Who's (Really) the Boss?
Perception of Situational Power in Written Interactions

Vinodkumar Prabhakaran[1] *Owen Rambow*[2] *Mona Diab*[2]
(1) Department of Computer Science, Columbia University, New York, NY 10027, USA
(2) Center for Computational Learning Systems, Columbia University, New York, NY 10115, USA
vinod@cs.columbia.edu, {rambow,mdiab}@ccls.columbia.edu

Abstract
We study the perception of situational power in written dialogs in the context of organizational emails and contrast it to the power attributed by organizational hierarchy. We analyze various correlates of the perception of power in the dialog structure and language use by participants in the dialog. We also present an SVM-based machine learning system using dialog structure and lexical features to predict persons with situational power in a given communication thread.

Keywords: Computational Sociolinguistics, Social Networks, Power Analysis, Dialog Acts, Dialog.

Title and Abstract in German

Wer ist (wirklich) der Chef?
Situative Macht in schriftlichen Dialogen

Wir untersuchen, wie situative Macht in schriftlichen Dialogen wahrgenommen wird. Situative Macht ist Macht, die nicht beständig ist, sondern nur zielbedingt während einer (vielleicht längeren) Interaktion existiert. Unsere Studie beruht auf Geschäftsemails. Wir kontrastieren situative Macht mit der Macht, die durch die organisatorische Hierarchie entsteht. Wir identifizieren verschiedene Korrelate der Wahrnehmung situativer Macht in der Dialogstruktur und im Sprachgebrauch der Dialogteilnehmer. Wir stellen ausserdem einen SVM-basierten maschinellen Lernalgorithmus vor, der Dialogstruktur und Wörter in den Emails benutzt, um Dialogteilnehmer mit situativer Macht zu identifizieren.

Keywords in German: Komputationalle Soziolinguistik, Soziale Netzwerke, Macht, Dialogakte, Dialog.

Proceedings of COLING 2012: Technical Papers, pages 2259–2274,
COLING 2012, Mumbai, December 2012.

1 Introduction

Within an interaction, there is often a power differential between the interactants. This differential is often drawn from a static source external to the interaction, such as a formal or informal power structure or hierarchy. Most computational studies (Creamer et al., 2009; Bramsen et al., 2011; Gilbert, 2012) that analyze power within interactions have used such an external power structure (namely, a corporate hierarchy) as the definition of the power differential. However, the power differential may also be dynamic and specific to the situation of the interaction. We define a person to have **situational power** if there is another person such that he or she has power or authority over the first person in the context of a situation or task. Such situational power may not always align with the external power structures, if one exists. For example, a Human Resources department employee within an organization will have power over an office manager when the interaction is about enforcing some HR policy. But the direction of power will be reversed if the interaction is about allocating office space for a new hire. Neither of these directions of power may be captured in the organizational hierarchy chart. In some cases, the situational power may even be in the opposite direction to the power relation reflected in the hierarchy — e.g., a subordinate organizing an office event.

Although power is a difficult concept to define, it is often recognizable from an interaction. As we show in this paper, one of the primary ways power is recognized in dialog is by the manner in which people participate in the dialog. Power relations sometimes constrain how one behaves when engaging in dialog; in some other cases, they enable one to constrain another person's behavior. And in some cases, the dialog behavior becomes a tool to express and even pursue power. By dialog behavior, we mean the choices a discourse participant makes while engaging in dialog. It includes choices with respect to the message content, like lexical choice or degree of politeness. It also includes choices participants make in terms of dialog structure, such as the choice of when to participate with how much and what sort of contribution, the number of questions to ask and which of those questions to answer. These manifestations may differ depending on whether the power differential is entirely hierarchical or situational or a mix of both - hierarchical power may be inert in a particular interaction, but situational power is a rather active form of power.

