Phrases That Signal Workplace Hierarchy

Eric Gilbert
School of Interactive Computing & GVU Center
Georgia Institute of Technology
gilbert@cc.gatech.edu

ABSTRACT

Hierarchy fundamentally shapes how we act at work. In this paper, we explore the relationship between the words people write in workplace email and the rank of the email's recipient. Using the Enron corpus as a dataset, we perform a close study of the words and phrases people send to those above them in the corporate hierarchy versus those at the same level or lower. We find that certain words and phrases are strong predictors. For example, "thought you would" strongly suggests that the recipient outranks the sender, while "let's discuss" implies the opposite. We also find that the phrases people write to their bosses do not demonstrate cognitive processes as often as the ones they write to others. We conclude this paper by interpreting our results and announcing the release of the predictive phrases as a public dataset, perhaps enabling a new class of status-aware applications.

Author Keywords

computer-mediated communication (CMC), email, natural language processing (NLP), text, status, power

ACM Classification Keywords

H5.3. Group and Organization Interfaces; Asynchronous interaction

INTRODUCTION

> **Email 1:**
> Please take a look at these spreadsheets and calc the gas usage by plant and by pipe in CA. Mike is telling us that most of these palnts [sic] will be shutting down in the next few weeks due to credit exposure. Let's discuss the impact on sendouts. Thanks.
>
> **Email 2:**
> Thank you! The itemization was absolutely no problem, and please let me know when I can do things like that to make your job go more smoothly. I know the market got chaotic late yesterday . . . So I thought I'd ask in the future, is it you I should come to, or real-time? Thanks again for your help.

Which email message comes from someone's boss and which goes to someone's boss? In the first message, we see softened calls to action in "please take a look" and expectations of future work in "let's discuss." In the second, we see confidence exuded in "absolutely no problem," offers of help in "please let me" and hedging in "so I thought." As you may have already guessed, the first message comes from the boss. This paper is about email phrases like these and what they reveal about corporate hierarchies.

Despite years of new social media platforms and experiences, email is still central to how we communicate, especially at work. Nielsen recently reported that Americans use email for a third of all their online communication [20]. Email is the most frequent mode of communication on mobile devices [20]. In a 2008 study, 37% of respondents said they check their work email "constantly," up from 22% in 2002 [17]. With smartphones now everywhere, we can only imagine this figure has gone up.

At the same time, email is not only a place where we chat and exchange information: it is a performance [10]. At work, we have a place within the hierarchy. We have bosses and perhaps people who work for us. The people around us expect that we act like someone who occupies the role. Bosses ask for things; employees provide them. And yet the boss versus employee dichotomy is a false one: we can occupy either role depending on who's around. At work, email is the performance of power and hierarchy captured in text.

In this paper, we search for signs of hierarchy in the phrases people use in email. We closely study the particular words and phrases people send to those above them in the corporate hierarchy versus those at the same level or lower. We adopt the Enron email corpus as our dataset, coupling it with a dataset of Enron employee job titles. By applying penalized logistic regression, we tease apart the relative power of certain words and phrases to signal hierarchy within workplace email. We find that certain phrases are strong predictors, such as the hedge "thought you would" (an upward phrase) and the aforementioned "let's discuss" (a lateral or downward one). Other intuitively good predictors of hierarchy, like "glad to" or "can be reached," carry little weight. Surprisingly, and perhaps disturbingly, we also find that upward phrases do not show evidence of active thinking as often as downward or lateral ones.

After presenting and reflecting on the most powerful phrases for signaling hierarchy, we announce the release of all 7,222 phrases (and associated β weights) as a public dataset. We hope to do for power and hierarchy what LIWC [23] has done for so many other categories. We also think the phrases dataset may lay the groundwork for status-aware CMC applications. For instance, an email client might analyze the content of your messages and notify you differently based on the inferred rank of the person sending the message.

CSCW'12, February 11–15, 2012, Seattle, Washington, USA.

RELATED WORK

Next, we review related work on power and hierarchy in the workplace. We also discuss analytic efforts similar to ours: work aimed at extracting socially relevant information from text. Finally, we conclude this section with research sparked by the Enron email corpus.

Power and Hierarchy

We focus on two bodies of research most relevant to this work: hierarchy and power in CSCW and linguistics research. From its earliest days, CSCW has been concerned with the relative power of individuals collaborating over networked systems (e.g., [3, 30]). For example, [3] reports on the role of power and status in an early CSCW system called *The Coordinator*. In recent years, we've seen power and hierarchy in the emerging social computing literature [2, 28]. For example, researchers have looked at Wikipedia through the lens of power, where people exercise it informally by marking territory with templates [28] and formally through the Wikipedia bureaucracy [2].

Social structures like power also leak into the words we use. (See [31] for an overview from a linguistic perspective.) For example, managers often employ directives (as might be expected), but also wrap those directives in hedging phrases to make them more palatable to those under them (e.g., "when you have time" as a euphemism for "now"). For years, researchers accepted the common wisdom that men use directives more when talking to subordinates, but recent work has shown that women use just as many when put in similar contexts [32]. Bosses will often inject humor to soften the blow of their words and to build loyalty [12]. They also use collective talk (e.g., "let's all give it a try") to build support for themselves as leaders [33]. We look for evidence of this theory later in the paper when we examine the structure of the most predictive phrases.

