Email Formality in the Workplace: A Case Study on the Enron Corpus

Kelly Peterson, Matt Hohensee, and Fei Xia
Linguistics Department
University of Washington
Seattle, WA 98195
{kellypet, hohensee, fxia}@uw.edu

Abstract

Email is an important way of communication in our daily life and it has become the subject of various NLP and social studies. In this paper, we focus on email formality and explore the factors that could affect the sender's choice of formality. As a case study, we use the Enron email corpus to test how formality is affected by social distance, relative power, and the weight of imposition, as defined in Brown and Levinson's model of politeness (1987). Our experiments show that their model largely holds in the Enron corpus. We believe that the methodology proposed in the paper can be applied to other social media domains and be used to test other linguistic or social theories.

1 Introduction

Email has become an important way of communication in our daily life. Because of its wide usage, it has been the subject of various studies such as social network analysis (e.g., (Leuski, 2004; Diesner et al., 2005; Carvalho et al., 2007)), deception detection (e.g., (Zhou et al., 2004; Keila and Skillcorn, 2005)), information extraction (e.g., (Culotta et al., 2004; Minkov et al., 2005)), and topic discovery (e.g., (McCallum et al., 2007)). In this study, we focus on email formality in various social settings; that is, we want to determine whether the choice of formality in email communication is affected by factors such as the social distance and relative power between the senders and the recipients.

While an early perspective of email communication held that email is a lean medium which lacks vital social cues (Daft and Lengel, 1986), other work has shown that senders of email exhibit a wide range of language and form choices which vary in different social contexts (Orlikowski and Yates, 1994). Through various theories of sociolinguistics, it is proposed that these changes take place in a predictable manner.

Brown and Levinson (1987) have proposed a model where in order to save the "face" or public self image of the hearer of a message, a speaker can employ a range of verbal strategies. Their model of politeness states that in social situations there are three factors which are considered in a decision whether or when to use communication techniques such as formality:

1. The "social distance" between the participants as a symmetric relation
2. The relative "power" between the participants as an asymmetric relation
3. The weight of an imposition such as a request

Abdullah (2006) examines email interactions from the perspective of Brown and Levinson's politeness model in a Malaysian corporation from over 180 participants and a corpus of 770 email messages. This work directly examines the factors mentioned previously which influence email formality. Unfortunately, the methodology and data were not provided for this study.

The goal of our work is to test whether Brown and Levinson's model holds in a real setting with a much larger data set. In this study, we chose the Enron Email Corpus as our dataset. We first built two classifiers: one labels an email as *formal* or *informal*

Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 86–95,
Portland, Oregon, 23 June 2011.

and the other determines whether an email contains a request. Next, we used the classifiers to label every email in the Enron corpus. Finally, we tested whether the three factors in Brown and Levinson's theory indeed affect formality in email communication. While we consider the work a case study, we believe that the methodology proposed in the paper can be applied to other social media domains and be used to test other linguistic or social theories.[1]

2 Overview of the Enron email corpus

The Enron email corpus, which consists of hundreds of thousands of emails from over a hundred Enron employees over a period of 3.5 years (1998 to 2002), was made public during the US government's legal investigation of Enron. The corpus was first processed and released by Klimt and Yang (2004) at Carnegie Mellon University (CMU), and this CMU dataset has later been re-processed by several other research groups. In this section, we briefly introduce the datasets that we used in our experiments.

2.1 The ISI dataset

The CMU dataset contains many duplicates. It was later processed and cleaned by Shetty and Adibi at ISI and released as a relational database. The ISI database comprises 252,759 messages from the email folders of 150 employees (Shetty and Adibi, 2004).[2] We use the ISI dataset as the starting point for all of our experiments except for the one in Section 5.1.