In this paper, we focus on situational power, how an outsider perceives it and what attributes of the interaction contributes to that perception. More specifically, we focus on why a dialog participant is perceived to have situational power within an interaction. Another related problem is whether a participant is perceived to have situational power over a *specified person* or not. We do not address it in this paper. We first analyze the notion of situational power in detail and find that its perception is subjective. We then define a set of features describing the choices participants make, including the dialog structure and the verbosity of dialog participants, and study how they correlate with the perception of situational power. We build a system to detect participants with situational power using a supervised machine learning approach. The main findings of this study are: a) situational power is a very different form of power than the power ascribed by hierarchy; b) the perception of situational power correlates with dialog structure as well as content-based features. The contributions of this paper also include an automatic situational power tagger based on dialog structure and content features. This study is conducted on email threads; however, we do not use any aspects of written dialog that are specific to email. Hence, we expect the methodology we use and the insights we gain to be applicable to other genres of online social media such as online discussion forums.

The rest of the paper is structured as follows: Section 2 presents related work in this field. Section 3 presents the data and describes various annotations present in the data, including situational power annotations. Section 4 compares the perception of situational power with hierarchical power. Section 5 defines the problem formally and Section 6 describes all dialog structure and verbosity features we used. Section 7 presents statistical significance measures of these features with respect to persons perceived to have situational power. Section 8 presents the machine learning experiments conducted on the data and discusses results. We finally conclude and discuss future work.

2 Related Work

Most definitions of power in the sociology literature — e.g., (Dahl, 1957; Emerson, 1962; Bierstedt, 1950) — 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). The five bases of power proposed by French and Raven (1959) (Coercive, Reward, Legitimate, Referent, and Expert) and their extensions are widely used in sociology to study power. Wartenberg (1990) makes the distinction between two notions of power: power-over and power-to. Power-over refers to hierarchical relationships between interactants, while power-to refers to the ability to exercise power an interactant possesses (maybe temporarily) and can use within the interaction. Our notion of situational power roughly corresponds to Wartenberg's notion of power-to, while the power attributed to an interactant by an organizational hierarchy can be considered an instance of power-over. Both can be considered special cases of French and Raven's notion of legitimate power.

It has long been established that there is a 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; Scherer, 1979; Brook 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. Reid and Ng (2000) explain this correlation by saying that "conversational turns function as a resource for establishing influence": discourse participants can manipulate the dialog structure in order to gain influence. This echoes a starker formulation by Bales (1970): "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." Simply successfully claiming the conversational floor represents a feat of power. The previous work just discussed was conducted entirely on spoken dialog. In this paper, we show that the core insight — conversation is a resource for power — carries over to written dialog. This means that we can predict powerful people from studying 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.

We now turn to the computational literature. Several studies have used Social Network Analysis (Diesner and Carley, 2005; Shetty and Adibi, 2005; Creamer et al., 2009) or email traffic patterns (Namata et al., 2007) for extracting social relations from online communication. 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 relatively a new area which has only recently become an active area of research (Bramsen et al., 2011; Gilbert, 2012; Danescu-Niculescu-Mizil et al., 2012).

Bramsen et al. (2011) and Gilbert (2012) address the same problem we are trying to solve: identifying social power relationships from online written communication. They both also use the Enron email corpus for their experiments. 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. Firstly, our focus is on situational power and how it is perceived by an outsider, while they focus on power from the organizational hierarchy. We show in Section 4 that the perception of situational power is not aligned with the organizational hierarchy. Secondly, our data unit is a naturally occurring thread, not data units assembled by the researchers, and a thread may or may not include a person who has power. Also, we focus on the structure of the dialog (which we can do since our unit is a thread, as opposed to a single message or an arbitrary aggregation of single messages).

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 based on context and perception. Strzalkowski et al. (2010) 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. They model power using notions of topic switching, exploiting mainly complex lexical features. Peterson et al. (2011) focus on formality in Enron email messages and relates it to social distance and power.

3 Data

3.1 Corpus

We use the email corpus presented in (Prabhakaran et al., 2012a) which contains manual annotations for various types of power relations between participants. In this study, we focus only on Situational Power (SP) and leave the other types of power for future work. The corpus contains 122 email threads with a total of 360 messages and 20,740 word tokens. This set of email threads is chosen from a version of the Enron email corpus with some missing messages restored from other emails in which they were quoted (Yeh and Harnly, 2006). Most emails are concerned with exchanging information, scheduling meetings, and solving problems, but there are also purely social emails. Table 1 presents some statistics on participants and messages in the corpus. We define an active participant of a given thread as someone who has sent at least one email message in the thread.

