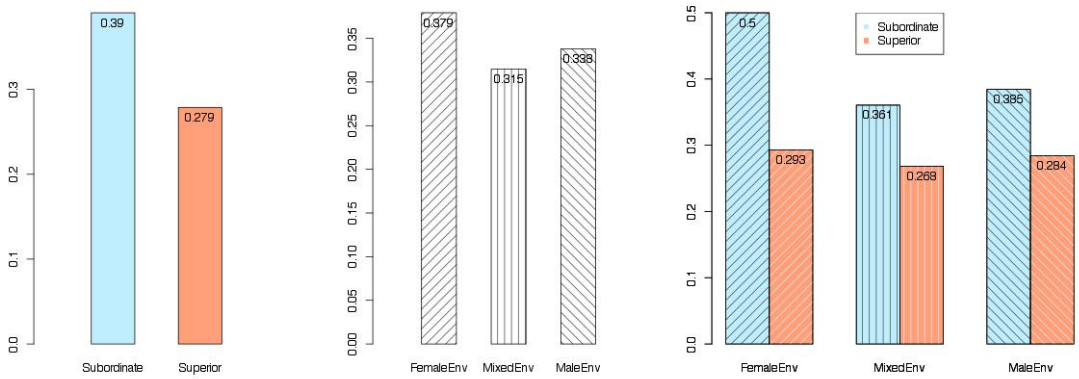



Figure B.21: Mean value differences along Gender Environment and Power: Initiator

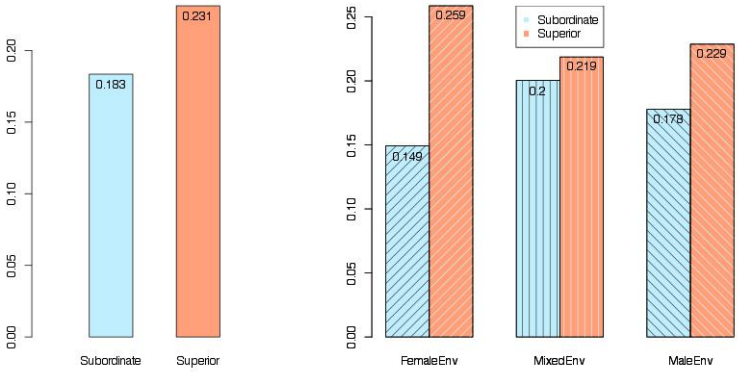



Figure B.22: Mean value differences along Gender Environment and Power: FirstMsgPos

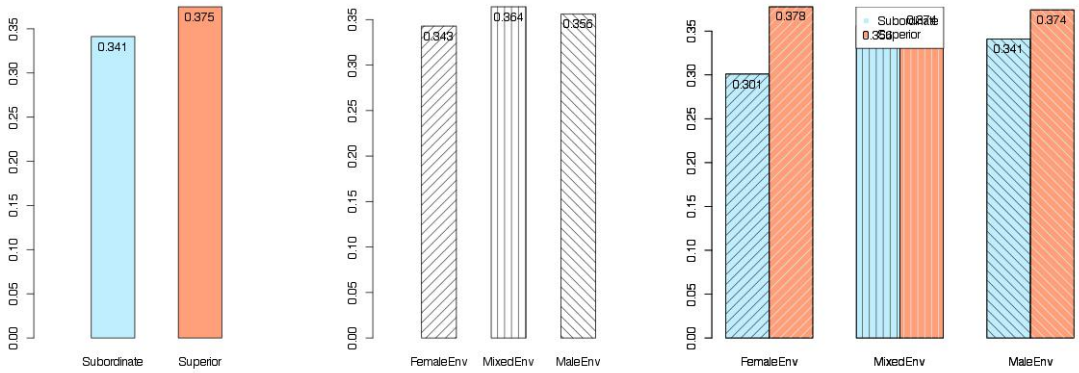



Figure B.23: Mean value differences along Gender Environment and Power: LastMsgPos