2.2 The Sheffield dataset

The Enron email corpus contains both personal and business emails. In 2006, Jabbari and his colleagues at the University of Sheffield manually annotated a subset of the emails in the CMU dataset with "Business" or "Personal" categories (Jabbari et al., 2006). The subset contains 14,818 emails and 3,598 of them (24.2%) are labeled as "personal".[3] We use this dataset in the personal vs. business experiment as described in Section 5.1.[4]

2.3 The ISI Enron employee position table

In addition to the ISI database, ISI also provided a table of 161 employees and their positions in the company.[5] In Section 5.3, we study the effect of seniority on the formality of a message, and we use this table to determine the relative seniority between senders and recipients of a given email.

3 Creating the gold standard

In this study, we build two classifiers: a formality classifier that determines whether an email is formal, and a request classifier that determines whether an email contains a request. In order to train and evaluate the classifiers, 400 email messages were randomly chosen from the Enron corpus and manually labeled for formality and request.

3.1 Formality annotation

Formality is a concept which is difficult to define precisely and human judgment on whether an email is formal can be subjective. To determine how much human annotators can agree on the concept, we asked three annotators to label 100 out of the 400 emails with four labels: "very formal", "somewhat formal", "somewhat informal", and "very informal".

Because formality is hard to define, we did not give annotators a concrete definition. Instead, we provided a few guidelines and asked annotators to follow the guidelines and their intuition. One of these guidelines was that the formality of an email should not necessarily be dictated by the relationship between the sender and the recipient if their relationship can be inferred from the message. Another guideline stressed that the nature of an email being business or personal should not necessarily dictate its formality. Other than these guidelines, annotators were asked to come up with their own criteria for formality while doing the annotation.

Table 1 shows the agreement between each annotator pair and the average score of the three pairs. For agreement, we calculated the accuracy, which

[1] Our data including annotations and results can be found at http://students.washington.edu/kellypet/enron-formality/

[2] The dataset can be downloaded from http://www.isi.edu/~adibi/Enron/Enron.htm

[3] The dataset is available at http://staffwww.dcs.shef.ac.uk/people/L.Guthrie/nlp/research.htm.

[4] The ISI dataset and the Sheffield dataset contain significant overlap as both were derived from the CMU dataset, but the former is not necessarily a superset of the latter.

[5] We downloaded the table in January 2011 from http://www.isi.edu/~adibi/Enron/Enron_Employee_Status.xls

Annotator pair	2-way Agreement (Acc/F1)	4-way Agreement (Acc)
A vs. B	87.3/77.8	53.7
A vs. C	85.4/77.2	40.6
B vs. C	84.5/72.9	36.1
Pairwise Ave	85.7/76.0	43.5

Table 1: Inter-annotator agreement for formality annotation

is the percentage of emails that receive the same label from the two annotators. *2-way agreement* means the agreement is calculated after the label *very formal* has been merged with *somewhat formal*, and *very informal* with *somewhat informal*; *4-way agreement* means that the agreement is calculated with the four formality labels used by the annotators. With the 2-way distinction (formal vs. informal), we also calculate the f-score for identifying *informal* emails, treating one annotation as the gold standard and the other as the system output. This table shows that, although the concept of formality is intuitive, the inter-annotator agreement on formality is pretty low (especially when making the 4-way distinction).

Finally, Annotator A, who had the highest agreement with other annotators, annotated the remaining 300 emails, and his annotation was treated as the gold standard for our formality classifier.

3.2 Request annotation

In order to train and evaluate our request classifier, we asked two annotators to go over the same 400 emails and label each message as *containing-request* or *no-request*. A message is considered to contain a request if it is clear that the sender of the message expects the recipient to take some action to respond to the message. For instance, if a message includes a question such as *what do you think?* or a request such as *please call me tomorrow*, it should be labeled as *containing-request* as the sender expects the recipient to call the sender or answer the question. Our definition is slightly different from the definition of *request* used in speech acts, and it can be seen as a synonym of *require-action*.