Apart from the thread level annotations for different types of power, the corpus also contains utterance level annotations for overt displays of power. The same corpus has been previously annotated with dialog act annotations (Hu et al., 2009). We utilize these annotations in our study

Statistic	Count / Mean (SD)
Number of email threads	122
Number of participants	1033
Ave. Participants / thread	8.47 (13.82)
Number of active participants	221
Ave. Active participants / thread	1.81 (0.73)
Number of messages	360
Ave. Messages / thread	2.95 (2.24)
Ave. Messages / active participant	1.45 (1.01)
Number of word tokens	20,740
Situational Power (SP)	81

Table 1: Corpus statistics

and will describe them in more detail in the following sections. We give an example thread and corresponding situational power annotations in Table 2. The example shows annotations for overt display of power (ODP) which will be explained in Section 3.1.2. The email body also contains dialog act annotations which will be explained in Section 3.1.3.

3.1.1 Situational Power Annotations

Person_1 is said to have situational power over person_2 if person_1 has power or authority to direct and/or approve person_2's actions in the current situation or while a particular task is being performed, based on the communication in the current thread. Situational power is independent of 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). For more details on the situational power annotations, see (Prabhakaran et al., 2012a), where we explain these annotations in more detail with examples and describe instructions given to the annotator.

In our example thread, the annotator judged *William* to be possessing situational power over *Barry* and *Barry* over *Stephanie*: in both cases, a task is assigned to another person, even if the language used is more direct in the case of *Barry* delegating his task to *Stephanie*.

3.1.2 Overt Display of Power (ODP) annotations

An utterer can choose linguistic forms in her utterance to signal that she is imposing constraints on the addressee's choice of how to respond, which go beyond those defined by the standard set of dialog acts. For example, if the boss's email is “Please come to my office right now”, and the addressee declines, he is clearly not adhering to the constraints the boss has signaled, though he is adhering to the general constraints of cooperative dialog by responding to the request for action. This roughly captures the “restriction of an interactant's action-environment", suggested as one of the key elements to identify exercise of power in interactions by Wartenberg (1990). An utterance is defined to have an **overt display of power (ODP)** if it is interpreted as creating additional constraints on the response beyond those imposed by the general dialog act. Syntactically, an ODP can be an imperative, a question, or a declarative sentence. The presence of an ODP does not presuppose that the utterer actually possess social power: the utterer could be attempting to gain power. Out of the 1734 utterances in our corpus, 86 were annotated to have an expression of ODP. In our example thread, utterances M2.2 and M2.6 were labeled as instances of ODP.

From: William S Bradford
To: Barry Tycholiz
CC: Michael Tribolet
Subject: Gas Inventories
Time: 2001-09-24 10:45:00

M1.1. Barry,
[Conventional]
M1.2. Let me know if you have any time to review.
[Inform]
M1.3. Bill
[Conventional]

From: Barry Tycholiz
To: Stephanie Miller
Subject: Gas Inventories
Time: 2001-09-24 11:11:05

M2.1. Steph,
[Conventional]
M2.2. further to our discussion, Pls review.
[Request-Action]
Flink2.2
M2.3. I took a quick look at the locations and most appear to be East based.
[Inform]
M2.4. You might want to use an analyst to figure this out.
[Inform]
M2.5. 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.
[Inform]
M2.6. Pls review and let's discuss asap.
[Request-Action]
Flink2.6
M2.7. BT
[Conventional]

Situational Power	William S Bradford –> Barry Tycholiz
	Barry Tycholiz –> Stephanie Miller
Overt Display of Power	M2.2, M2.6

Table 2: Example thread and annotations; note that the dialog act for M1.2 appears to be incorrectly labeled as Inform, instead of Request-Action or perhaps Request-Information (we did not change any dialog act labels)

For a further discussion of the annotation of ODP, see (Prabhakaran et al., 2012b), where we further define the notion of ODP, give inter-annotator agreement numbers, and present initial work on building an automatic classifier for ODP.