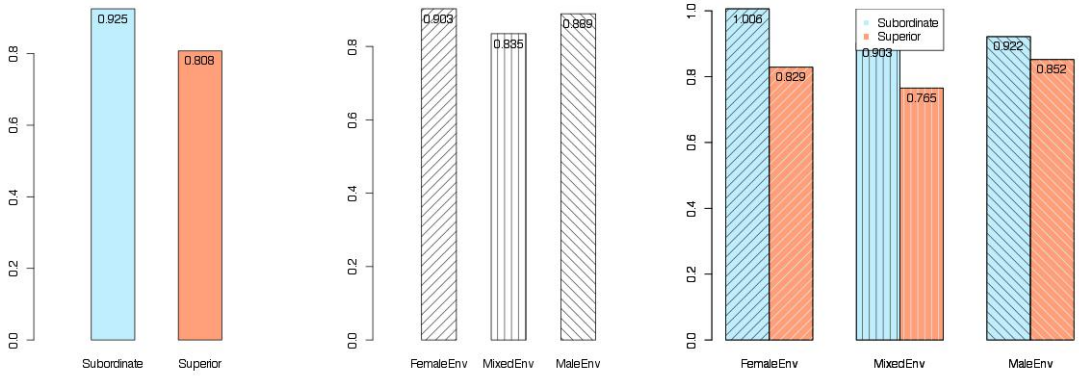



Figure B.24: Mean value differences along Gender Environment and Power: MsgCount

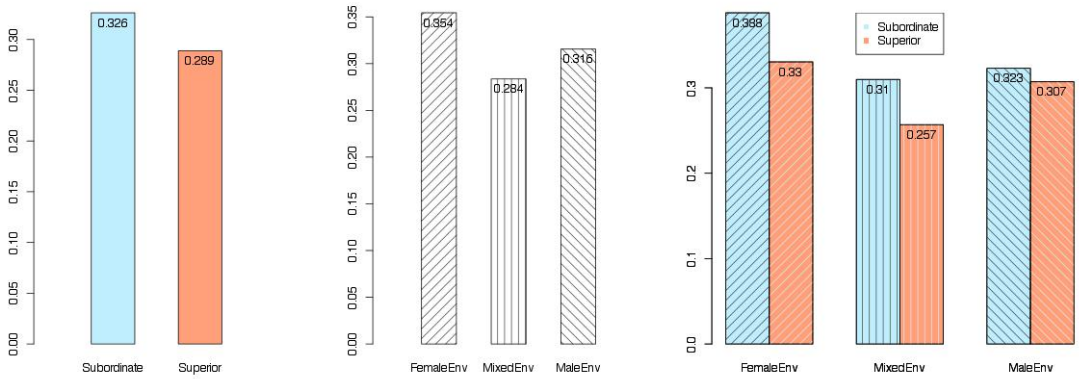



Figure B.25: Mean value differences along Gender Environment and Power: MsgRatio

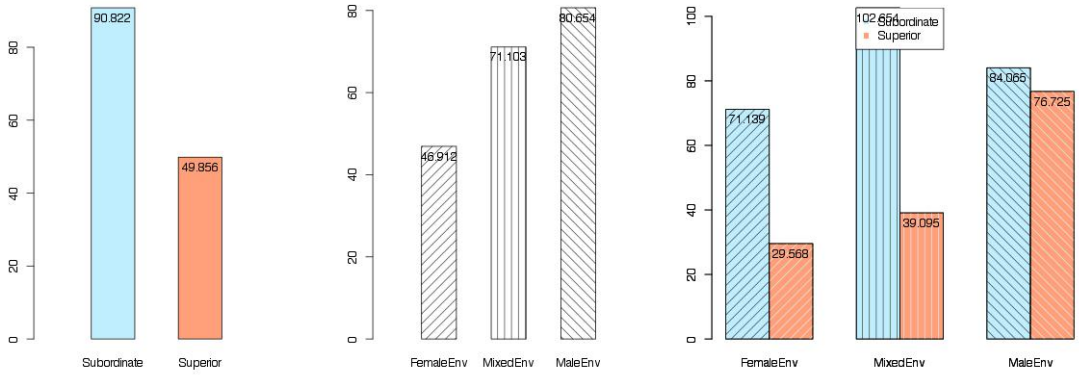



Figure B.26: Mean value differences along Gender Environment and Power: TokenCount