While some emails clearly contain requests and others clearly do not, there is some gray area in between, which results in the disagreement between the annotators. Many of the disagreed emails include sentences such as *Let me know if you have any questions*. This very commonly used expression is itself ambiguous between the meanings *"Let me know whether you have any questions"* and *"If you have any questions, please inform me of that fact"*. Furthermore, this sentence often appears as a marker of politeness or an offer to clarify further, rather than a request for action. So the correct label of an email containing this expression depends on the context. For the 400 messages, the two annotators agreed on 361 messages, for an inter-annotator agreement of 90.3% and a F1-score of 87.9% for identifying emails that contain requests.

4 Building classifiers

In this section, we discuss the feature sets used for the two classifiers and report their performances.

4.1 Data pre-processing

Before forming the feature vectors for the classifiers, we preprocessed all the emails in the ISI and Sheffield dataset in several steps. First, we removed any replied or forwarded message from the email body as we want to use only the text written by the sender. If the email body becomes empty after this step, the email is excluded from the analysis conducted in Section 5. After this step, the size of the ISI dataset reduces from 252,759 to 232,815 emails, and the size of the Sheffield dataset changes from 14,818 to 13,882 emails. Second, the email messages were segmented into sentences and tokenized with tools in the NLTK package (Bird et al., 2009).

4.2 Formality classifier

For the formality classifier, we use two labels: *formal* and *informal*.

4.2.1 Features for formality

During formality annotation, after the 100 emails had been annotated, the three annotators were asked to provide a few paragraphs describing their criteria for formality. In these criteria, more cues are indicators of informality (e.g., the use of vulgar words) than indicators of formality. We use the following features to capture the informal "style" of the

emails:[6]

F1: Informal Word Features, which check the occurrences of *informal words* (see the next section for detail)

F2: Punctuation Features:

- Exclamation Points ('!')
- Absence of sentence final punctuation
- Frequency of ellipsis ('. . . ')

F3: Case features:

- All lowercase Subject line
- Frequency of sentences which were entirely lower case
- Frequency of sentences whose first word is lower case

4.2.2 Informal words

We designed a simple heuristic method to extract a list of informal words from the Enron corpus. First, we collect all the unigrams in the Enron corpus. Second, we retrieve the information about each unigram from Wordnik,[7] a website that provides access to retrieve word definitions from multiple source dictionaries. Among the several dictionaries crawled by Wordnik, we find Wiktionary to be the best source for our task since its labels on word definitions such as 'informal', 'offensive', 'vulgar', 'colloquial' and 'misspelling' were the most consistent and relevant to our definition of "formality". In addition to these labels, the part of speech category for 'interjection' was also used to determine if a word might be considered informal in email communication. Third, we use the gathered word definitions to determine whether a word is *informal*.

One issue with the last step is that words often have multiple meanings and some meanings are informal and others are not. For instance, the word *bloody* can be formal or informal depending on which meaning is used in an email. As word sense disambiguation is out of the scope of this work, we use some simple heuristics to determine whether a word should be treated as informal or not. In essence, the process treats a word as informal if a large percentage of definitions for the word have certain labels (e.g., *vulgar*, *offensive*, and *misspelling*) or certain part-of-speech tags (e.g., *interjection*).[8]

4.2.3 Performance of the formality classifier

We trained a Maximum Entropy (MaxEnt) classifier in the Mallet package (McCallum, 2002). Table 2 shows classification accuracy and precision, recall, and F1-score for identifying informal emails. The baseline system labels every email as *formal* because 62.7% of the emails in the dataset were annotated as formal; its F1-score is zero as the recall is zero. The numbers for the inter-annotator agreement row are copied from the pairwise average of the 2-way agreement in Table 1. The table shows that, with very few features, the performance of the formality classifier is much better than the baseline and is close to inter-annotator agreement. All three types of features beat the baselines and combining them provides additional improvement.