Processing Text for Social Information

Social scientists have been interested in the interpersonal dimensions of text for decades. Much of this work, including the well-known LIWC [23], descends from Harvard's General Inquirer [27], a dictionary for measuring social science concepts in unstructured text. In recent years, researchers have applied more refined and targeted dictionary techniques (e.g., [6, 9, 11]). For example, in [6] the authors demonstrate that a dictionary-based method can compute happiness over a wide variety of modern text corpora, like blogs.

Over the last decade or so, roughly corresponding to the rise of the social internet, the natural language processing community has also moved into this space. Whereas the methods above employ dictionaries vetted by experts, machine learning research applies algorithms to learn its social concepts directly from data. Most notably, techniques for inferring sentiment have exploded. (See [22] for a review.) Meta projects like SentiWordNet have fused the dictionary and machine learning approaches, generating reusable dictionaries by overlaying many experiments that predict the same dependent variable (i.e., sentiment) [7]. Our work follows in this tradition: we aim to learn from existing data and produce a reusable dictionary of power and hierarchy.

CEO	President
Vice President	Director
In-House Lawyer	
Manager	Trader
Specialist	Analyst
Employee	

Figure 1. A visual depiction of the hierarchy of Enron job titles. We use the job titles of senders and recipients to determine whether an email goes up or down the hierarchy.

The Enron Corpus

The purchase of the Enron corpus after the company's collapse [15] sparked many new email studies. ([25] presents the corpus's basic descriptive statistics.) Using the corpus, researchers have inferred important nodes in social networks [26], improved spam filtering [18] and developed new NLP techniques for name resolution [4].

We are not the first to search the corpus for signs of power and hierarchy. Relevant to the present work, [5] and [24] show how social network features (computed across the network inferred from all messages) can signal power relationships. In [19], the researchers show that a small set of unigrams (single words) have predictive information for inferring power relationships. [19] inspired this work. We build on it by closely studying the power of words and phrases, aiming for insight into why and how people construct hierarchy through CMC. We think this approach (i.e., features rather than black box accuracy) is more relevant to the CSCW community.

METHOD

To search for hierarchy in text, we turn to two datasets: the Enron email corpus [15] and an Enron job titles dataset. The Enron email corpus is the only large email dataset available to researchers. It contains 517,431 email messages sent by 151 people over the span of nearly four years [25]. The job title dataset, formed from trial documents by researchers at Johns Hopkins and USC[1], contains titles for 132 Enron employees. For example, it maps *Jeffrey Skilling* to *CEO* and *Michelle Lokay* to *Employee, Administrative Assistant.*

Pairing this dataset with an account of the ranks of each job title within Enron's corporate culture [21], we were able to fit each employee into a rank within the company. Figure 1 presents the relative ranks of job titles. CEOs and Presidents have the highest rank; Vice Presidents and Directors report to CEOs[2]; In-House Lawyers follow next; Managers and Traders form the next two levels; Specialists and Analysts sit at the hierarchy's second lowest level, above Employees. By combining all three sources of data, we can say that an email

[1] http://www.isi.edu/~adibi/Enron/Enron.htm

[2] We made custom rules for Sally Beck, Rod Hayslett, Rick Buy and Jeff Dasovich, all of whom held special positions within Enron. Dasovich was an appointed liaison between Enron and government investigators; we discarded his email entirely.

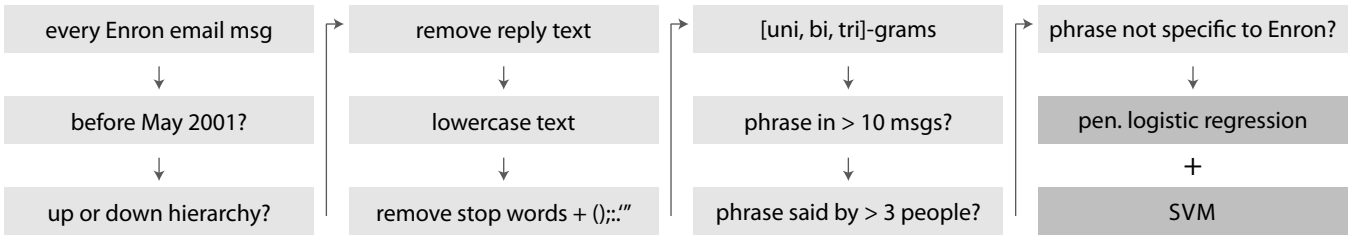

Figure 2. A map of the steps taken in this paper to prepare text for modeling. All phrases go to the penalized logistic regression model and the SVM. We cover these steps in detail in the main body of the paper, but include this figure for overview and reference.

from Michelle Lokay to Jeffrey Skilling went six levels up the corporate hierarchy.

The unit of analysis in this paper is an individual email message combined with status information. We will treat an email as doing one of three things: traveling up the hierarchy (e.g., Manager to VP), staying at the same level (e.g., Employee to Employee) or going down the hierarchy (e.g., CEO to Trader). Furthermore, the models presented here will try to predict whether an email message simply goes up the corporate hierarchy or not. In other words, we will treat email messages which stay at the same level or go down the hierarchy as the same. This simplification means that we can apply traditional statistical techniques but still induce an ordering of employees. In other words, we can reproduce the corporate hierarchy without making our models unnecessarily complex.