3.1.3 Dialog Act annotations

The corpus we used also contains manual dialog act annotations as described in Hu et al. (2009). We use these annotations to model the dialog structure of the communication thread. Each message in the thread is segmented into Dialog Functional Units (DFUs). A DFU is a contiguous subset of a turn (i.e., in our corpus, of an email message) which has a coherent communicative intention. Each DFU is assigned a Dialog Act (DA) label which is one of the following: **Request for Action**, **Request for Information**, **Inform**, **Inform-Offline**,[1] **Conventional**, and **Commit**.

In addition, DFUs are interlinked by three types of links to reflect the dialog structure. These links capture the patterns of local alternation between an initiating dialog act and a responding one. A **forward link** (Flink) is the analog of a "first pair-part" of an adjacency pair, and is similarly restricted to specific speech act types. All Request-Information and Request-Action DFUs are assigned Flinks. The responses to such requests are assigned a **backward link** (Blink). If an utterance can be interpreted as a response to a preceding DFU, it gets a Blink even where the preceding DFU has no Flink. The preceding DFU taken to be the "first pair-part" of the link is assigned a **secondary forward** link (SFlink).

3.2 How Subjective is the Perception of Situational Power?

In this section, we investigate how subjective the perception of situational power is. We performed an independent study of annotator perceptions on a subset of 47 threads from the corpus. We trained two additional annotators — AnnA and AnnB — using the same annotation manual described in (Prabhakaran et al., 2012a) and compared the annotations they produced for situational power on the selected threads. Both AnnA and AnnB are undergraduate students, one from the Arts Department and the other from the Engineering Department.

The cognitive process behind labeling a participant to have situational power is not a binary decision the annotator makes for each participant. Annotators read the entire thread before performing the annotations. 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 annotations about situational power 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 situational power or not. There were 289 participants in the selected 47 threads. AnnA found 19 of these participants to have situational power while AnnB found only 13 to have situational power, 8 of which were also found by AnnA. We obtained a κ value of 0.472 which is only a moderate agreement. The fact that we don't obtain a higher agreement 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

[1] Sometimes, the Inform act refers to a previous act of communication that did not happen in the email thread itself. Such cases are marked as Offline.

him or her. However, these 50 recipients will be added as data points in our κ calculation, thereby increasing the expected agreement and decreasing the κ value. Another reason could be just that the task by itself is subjective. The indicators that are noticed by each annotator may underspecify how they can be interpreted in the power narrative (and subsequently the situational power annotation). 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, in our example thread, 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 annotation in (Prabhakaran et al., 2012a) adopted the latter narrative whereas both AnnA and AnnB adopted the former. In our investigation of 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. We obtained similar moderate (>0.3) agreement between AnnA and AnnB and the original annotations for the subset of threads that were triply annotated. The moderate agreement suggests that there must be some core indicators of situational power that we could obtain by combining multiple perceptions. We leave that to future work. For the rest of this paper, we rely on the original annotations for the perception we are modeling. In Section 8.3, we explicitly address the issue of whether the original annotator's perceptions are internally consistent.

4 Are they really the bosses?

We explore how the perception of situational power compares with the organizational hierarchy. For this purpose, we utilize the gold organizational hierarchy for Enron released by (Agarwal et al., 2012). It contains relations between 1,518 employees, and 13,724 dominance pairs (pairs of employees such that the first dominates the second in the hierarchy, not necessarily immediately). We labeled a participant to have hierarchical power within a thread if there exist a dominance pair in the gold hierarchy where he/she is dominating over any other participant in the same thread.

According to the gold hierarchy, 113 out of the 1033 participants in our corpus have hierarchical power within the interaction. But only 12 of them (10.6%) were perceived to have situational power by our annotator. In other words, bosses act bossy rather rarely. Also, a total of 107 participants were judged to have situational power. The 12 of those who also had hierarchical power amounts to only 11.2% of them. In other words, you don't have to be the boss to be bossy.

The above findings are particularly important since most previous computational approaches have concentrated on modeling power purely in hierarchical terms. These findings show the importance of thinking beyond hierarchy and about other types of power.