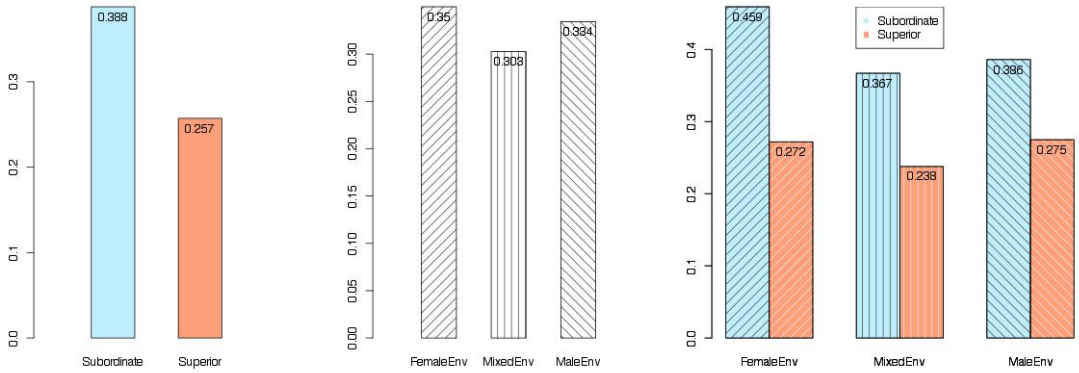



Figure B.27: Mean value differences along Gender Environment and Power: TokenRatio

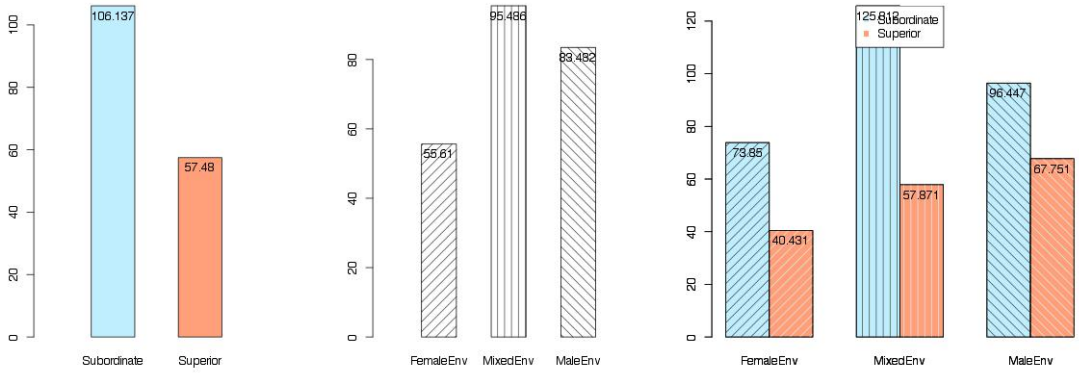



Figure B.28: Mean value differences along Gender Environment and Power: TokenPerMessage

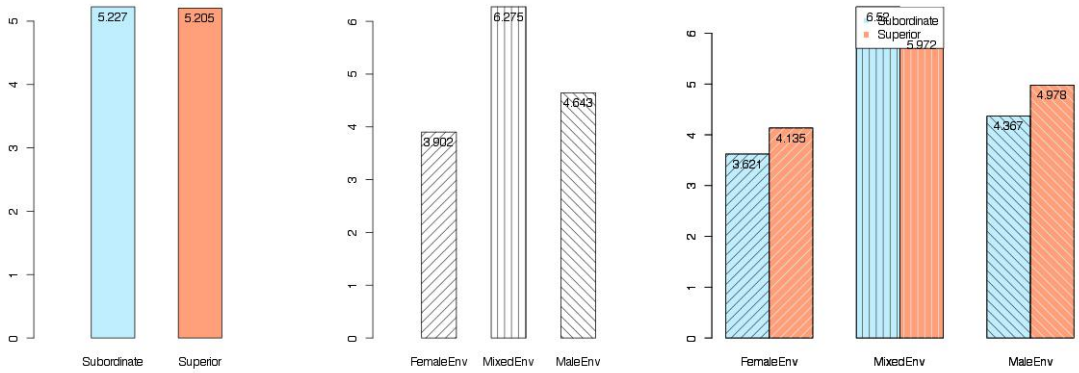



Figure B.29: Mean value differences along Gender Environment and Power: AvgRecipients