The Enron Confound

This entire paper hinges on the idea that our models reproduce a power relationship between two people in a company. And yet the models build on data from a profoundly dishonest company which ultimately fell apart. At the same time, the Enron email corpus is without parallel in the research community. Nowhere else can you find such a rich, complex and naturally occurring email dataset.

Therefore, we have taken steps to guard against this problem. It seems reasonable that up until a certain point—-the point when everything started to fall apart—-Enron behaved like a normal company internally. Even if certain people knew of or suspected malfeasance, it seems likely that they behaved towards their coworkers the way other people behave toward their coworkers. The trick is finding the point after which all of that may have changed. For example, perhaps as word got out regarding how executives had steered the company into the ground, lower level employees started looking for parachutes and challenging the authority of their bosses.

After reviewing a history of Enron [13], we decided to discard all data after May 1, 2001. The SEC did not launch its investigation until many months later in October (the first by any agency). Enron executives did not begin to sell their stock until later—something we only learned after Enron's fall. By contrast, in February 2001 Fortune magazine named Enron the "most innovative company in America" and Enron executives gave well-received presentations proclaiming the company's bright future. With the private selling of Enron stock by executives, we think May 1 was a sea change moment: executives admitted (to themselves) that Enron would probably collapse. Before, on the other hand, it seems likely that Enron employees behaved toward one another the way people do in countless other companies.

Preparing the Email Text

We include in our corpus only those email messages which clearly go up the Enron hierarchy or clearly do not. In other words, we label an email message as *upward* only when every recipient outranks the sender. Conversely, we label an email message as *not-upward* only when every recipient does not outrank the sender. (We make use of both *To:* and *Cc:* headers.) At first it seemed attractive to allow mixtures of ranks (e.g., an email from a Trader to both a VP and a Specialist) because it results in a bigger dataset. But it is possible that the sender speaks differently to each person, perhaps even addressing each one individually. This would confuse any model and cloud the results. While we approach classifying messages conservatively, the upside is that we have greater confidence in the phrase findings. After filtering this way and removing duplicate messages (the Enron corpus has many duplicates), our corpus has 2,044 email messages.

From each message, we discard any text that looks like a quoted reply or a forwarded message by searching for conventional textual markers (e.g., *Original Message*, *— Forwarded by* and lines beginning with >). Next, we convert all text to lowercase and remove the punctuation characters *();:.'"*, while letting the punctuation characters *?!,* remain. We allowed these particular characters to stay because we hypothesized they may signal hierarchy, whereas we thought others—like the period—would not. Keeping punctuation marks can sometimes degrade model performance. For example, keeping all punctuation marks would partition predictive power between "however" and "; however." Therefore, we use punctuation sparingly. From here we adopt a trigram "bag of words" model, a common approach in computational linguistics. That is, we use single words (unigrams), two-word phrases (bigrams) and three-word phrases (trigrams) as the inputs to our models.

However, we cannot include every possible unigram, bigram and trigram. This is not because there would be too many, but because many phrases subtract information or even endanger the validity of the results. Following convention, we discard all phrases consisting solely of "stop words" like "at," "it" and "here." We also look for phrases too obscure to matter in other domains or in other companies: we discard any phrase which does not appear in at least ten email messages, ensuring that we build models only around common English phrases.

More subtly, a phrase could occur at least ten times, but only matter to energy companies like Enron. We could not devise a way to handle this case automatically with code, so two independent raters familiar with Enron looked over every phrase and marked any that appeared specific to Enron's business. Both raters followed the company during its demise, read a history of the scandal and had reviewed trial documents. (We explored using the Google 1T corpus [1] to mark Enron-specific phrases, but it measures web text—not email text—and produced poor results.) Because our dataset is so large and so asymmetric (both in terms of proportions and weights assigned to mistakes), traditional metrics of agreement like Cohen's κ are inappropriate. Instead, we again take a conservative approach: any phrase marked by either rater as Enron-specific was deleted from the list of possible phrases. This process removed proper names like "Frank" and "Stacey," as well as phrases like "gas markets."

Since our models will predict relationships, we also have to guard against phrases that uniquely identify a single relationship. For example, a particular trader and assistant might use the phrase "transfer the memo" uniquely across the corpus. Any model would interpret "transfer the memo" as an important phrase, even though it only serves to identify the trader and assistant. To guard against this, we discard any phrase not written by at least three different people. After these steps, our models have 7,222 different phrases available to them. Figure 2 presents an overview of the entire process.

Statistical Methods

In this paper, we employ two techniques to predict the dependent variable *upward*. Each one has a different objective. Our statistical technique, penalized logistic regression [8], allows us to determine the relative importance of each phrase for predicting hierarchy. Implemented in the R package `glmnet`[3], penalized logistic regression predicts a binary response variable while guarding against the collinearity of phrases, something traditional logistic regression does not do. This is important in our context since English phrases exhibit highly correlated behavior: the word *cone* will appear after the phrase *ice cream* much more often than the word *house*. This implementation of penalized logistic regression handles correlated predictors by shifting most of a coefficient's mass into the most predictive feature, often leaving others out of the model all together. The implementation also handles sparsity well, an important feature in natural language processing where a message only contains a small percentage of the possible phrases.