5 Problem Definition and Approach

Now we move on to finding correlates in the interaction that could help predict participants with situational power. Given a communication thread $\mathscr{T}$ and an active participant $\mathscr{X}$, we would like to predict whether $\mathscr{X}$ has situational power over some person $\mathscr{Y}$ in the thread.[2] In this paper, we restrict ourselves to features relative to messages sent by the participant $\mathscr{X}$. Hence, we consider each active participants in the thread as a data point and extract features with respect to them. There were 221 active participants in our corpus out of which 81 were annotated to have situational power. Our approach is to build a binary classifier predicting whether or not $\mathscr{X}$ has situational power based on features with respect to $\mathscr{X}$ in the context of the given thread $\mathscr{T}$.

6 Feature Sets

In this section, we describe six sets of features we use to capture the way interactants participate in dialog. The first four sets of features — dialog act (DAP), dialog link (DLC), positional (PST) features and verbosity (VRB) — relate to the whole dialog and its structure, whereas lexical (LEX) features and overt display of power (ODP) are features related to the form and content of individual messages. PST, VRB, and LEX are readily derivable from the data, while we use the gold annotation for DAP, DLC, and OSP.

6.1 Dialog act features - DAP

We have six features: **ReqAction**, **ReqInform**, **Inform**, **InformOffline**, **Conventional**, and **Commit**, denoting the percentage of each of these dialog act labels aggregated over all messages sent by the participant within the thread.

6.2 Dialog structure link features - DLC

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. **Flink**, **SFlink** and **Blink** denote the total number of Flinks, SFlinks and Blinks in messages sent by the participant. **Clink** denotes the number of Blinks by other people connected back to DFUs in messages sent by the participant. This includes both Flinks and SFlinks that are connected. **Dlink** denotes the number of Flinks by the participant that were not connected back via Blinks by other people ("dangling Flinks"). These are requests with no responses. **DlinkRatio** denotes Dlinks as a percentage of the number of Flinks by the participant.

6.3 Positional features - PST

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. We used two other features: **FirstMsg** and **LastMsg**, to denote the position where the participant sent his/her first and last message, normalized by the total number of messages in the thread.

6.4 Verbosity features - VRB

We use features denoting how verbose the participant is within the thread. **MsgCount** denotes the number of messages sent by the participant. **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.

[2] The related problem to predict if person $\mathscr{X}$ has situational power over a specified person $\mathscr{Y}$ is not addressed in this paper.

Set	Features	SP
DAP	ReqAction	$\mathbf{0.07/0.01}_{0.01}$
	ReqInform	$0.10/0.12_{0.70}$
	Inform	$0.56/0.63_{0.10}$
	InformOffline	$0.003/0.005_{0.62}$
	Conventional	$0.25/0.23_{0.35}$
	Commit	$0.001/0.003_{0.51}$
DLC	Flink	$\mathbf{0.98/0.59}_{0.03}$
	SFlink	$\mathbf{0.49/0.24}_{0.02}$
	Blink	$0.72/0.59_{0.40}$
	Clink	$\mathbf{0.83/0.44}_{7.1E-3}$
	Dlink	$0.64/0.39_{0.08}$
	DlinkRatio	$\mathbf{0.33/0.21}_{0.05}$
PST	Initiator	$\mathbf{0.68/0.48}_{3.3E-3}$
	FirstMsg	$\mathbf{0.13/0.24}_{1.1E-3}$
	LastMsg	$0.41/0.36_{0.21}$
VRB	MsgCount	$\mathbf{1.68/1.32}_{0.03}$
	MsgRatio	$0.54/0.50_{0.18}$
	TokenCount	$\mathbf{113.04/74.19}_{0.02}$
	TokenRatio	$\mathbf{0.62/0.47}_{2.1E-3}$
	TokensPerMsg	$73.22/54.76_{0.07}$
ODP	ODPCount	$\mathbf{0.78/0.14}_{6.0E-8}$

Table 3: Statistical significance measures; values with $p \leq 0.05$ are boldfaced

6.5 Lexical features - LEX

The LEX feature set contains lexical features extracted from the content of messages sent by the participant. We aggregated all messages sent by the participant in the thread and extracted ngram counts for word lemmas. We experimented with **Unigram** counts, **Bigram** counts and a combination of both. We found that unigram counts performed better than the other two. Higher order ngrams were found to decrease the performance of the system.