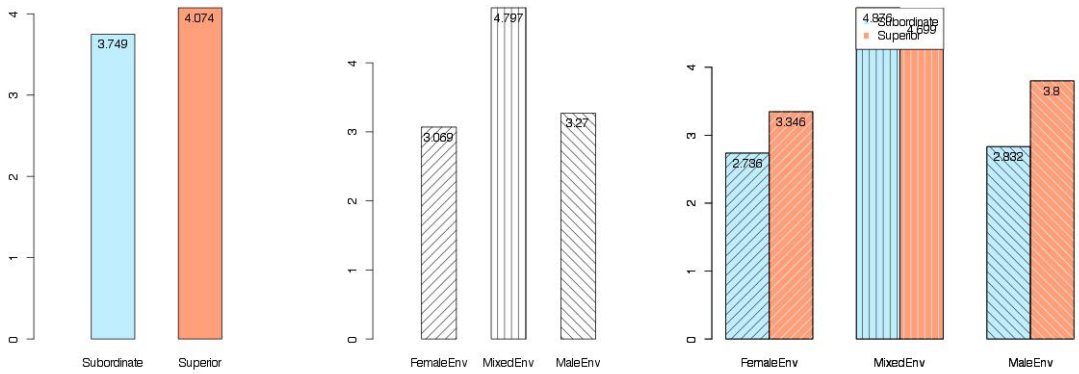



Figure B.30: Mean value differences along Gender Environment and Power: AvgToRecipients

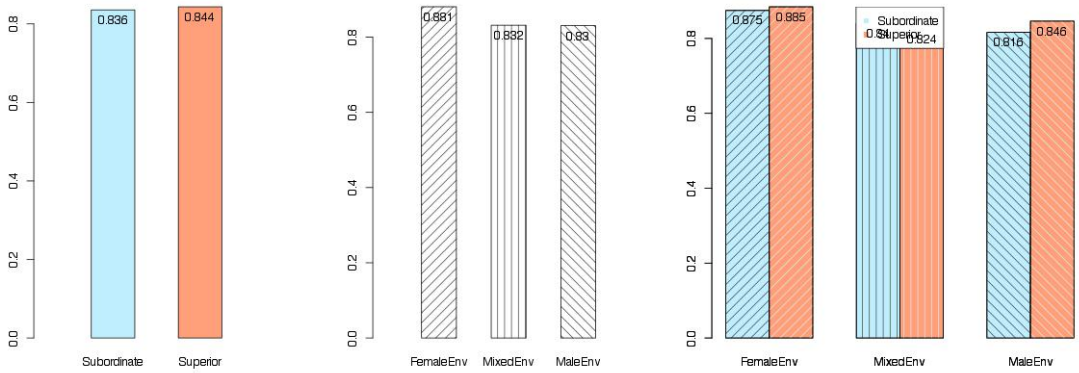



Figure B.31: Mean value differences along Gender Environment and Power: InToList%