In the penalized logistic regression model, we include fixed effects for the sender of each message. We instantiate this by including dummy variables in the model representing each sender. This means that any predictive power assigned to the phrases comes after controlling for who said it. The fixed effects also guard against "catchphrases," phrases often associated with one person more than anyone else. In the Results section, we use a model comprised only of sender variables as a substitute for the null model to make a stronger claim about the predictive utility of email phrases. We cover this in more detail later.

The penalized logistic regression model allows close inspection of the relative power of the phrases. Yet, it is not the strongest purely predictive model. It cannot, for instance, support higher-order interaction terms without considerably more data than we have available to us. (Here, a higher-order interaction might look something like a *thanks* phrase co-occurring with a *best regards* without *see attached*.) To explore how much information phrases have in purely predictive terms, we also employ a Support Vector Machine (SVM) model, validating it using three-fold cross-validation. We use the well-known `SVMlight` implementation[4]. The NLP literature documents many instances where SVMs outperform other machine learning techniques on text [14].

Model	Dev. (Acc. %)	df	χ^2	p
Null	2,722.66	0		
Senders-only	1,628.60	98	1,094.10	$< 10^{-15}$
Phrases + senders	224.21	974	1,404.39	$< 10^{-15}$
SVM	(70.7%)	7,222	43.35	$< 10^{-10}$

Table 1. A summary of our different model fits for the progression of models in the Results section. *Null* refers to an intercept-only model. *Dev.* refers to deviance, a measure of the goodness of fit similar to the better-known R^2.

RESULTS

Instead of comparing our phrases model against a null model (i.e., an intercept-only model), we use as our baseline a model that only knows a message's sender. Whereas the null model has deviance 2,722.66, the sender-only penalized logistic regression model has deviance 1,628.6. (The deviance is related to a model's log-likelihood. It is an analog of the R^2 statistic for linear models.) The difference in deviances approximately follows an χ^2 distribution. Simply knowing the sender of an email message provides considerable explanatory power: $\chi^2(98, N{=}2{,}044) = 2{,}722.66 - 1{,}628.6 = 1{,}094.1$, $p < 10^{-15}$.

Adding phrases to the penalized logistic regression model, we find that the model undergoes another dramatic reduction in deviance. The model containing both the phrases and the senders has deviance 224.21. Comparing this to the sender-only model above, the phrases model has significantly more explanatory power: $\chi^2(974 - 98, N{=}2{,}044) = 1{,}628.6 - 224.21 = 1{,}404.39$, $p < 10^{-15}$. The phrases add considerable predictive information after controlling for the identity of the sender (i.e., after controlling for fixed effects). The `glmnet` implementation of penalized logistic regression only activates those variables which have an effect on the dependent variable. As with most computational linguistics work, most phrases do not affect *upward*: only 974 of the 7,222 possible phrases have coefficients significantly different from zero (at the 0.001 level). Table 1 presents a summary.

Table 2 presents the 100 phrases with the most positive β weights. Table 3, on the other hand, shows the 100 phrases

[3] `http://cran.r-project.org/web/packages/glmnet`

[4] `http://svmlight.joachims.org`

↑ phrases	β	↑ phrases	β
the ability to	6.76	attach	6.72
I took	6.57	that we might	6.54
are available	6.52	the calendar	6.06
kitchen	5.72	can you get	5.72
thought you would	5.65	driving	5.61
, I'll be	5.51	thoughts on	5.51
looks fine	5.50	shit	5.45
voicemail	5.43	we can talk	5.41
tremendous	5.27	it does	5.21
will you	5.17	involving	5.15
left a	5.07	the report	5.04
I put	4.90	please change	4.88
you ever	4.80	issues I	4.76
I'll give	4.69	is really	4.65
okay ,	4.60	your review	4.56
to send it	4.48	europe	4.45
communications	4.38	weekend .	4.35
a message	4.35	have our	4.33
one I	4.28	interviews	4.28
can I get	4.28	you mean	4.26
worksheet	4.15	haven't been	4.10
liked	4.07	me . 1	4.07
I gave you	3.95	tiger	3.94
credit will	3.88	change in	3.88
you make	3.86	item	3.84
together and	3.82	a decision	3.82
have presented	3.78	a discussion	3.74
think about	3.71	sounds good	3.65
lot to	3.64	units	3.62
bills	3.61	you are the	3.61
october	3.57	proceed	3.56
keeping	3.55	agreement for	3.50
anything we	3.49	you have an	3.47
don't know what	3.47	february	3.44
the email	3.43	do we want	3.40
in the process	3.34	me or	3.32
head	3.29	, yes ,	3.24
be a great	3.21	case of	3.18
be my	3.17	remedy	3.16
administration	3.15	invite you	3.13
worked on	3.12	conflict	3.11
is doing	3.11	by our	3.11
compensation	3.10	asked if	3.08
candidate	3.08	that night	3.07
this afternoon	3.05	listed	3.04
thanks a	3.03	excellent	3.00
you may	3.00	were pulling	2.99
here's	2.99	factor	2.96
change my	2.95	final draft	2.95
looked at	2.95	wed	2.93

Table 2. The 100 most powerful phrases for predicting that an email message goes up the corporate hierarchy. The table flows left to right, then top to bottom. All phrases are significant at the 0.001 level.