6.6 Overt Display of power - ODP

We used a feature **ODPCount** to denote the number of instances of ODP in messages sent by the participant. For this study, we used the gold ODP tags (as we use the gold dialog annotations).

7 Statistical Significance Study

In this section, we present the statistical significance study of dialog features with respect to persons with situational power. We consider two populations of people who participated in the dialog – $\mathscr{P}_p$, those judged to have situational power and $\mathscr{P}_n$, those not judged to have situational power. Then, for each feature, we performed a two-sample, two-tailed t-test comparing means of feature values of $\mathscr{P}_p$ and $\mathscr{P}_n$. Table 3 presents means of each feature value for both populations $\mathscr{P}_p$ and $\mathscr{P}_n$ (as mean($\mathscr{P}_p$)/mean($\mathscr{P}_n$)) along with the p-value associated with the t-test as the subscript. For p-values less than 0.05, we reject the null hypothesis and consider the feature to be statistically significant. We have highlighted the statistically significant features in Table 3.

7.1 Significant Features

We find many features to be statistically significant, which suggests that situational power is reflected in the dialog structure and content of messages. For example, persons with situational power tend to utter requests for action (ReqAction) significantly more than those without. They have significantly more connected links (Clink, SFlink); but the ratio of dangling links is also significantly higher for them, probably because they issue significantly more forward links (Flink). They also tend to be the initiators of the thread (Initiator) or start participating in the thread closer to the beginning (FirstMsg). They talk more within a thread (TokenRatio) and send significantly more (MsgCount) and longer (TokensCount, TokensPerMsg) messages. They also have significantly more instances of overt displays of power (ODPCount) than others.

7.2 Multiple Test Correction

The statistical measures presented in the previous section are exploratory in nature, presenting tests on all features, to gain a better understanding of their interaction with the situational power. 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 in the previous section. However, on applying the Bonferroni correction to adjust the p-value for the number of tests performed (threshold = 0.05/21 = 0.0024), three features (one each from PST (FirstMsg), VRB (TokenRatio) and ODP (ODPCount)) still remain statistically significant. Hence the global null hypothesis that the dialog structure and language use do not interact with situational power would still be rejected.

8 Automatic Situational Power Tagger

In this section, we present machine learning experiments to predict persons with situational power using features described in Section 6. We train a binary classifier to predict whether a participant has situational power or not.

8.1 Machinery

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, POS tagger, lemmatizer and SVMLight (Joachims, 1999) wrapper. We balanced our dataset by up-sampling minority class instances in the training step. This has proven useful previously in cases of unbalanced datasets (Japkowicz, 2000). All results presented below have been obtained after balancing the training folds in cross validation; the test folds remain unchanged.

8.2 Experiments

Since we defined sets of dialog features (DAP, DLC, PST and VRB) in Section 6 based on separate aspects of communication they capture, we assume that these feature sets are reasonably coherent and individual features within a set interact with each other more strongly than with features from

Feature set	Features	P	R	F
Random		36.7	49.4	42.1
AlwaysTrue		36.7	100.0	53.6
VRB	TokenRatio, MsgRatio	43.9	70.4	54.0
PST	FirstMsg	45.1	67.9	54.2
DAP	ReqInform, Inform-Offline, Conventional, Commit	40.9	75.3	53.0
DLC	Blink, Flink, Clink, SFlink	49.6	**75.3**	59.8
LEX	Unigrams	54.9	55.6	55.2
ODP	ODPCount	**71.2**	51.9	**60.0**
BEST	DLC, ODP	59.4	70.4	64.4

Table 4: Cross validation results (P: Precision, R: Recall, F: F-measure) VRB: Verbosity, PST: Positional, DAP: Dialog acts, DLC: Dialog links, LEX: Lexical, ODP: Overt display of power

other feature sets. So, first we find the best performing subset of features for each feature set by exhaustive search within the set. The small cardinality of these feature sets (max of 6 for DAP and DLC) makes exhaustive search computationally feasible. For LEX, we found the best performing feature set to be Unigrams (Section 6.5), and ODP contains only one feature - ODPCount. 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 tried various feature selection methods such as information gain, which did not improve results. We believe this is because these methods rank each feature individually, and thus important interactions between features are not captured.