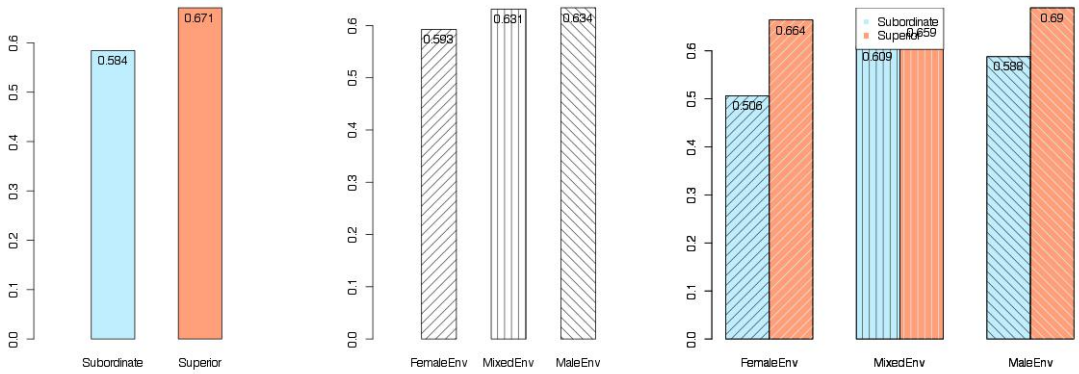



Figure B.32: Mean value differences along Gender Environment and Power: AddPerson

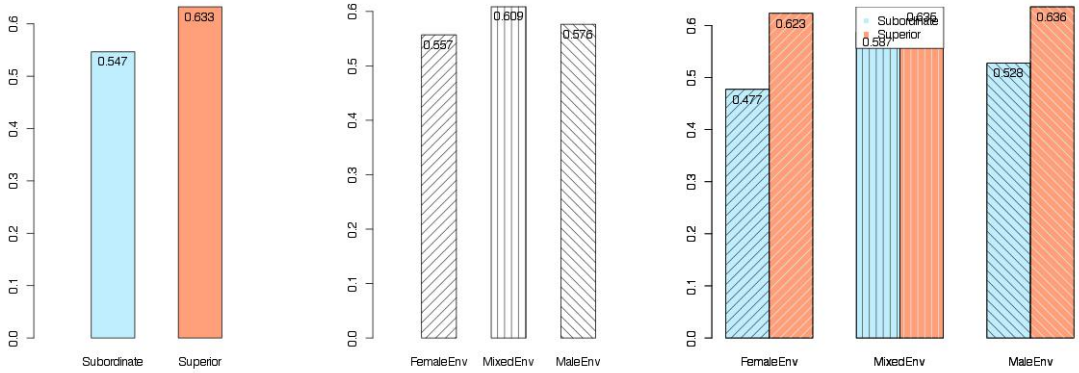



Figure B.33: Mean value differences along Gender Environment and Power: RemovePerson

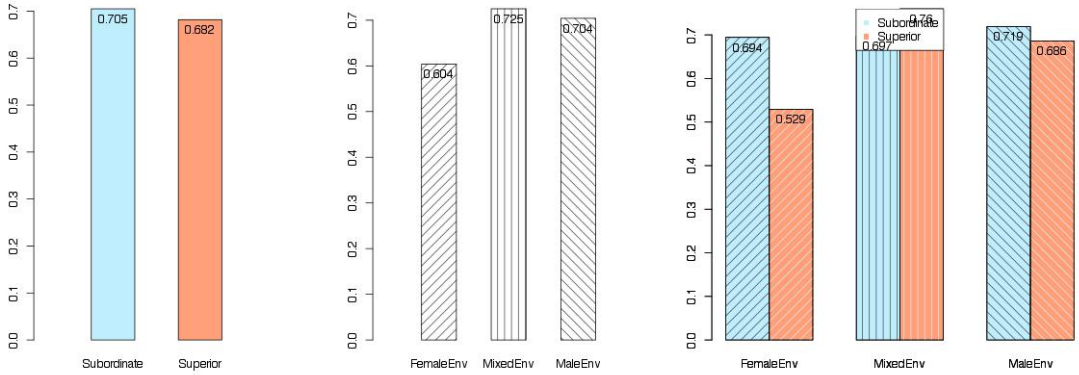



Figure B.34: Mean value differences along Gender Environment and Power: ReplyRate