↔↓ phrases	β	↔↓ phrases	β
have you been	-8.46	to manage the	-6.66
you gave	-6.64	let's discuss	-5.72
we are in	-5.44	publicly	-5.24
title	-5.05	promotion	-5.02
need in	-4.80	good one	-4.62
opened	-4.57	determine the	-4.47
initiatives	-4.38	is difficult	-4.36
. I would	-4.34	man	-4.26
we will probably	-4.12	number we	-4.11
any comments	-4.06	contact you	-4.05
you said	-3.99	the problem is	-3.97
I left	-3.88	you did	-3.78
can you help	-3.68	cool	-3.54
send this	-3.47	your attention	-3.44
whether we	-3.44	to think	-3.44
the trade	-3.40	addition to the	-3.30
and I thought	-3.28	great thanks	-3.24
should include	-3.19	selected	-3.16
please send	-3.14	ext	-3.13
existing	-3.06	and let me	-3.05
mondays	-3.02	security	-3.01
presentation on	-2.95	got the	-2.94
let's talk	-2.94	get your	-2.88
the items	-2.78	this week and	-2.77
i hope you	-2.77	team that	-2.75
did it	-2.75	a deal	-2.71
test	-2.69	yours .	-2.68
be sure	-2.65	briefing	-2.60
fri	-2.53	notes	-2.51
forgot to	-2.50	funny	-2.48
confirmations	-2.45	sessions	-2.43
pay the	-2.39	your group	-2.37
implement	-2.35	resolve	-2.34
would need to	-2.34	will be making	-2.33
enter into	-2.32	numbers and	-2.28
and i discussed	-2.28	are you	-2.27
should look	-2.27	calendar	-2.27
helping	-2.26	email	-2.24
doing a	-2.21	to suggest	-2.19
use the	-2.19	the confusion	-2.19
and I am	-2.17	fyi I	-2.16
months to	-2.15	in charge	-2.15
look for	-2.13	meeting will	-2.10
fyi ,	-2.09	that we should	-2.06
know this	-2.05	sent this	-2.04
confirming	-2.03	give me	-2.03
included on	-2.00	prior to the	-1.99
problem with	-1.90	, I thought	-1.89
location	-1.89	supposed to be	-1.88
you take a	-1.85	. I just	-1.83

Table 3. The 100 most powerful phrases for predicting that an email message *does not* go up the corporate hierarchy. (Table flows same as Table 1; all phrases significant at the 0.001 level.)

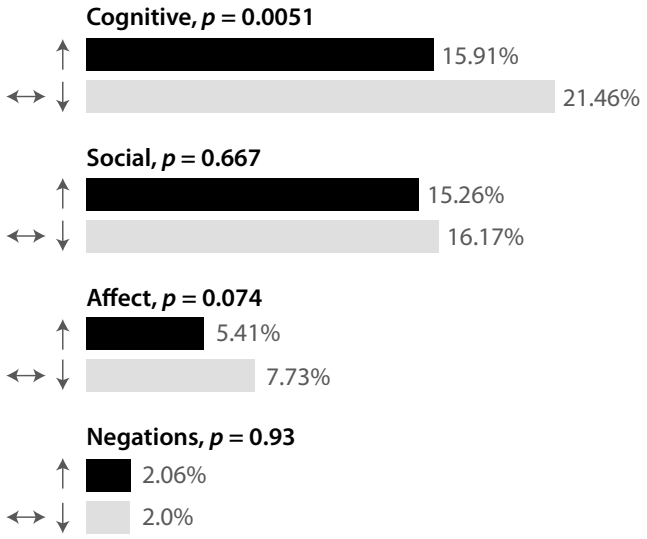

Figure 3. Testing the predictive phrases for structure. We use the LIWC program to test phrases for membership in seven categories. Four are shown here. After a Bonferroni correction, we find that the phrases you say to your boss do not demonstrate cognitive processes as often as the phrases you say to others.

with the most negative β weights. (We have deleted the senders from these lists to showcase the more generalizable set of phrases. 23 senders appeared in the top 100 positive predictors; 12 senders appeared in the top 100 negative predictors.) Tables 2 and 3 affect a message's likelihood of having gone *upward* most strongly—when they appear. It is important to note that a phrase's β corresponds not only to its discriminative power, but also to its obscurity. Recall that we limited the set of possible phrases to only those that appeared in at least ten messages. We should expect that more obscure phrases (i.e., longer phrases) will have higher β coefficients since they occur fewer times and therefore may more easily skew toward one side or the other.