We used 5-fold cross validation on the data to evaluate the prediction performance for different feature subsets. 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. We report (R)ecall, (P)recision and (F)-measure ($\beta = 1$). 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.

8.3 Results

Table 4 shows cross validation results for each set of features. We present two baseline measures - **Random** and **AlwaysTrue**. In the Random baseline, we predict an active participant to have situational power at random. In AlwaysTrue baseline, we always predict an active participant to have situational power. The table 4 lists the best performing feature subset and corresponding precision/recall/f-measure for each set of features. As described in Section 8.2, the BEST feature set is found by doing exhaustive search on all combinations of best performers of VRB, PST, DAP, DLC, LEX and ODP.

The best performing individual feature sets are ODP and DLC, both at or near 60.0 F-measure. The random and AlwaysTrue baselines yield F-measures of 42.1 and 53.6, respectively. The best performers of all feature sets except DAP outperformed these baselines. The simple ngram based model obtained an F-measure of 55.2. While ODP results in a high precision (71.2) model, DLC yields a high recall (75.3) model; combination of both gave the best performing system with an

F measure of 64.4. The best performing single feature was ODPCount, which by itself gave an F measure of 60.0. The results we obtained are in line with the findings from the statistical significance study presented in Section 7. For example, DAP contained the least significant features while ODP contained the most significant feature. The tagger also performed worst when only DAP features were used and best when ODP was used. We assessed the statistical significance of F-measure improvements over baseline, using the Approximate Randomness Test (Yeh, 2000; Noreen, 1989).[3]. We found the improvements to be statistically significant (p = 0.001).

For the best performing feature set — DLC+ODP — we obtained a mean F-measure (macro-average) of 64.92 with a standard deviation of 8.82 (please note that Table 4 reports micro-averaged F-measure: 64.4). The low standard deviation suggests that the model built in this setting will obtain comparable performances for new unseen data. This also means that the data from which it was trained on the different folds of the cross-validation is sufficiently consistent to learn a model with predictive power. Put differently, our annotator's perception of situational power is coherent.

Conclusion and Future Work

In this paper, we studied the perception of situational power within written interactions. We have shown that situational power is not aligned with the organizational hierarchy. We have also shown that the perception of situational power correlates with various dialog structure and linguistic features. We presented an automatic situational power tagger to detect persons with power in written interactions. The methodology presented in this study can be applied to other forms of written interactions like discussions in online forums and blogs. We expect to find similar patterns of correlation in other genres.

As future work, we intend to study power relations between pairs of participants. It would be interesting to see how dialog features correlate with the other direction of power; that is from a submitter to an exerciser of power. Also, our current approach of aggregating features at the participant level is prone to noise. For example, let $\mathscr{X}$,$\mathscr{Y}$,$\mathscr{Z}$ be active participants such that $\mathscr{X}$ has power over $\mathscr{Y}$, who has power over $\mathscr{Z}$. When we aggregate features with respect to $\mathscr{Y}$, we are introducing noise from the part of communication between $\mathscr{X}$ and $\mathscr{Y}$. Extending our work to the person pair level would prevent this noise.

Our predictions are done using some gold features. Some features we used (verbosity, position) are readily derivable from the text; but others require processing. We will investigate using automatic taggers (such as a dialog act tagger and link predictor (Hu et al., 2009), an ODP tagger (Prabhakaran et al., 2012b)) to extract these features to predict power. However, one main contribution of this paper is to show the interaction between these dimensions of the dialog (like dialog structure and ODP) and situational power, which is an important first step towards solving the problem. Finally, in future work we will further study the manner in which different annotators interpret ambiguous threads in their power narratives, and identify different levels of certainty of situational power.

Acknowledgments

This work is supported, in part, by the Johns Hopkins Human Language Technology Center of Excellence. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsor. We thank Weiwei Guo for his valuable suggestions on this paper. We also thanks several anonymous reviewers for their constructive feedback.

[3]http://www.clips.ua.ac.be/ vincent/scripts/art.py

References

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

Bales, R. F. (1970). Personality and interpersonal behaviour. New York: Holt, Reinhart, and Winston.

Bales, R. F., Strodtbeck, F. L., M., M., T., and Roseborough, M. (1951). Channels of communication in small groups. *American Sociological Review*, pages 16(4), 461–468.