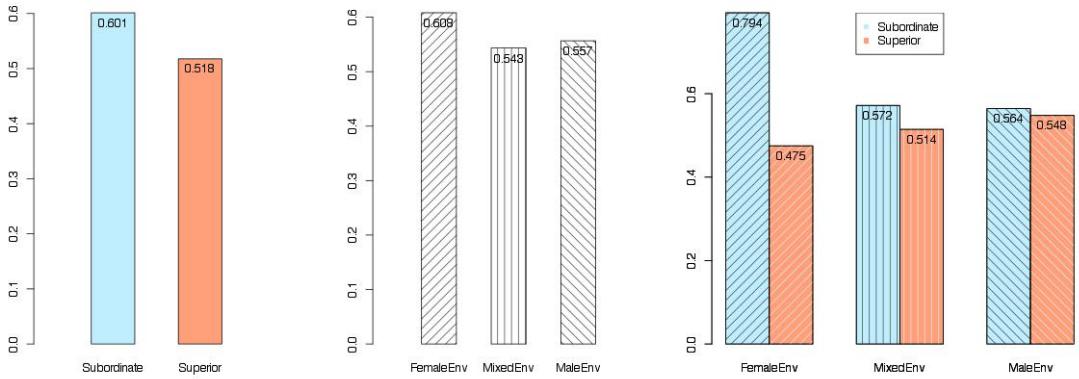



Figure B.35: Mean value differences along Gender Environment and Power: ConventionalCount

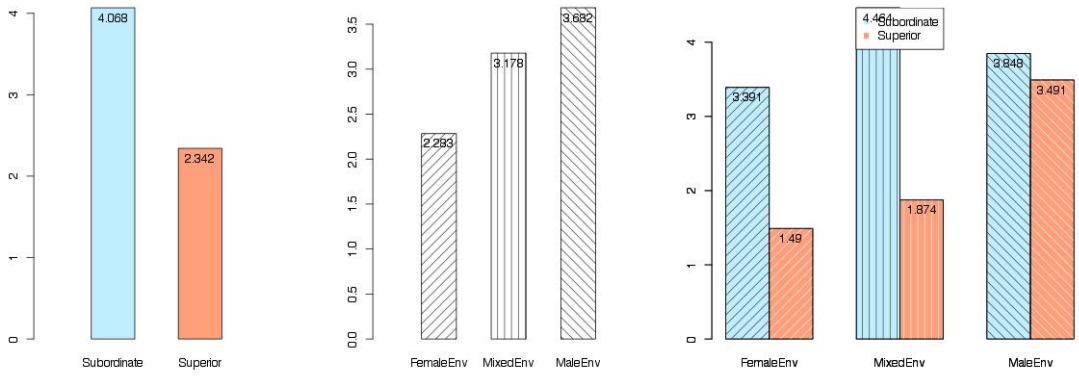



Figure B.36: Mean value differences along Gender Environment and Power: InformCount

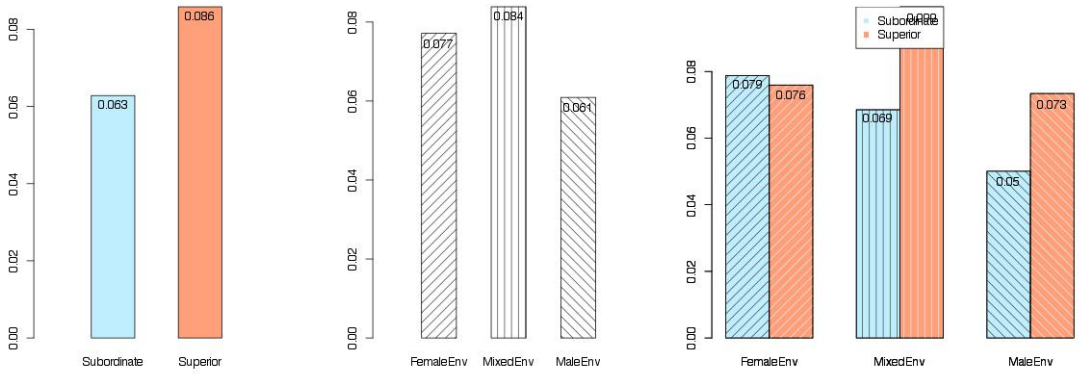



Figure B.37: Mean value differences along Gender Environment and Power: ReqActionCount