Do the phrases in Tables 2 and 3 cluster together in some meaningful way? To explore this question, we used the LIWC program [23] to compute categories to which the predictive features belong. After reviewing the LIWC categories, we picked seven we hypothesized might relate to workplace hierarchy: *Cognitive Processes, Social, Affect, Negations, Certainty, Money* and *Assent*. Due to these seven simultaneous tests, we allow for a Bonferroni correction, letting $\alpha = 0.05/7 = 0.0071$. Figure 3 presents proportions of membership in four of the LIWC categories along with p-values associated with the corresponding χ^2 test. We find that *Cognitive Processes* yields the only non-random result: 21.46% of the *not-upward* phrases belong to this category while only 15.91% of the *upward* phrases do, $\chi^2(1, N{=}974) = 7.83$, $p = 0.0051$. The three categories not shown in Figure 3—*Certainty, Money* and *Assent*—have random results, $p = 0.88$, $p = 0.34$ and $p = 0.73$, respectively.

Figure 4 provides a deeper view into the structure around the predictive phrases. It shows Word Tree visualizations [34] of searches for "thank" and "talk" in the two halves of the corpus. For example, Figure 4 shows that the phrase "thank you for your" appears many times in the corpus across multiple email messages. We think this is an intriguing, deeper look into the text behind our statistical techniques. The visualizations come from the online site Many Eyes [29].

Support Vector Machine Approach

The penalized logistic regression model allows us to inspect which phrases have the most impact on the dependent variable *upward* while controlling for senders' identities. However, as noted earlier, this approach does not allow us to get the most predictive information out the phrases. We now turn to an SVM to see how well the phrases can predict hierarchy when we give up introspection and let phrases interact in high-dimensional (and opaque) ways.

The SVM here does not have access to the identity of the sender; it only has access to the phrases. The reason behind this is that an SVM would project lots of predictive information onto the identities (i.e., it would know the *upward* baselines for each person). The SVM performs with 70.7% accuracy under three-fold cross-validation, a non-random improvement over the baseline of 60.5%, $\chi^2(1, N{=}1{,}900) = 43.35$, $p < 10^{-10}$. The 60.5% baseline corresponds to picking the most likely class, *not-upward*, every single time. We employ *three-fold* cross-validation—instead of a more standard technique like ten-fold cross-validation—because it allows us to collapse all relationships in the test set onto a single prediction. By way of an example, suppose in the test set the same relationship occurs three times. Our SVM makes an independent prediction for each one, and they are often different. In a post-hoc step, we unify all predictions about the same relationship by voting.

DISCUSSION

We find that certain words and phrases signal hierarchy while others do not. Tables 2 and 3 provide an intriguing look into how we express power and hierarchy through email at work. As is perhaps to be expected, "attach," a top predictor of *upward* ($\beta = 6.72$), suggests that the majority of documents flow up through organizations. "thought you would," also a top *upward* predictor ($\beta = 5.65$), has the ring of hedging and politeness. Some words unexpectedly signal one way or the other. For instance, "weekend" surprisingly suggests an *upward* email—perhaps used to signal how hard the employee works. "sounds good," on the other hand, unsurprisingly predicts a *upward* relationship.

> Hi Carol, my home number is [redacted] — **weekends** [$\beta = 4.35$] would work fine so give me a call any time. (↑76)

> Louise, thanks for your reply to my email. The fixes implemented over the **weekend** appear to have worked extremely well. I am sure that the changes your team made also contributed to the record 147 external and 263 total trades that I executed today. Thanks for all your help. (↑111)

> Hey, sorry I haven't read my e-mail in the past day and a half. Too busy selling, but that **sounds good** [$\beta = 3.65$] Fletch. (↑778)

(In the quotes above and below, the number corresponds to a message's place in either the *upward* or *not-upward* dataset.) Many of the most predictive *not-upward* phrases relate to checking on a subordinate's current status, exemplified by "have you been" and "I hope you." "I hope you" subtly communicates expectations from a boss to a subordinate, whereas "have you been" is more direct.

Figure 4. Word Tree visualizations of searches in the two halves of our dataset. At top, a visualization depicting a search for "thank" in only the *upward* part of our corpus. The visualization shows phrases that branch off from "thank" across every message in this part of the corpus. A larger font-size means that the word occurs more often in the corpus. "thank" appeared in the *upward* part of the corpus 16 times and is present in the most powerful predictors listed in Table 2. At bottom, a Word Tree visualization of a search for "talk" in the *not-upward* part of our corpus. Here, we show all phrases that terminate in "talk," with "need to talk" and "like to talk" the most likely trigrams. "talk" appeared in the *not-upward* part of the corpus 64 times. The visualizations come from the site Many Eyes.

> Did Tim already send you these documents for your review? If he did, **have you been** [β = -8.46] working with the Houston group on them or do you want to tell me about any problems? (↔↓412)
>
> How **have you been**? I have not had much of your mail lately, but am guessing this one is yours. Have a great weekend! (↔↓1,080)
>
> What **have you been** smoking? (↔↓716)
>
> Dan, **I hope you** [β = -2.77] haven't wasted much time on this one so far. Let me know where this stands. C (↔↓65)
>
> Aaron, I sent you a file a few days ago. **I hope you** could open it. Vince (↔↓1,188)

A limitation in our text-cleaning algorithms highlighted an unexpected finding. Many messages are tagged at the end with a short version of the message's date, outside the traditional *Date:* email header. Coming right after the message's signoff (e.g., "Best regards, Jim"), the models had access to these words. Wednesday's ("wed," β = 2.94) saw more messages going *upward*, while Fridays ("fri," β = -2.53) saw more messages going the other way. This is a finding we would like to see examined in more detail in the future.