Bierstedt, R. (1950). An Analysis of Social Power. *American Sociological Review*.

Bramsen, P., Escobar-Molano, M., Patel, A., and Alonso, R. (2011). Extracting social power relationships from natural language. In *ACL*, pages 773–782. The Association for Computer Linguistics.

Brook, M. and Ng, S. H. (1986). Language and social influence in small conversational groups. *Journal of Language and Social Psychology*, pages 5(3), 201–210.

Creamer, G., Rowe, R., Hershkop, S., and Stolfo, S. J. (2009). Segmentation and automated social hierarchy detection through email network analysis. In Zhang, H., Spiliopoulou, M., Mobasher, B., Giles, C. L., Mccallum, A., Nasraoui, O., Srivastava, J., and Yen, J., editors, *Advances in Web Mining and Web Usage Analysis*, pages 40–58. Springer-Verlag, Berlin, Heidelberg.

Dahl, R. A. (1957). The concept of power. *Syst. Res.*, 2(3):201–215.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., and Kleinberg, J. M. (2012). Echoes of power: language effects and power differences in social interaction. In Mille, A., Gandon, F. L., Misselis, J., Rabinovich, M., and Staab, S., editors, *WWW*, pages 699–708. ACM.

Diesner, J. and Carley, K. M. (2005). 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.

Emerson, R. M. (1962). Power-Dependence Relations. *American Sociological Review*, 27(1):31–41.

French, J. R. and Raven, B. (1959). The Bases of Social Power. In Cartwright, D., editor, *Studies in Social Power*, pages 150–167+. University of Michigan Press.

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

Hu, J., Passonneau, R., and Rambow, O. (2009). 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. Association for Computational Linguistics.

Japkowicz, N. (2000). Learning from imbalanced data sets: Comparison of various strategies. In *AAAI Workshop on Learning from Imbalanced Data Sets*.

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

Namata, Jr., G. M. S., Getoor, L., and Diehl, C. P. (2007). Inferring organizational titles in online communication. In *Proceedings of the 2006 conference on Statistical network analysis*, ICML'06, pages 179–181, Berlin, Heidelberg. Springer-Verlag.

Ng, S. H., Bell, D., and Brooke, M. (1993). Gaining turns and achieving high in influence ranking in small conversational groups. *British Journal of Social Psychology*, pages 32, 265–275.

Ng, S. H., Brooke, M., , and Dunne, M. (1995). Interruption and in influence in discussion groups. *Journal of Language and Social Psychology*, pages 14(4),369–381.

Noreen, E. W. (1989). *Computer-Intensive Methods for Testing Hypotheses : An Introduction*. Wiley-Interscience.

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

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

Pfeffer, J. (1981). *Power in organizations*. Pitman, Marshfield, MA.

Prabhakaran, V., Rambow, O., and Diab, M. (2012a). Annotations for power relations on email threads. In *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Prabhakaran, V., Rambow, O., and Diab, M. (2012b). Predicting overt display of power in written dialogs. In *Human Language Technologies: The 2012 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Montreal, Canada. Association for Computational Linguistics.

Reid, S. A. and Ng, S. H. (2000). Conversation as a resource for in influence: evidence for prototypical arguments and social identification processes. *European Journal of Social Psychology*, pages 30, 83–100.

Scherer, K. R. (1979). 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.

Shetty, J. and Adibi, J. (2005). 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. ACM.

Strzalkowski, T., Broadwell, G. A., Stromer-Galley, J., Shaikh, S., Taylor, S., and Webb, N. (2010). Modeling socio-cultural phenomena in discourse. In *Proceedings of the 23rd International Conference on COLING 2010*, Beijing, China. Coling 2010 Organizing Committee.

Wartenberg, T. E. (1990). *The forms of power: from domination to transformation*. Temple University Press.

Yeh, A. (2000). 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. Association for Computational Linguistics.

Yeh, J.-Y. and Harnly, A. (2006). Email thread reassembly using similarity matching. In *CEAS 2006 - The Third Conference on Email and Anti-Spam, July 27-28, 2006, Mountain View, California, USA*, Mountain View, California, USA.