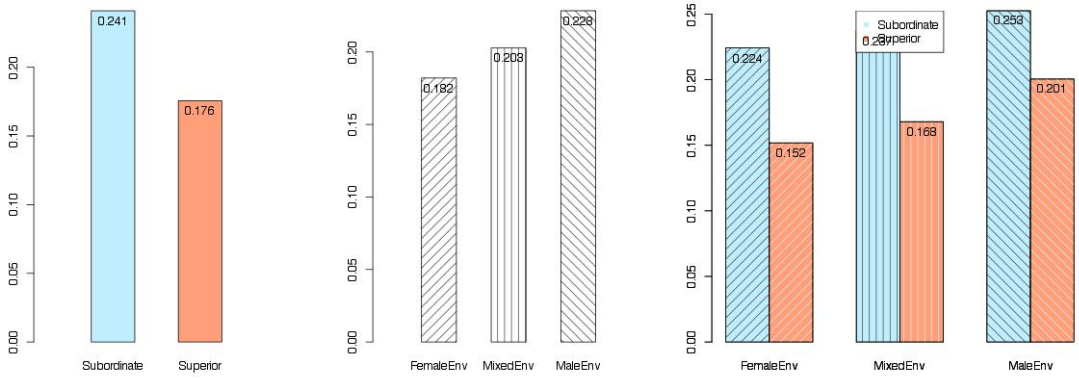



Figure B.38: Mean value differences along Gender Environment and Power: ReqInformCount

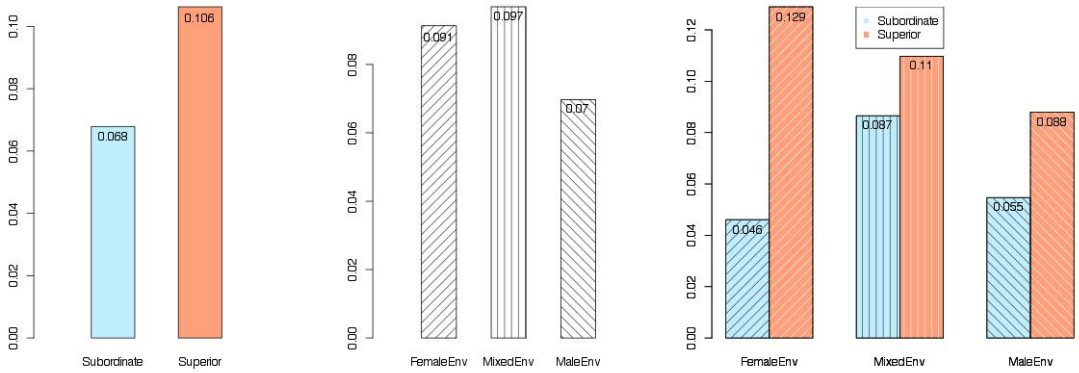



Figure B.39: Mean value differences along Gender Environment and Power: DanglingReq%

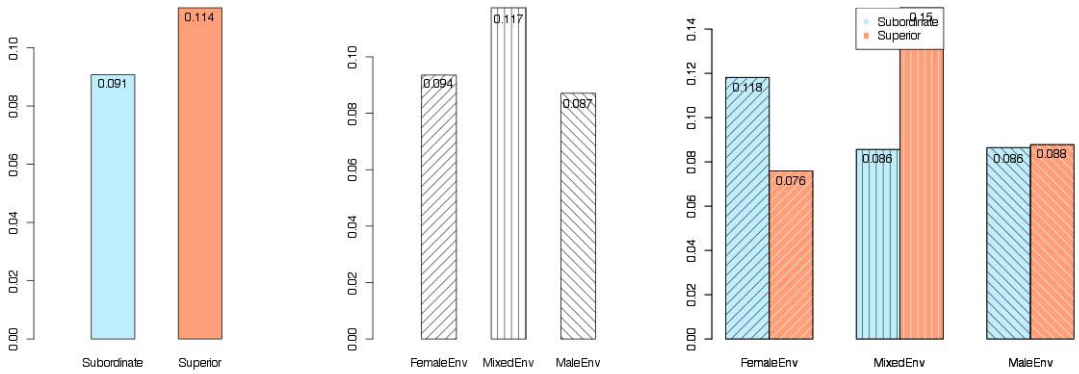



Figure B.40: Mean value differences along Gender Environment and Power: ODPCount