A few of the words and phrases in Tables 2 and 3 raised concern. For example, consider "October" (β = 3.57). In reviewing where it appears in the corpus, it often occurs in the context of travel:

> Jeff, Christ, Mark and myself are planning to visit Tom Piazze in **October**. I talked to Christy about Wharton and she will be calling Tom to set it up. (↑661)

Despite our efforts to control for it, some of the phrases may reflect Enron culture more than corporate culture. At the same time, perhaps "October" signals an uptick in travel associated with the new fiscal year and the money that comes with it. Future work may need to address these tricky issues.

Figure 3 reveals a surprising and somewhat concerning finding: people demonstrate active thinking more often to those below them than they do to their bosses. (Depending on your perspective, however, this isn't surprising at all—perhaps only surprising that NLP can detect it.) Now, that means that LIWC sees less evidence in the text of actively working things out in email, as measured by its internal word-stem list named *Cognitive Processes*. It seems reasonable to think that people only go to their bosses when they have an answer, not when they want to work out the answer. In pouring over our data, we also noticed a seemingly disproportionate number of misspellings in the *not-upward* emails. The example that opens this paper contains one. We suggest that future researchers explore misspellings as a feature: they may indicate someone has little incentive to proofread.

While we find reliable signals of hierarchy in certain phrases, the SVM did not perform particularly well. We see two possible explanations for it. First, a good amount of predictive information lies elsewhere: in the interaction timing patterns, social network data, etc. Or, we constrained the data so tightly that the SVM has a hard time not over-training. This seems more likely to us, because when we expand the dataset to go past May 2001 (yielding 11K messages total), the SVM's accuracy goes to 91%. In our focus on the phrases, we probably hamstrung the SVM.

not-[↑↔↓] phrases

have been working	opportunity to	agreement	comments
in the meantime	we discussed	remember	customers
I am forwarding	can't believe	available	suggested
can be reached	the meeting	decision	everyone
would like to	interested in	exceeds	private
in the office	your new	money	annual
take a look	can talk	suite	wait

Table 4. A small sample of phrases with no power to signal hierarchy (i.e., β = 0). While many seem like intuitively good predictors, the predictive power lies in phrases like those in Tables 2 and 3.

Phrases Dataset

We have released the predictive phrases that form the basis of the penalized logistic regression model[5]. We extracted the phrases in Tables 2, 3 and 4 from it. It contains every phrase available to the penalized logistic regression model plus the identities of the senders in the corpus. Each line in the file corresponds to a phrase and its β weight, sorted from smallest to largest. Most of the phrases have zero β weights; we include them so that researchers can also see which words and phrases *do not* signal hierarchy. Table 4 presents a sample of these β = 0 phrases.

Theoretical Implications

Even today, most CMC generates text [16]. Methods that can transform raw CMC text into meaningful inferences about social life are valuable to researchers. Compared to things like positive and negative sentiment, we know very little about how people express power and hierarchy through CMC. We hope that this work open lines of research in the CSCW community unavailable before. For example, imagine a study investigating who in a organization disproportionately attracts *upward* language. Do they move up the ladder faster?

By releasing the model's predictive phrases, we hope to do for power and hierarchy what LIWC [23] has done for other categories: reveal hidden social processes by analyzing the text we write to one another. While CSCW research has often incorporated power and status in its studies, we have not had a way to scalably operationalize that concept. That's the reason we performed such a close study of the words and phrases behind hierarchy: we wanted to build a portable dictionary capable of operationalizing hierarchy across organizations.

We studied a formal corporate hierarchy in this paper, marked by clear reporting lines. You might imagine that by applying the phrases in this paper to a corpus of workplace email you could discover informal power and reporting structures. Sometimes power exists in unlikely, undocumented places within companies. The phrases dataset may highlight them.

Design Implications

We also believe this work may enable a new class of status-aware applications. For example, imagine a mobile email client that can guess when somebody above you in the hierarchy just sent you a message. You might set the client to

[5] http://comp.social.gatech.edu/hier.phrases.txt

interrupt you at any time after 5 p.m. if the message comes from above you in the hierarchy (i.e., you might want to seem always available to them). Otherwise, however, you might set the client to hold notifications until after business hours. These notions of power and hierarchy are so crucial within workplace applications, yet no current tools know about them without considerable manual effort: you would have to input the org chart. Furthermore, organizations change all the time, and a software provider would want to build a product that scales to many companies without rebuilding the application for every client. We believe this textual work could form the basis of such a product.

However, more work may need to be done. As our SVM results showed, there is probably predictive information outside the text. Perhaps future work building on these purely textual results could improve the accuracy. Predictive features like response time, social network descriptors and amount of text come to mind. We focused on text because it is so universal and an application would have access to it (within your email archives). An application might not have access to the entire corporate network structure, on the other hand.

CONCLUSION

We believe this work addresses a fundamental force in the workplace: hierarchy. In this paper, we have revealed phrases which signal power and hierarchy in workplace email. These phrases shine a light on the ways CMC captures embedded and often undiscussed aspects of social life. We hope other researchers will be able to address new theoretical questions using the results in this paper, perhaps by making use of the phrases made available as a public dataset. Our results may also lay the groundwork for new status-aware applications, such as email clients that can differentially notify you of new messages based on the rank of the sender.

ACKNOWLEDGEMENTS

We would like to thank the comp.social lab at Georgia Tech for reviewing early drafts of this work. Aaron Bobick contributed valuable feedback on our analytic approach. Google helped support this work.

REFERENCES

1. T. Brants and A. Franz. Web 1T 5-gram Version 1. *Linguistic Data Consortium, Philadelphia*, 2006.
2. M. Burke and R. Kraut. Mopping up: Modeling wikipedia promotion decisions. In *Proc. CSCW*, pages 27–36, 2008.
3. R. Carasik and C. Grantham. A case study of cscw in a dispersed organization. In *Proc. CHI*, pages 61–66, 1988.
4. C. Diehl, L. Getoor, and G. Namata. Name reference resolution in organizational email archives. In *SIAM International Conference on Data Mining*, pages 20–22, 2006.
5. C. Diehl, G. Namata, and L. Getoor. Relationship identification for social network discovery. In *Proc. NCAI*, volume 22, page 546, 2007.
6. P. Dodds and C. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, pages 1–16, 2009.
7. A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining.
8. J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.
9. E. Gilbert and K. Karahalios. Widespread Worry and the Stock Market. In *Proc. ICWSM*, 2010.
10. E. Goffman. *The Presentation of Self in Everyday Life*. 1959.
11. J. Hancock, C. Landrigan, and C. Silver. Expressing emotion in text-based communication. In *Proc. CHI*, pages 929–932, 2007.
12. J. Holmes and S. Schnurr. Politeness, humor and gender in the workplace: negotiating norms and identifying contestation. *Journal of Politeness Research. Language, Behaviour, Culture*, 1(1):121–149, 2005.
13. Z. Jelveh and K. Russell. Interactive timeline: The rise and fall of enron. *The New York Times*, 2006.
14. T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, pages 137–142, 1998.
15. B. Klimt and Y. Yang. Introducing the Enron corpus. In *First conference on email and anti-spam (CEAS)*, 2004.
16. A. S. Lee Rainie, Kristen Purcell. The Social Side of the Internet. Technical report, Pew Internet & American Life Project, 2011.
17. M. Madden and S. Jones. Networked Workers. Technical report, Pew Internet & American Life Project, 2008.
18. V. Metsis, I. Androutsopoulos, and G. Paliouras. Spam filtering with naive bayes. In *Proc. CEAS*, pages 125–134, 2006.
19. G. Namata, L. Getoor, and C. Diehl. Inferring formal titles in organizational email archives. In *Proc. of the ICML Workshop on Statistical Network Analysis*. Citeseer, 2006.
20. Nielsen. What Americans Do Online: Social Media And Games Dominate Activity. Technical report, 2010.
21. S. Palus, P. Bródka, and P. Kazienko. How to analyze company using social network? *Knowledge Management, Information Systems, E-Learning, and Sustainability Research*, pages 159–164, 2010.
22. B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.
23. J. W. Pennebaker and M. E. Francis. *Linguistic Inquiry and Word Count*. Lawrence Erlbaum, August 1999.
24. R. Rowe, G. Creamer, S. Hershkop, and S. Stolfo. Automated social hierarchy detection through email network analysis. In *Proc. WebKDD*, pages 109–117, 2007.
25. J. Shetty and J. Adibi. The Enron email dataset: database schema and brief statistical report. Technical report, University of Southern California, 2004.

26. J. Shetty and J. Adibi. Discovering important nodes through graph entropy: the case of enron email database. In *Proceedings of the 3rd international workshop on Link discovery*, pages 74–81, 2005.
27. P. Stone, R. Bales, J. Namenwirth, and D. Ogilvie. The General Inquirer: a computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7(4):484–498, 1962.
28. J. Thom-Santelli, D. Cosley, and G. Gay. What's mine is mine: territoriality in collaborative authoring. In *Proc. CHI*, pages 1481–1484, 2009.
29. F. Viegas, M. Wattenberg, F. Van Ham, J. Kriss, and M. McKeon. Many Eyes: a site for visualization at internet scale. In *InfoVis*, pages 1121–1128. Published by the IEEE Computer Society, 2007.
30. S. Viller. The group facilitator: a cscw perspective. In *Proceedings of the second conference on European Conference on Computer-Supported Cooperative Work*, pages 81–95, 1991.
31. B. Vine. *Getting things done at work: The discourse of power in workplace interaction*. John Benjamins Publishing Company, 2004.
32. B. Vine. Directives at work: Exploring the contextual complexity of workplace directives. *Journal of Pragmatics*, 41(7):1395–1405, 2009.
33. B. Vine, J. Holmes, M. Marra, D. Pfeifer, and B. Jackson. Exploring co-leadership talk through interactional sociolinguistics. *Leadership*, 4(3):339, 2008.
34. M. Wattenberg and F. Viégas. The Word Tree, an interactive visual concordance. In *InfoVis*, pages 1221–1228, 2008.