	Acc	Prec	Rec	F1
Baseline	62.7	-	-	-
Inter-annotator agreement	85.7	89.5	66.8	76.0
F1: Informal words	69.2	75.0	26.7	39.3
F2: Punctuation	74.4	82.5	45.8	58.9
F3: Case features	69.7	80.0	26.5	39.8
F1+F2	76.4	77.3	51.1	61.5
F1+F3	72.8	74.3	39.4	51.5
F2+F3	80.3	85.2	59.7	70.2
F1+F2+F3	80.6	85.7	62.1	72.0

Table 2: Performance of the formality classifier. We use 10-fold cross validation on the 400 emails. Baseline: label every email as *formal*.

4.3 Request classifier

The request classifier uses two labels: *containing-request* and *no-request*.

[6] We did not use ngram features as they might be too specific to the small training data we have and might not work well when applied to other emails in the Enron corpus or emails in other domain.

[7] http://www.wordnik.com

[8] We manually checked the list of informal words extracted and estimated that the number the false positives is less than 1%. However, the list is definitely not complete as many informal words in the Enron corpus do not appear in the dictionaries used by Wordnik.

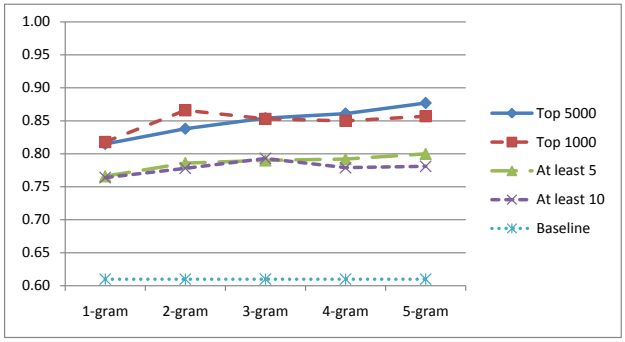

Figure 1: Accuracy of the request classifier with different feature sets

4.3.1 Features for request

There has been considerable research into categorizing email messages by function. Cohen, Carvalho, and Mitchell (2004) described the classification of email into 'email speech acts', building on the speech act theory of Searle (1975). Carvalho and Cohen (2006) achieved high-precision results categorizing messages into categories such as 'request' and 'proposal' when preprocessing the text in certain ways and using unigram, bigram, and trigram features only.

Unlike formality, which is more about the style of the messages (e.g., whether the email is all in lowercase), the content words are more relevant for identifying requests. Following the work in (Carvalho and Cohen, 2006), we used word ngrams as features. To prevent the features from being too specific to the small training data, we experimented with two ways of feature selection: by feature counts and by chi-square scores. N-grams were extracted from the email body only. For pre-processing, in addition to the pre-processing step mentioned in 4.1, we also replaced some name entities (e.g., numbers and dates) with special labels and lowercased the text.

4.3.2 Performance of the request classifier

We trained a MaxEnt classifier and ran 10-fold cross validation on the 400-email dataset. Figure 1 shows the accuracy of the classifier with different feature sets. The bottom dotted line is the baseline result. In the 400 emails, 244 are labeled as *no-request*, so a baseline system that labels everything as *no-request* has an accuracy of 61%. The middle two lines are the accuracy with features that occur no fewer than 5 or 10 times. For the top two curves, features are sorted according to the chi-square scores, and the top one thousand or five thousand are kept. X-axis shows the value of n for word ngrams; e.g., 3-gram means features include word unigrams, bigrams, and trigrams. Figure 1 shows that chi-square scores outperform feature counts for feature selection, and varying the value of n does not affect the accuracy very much.

Table 3 shows classification accuracy and precision, recall, and F1-score for identifying request-containing emails when n is set to 3. The table shows that our classifier, regardless of methods used for feature selection, greatly outperforms the baseline system, and there is a small gap between the performance of our classifier and the inter-annotator agreement. For the rest of our experiment, we will use 3-gram, Top5000 as the feature set for the request classifier.

	Acc	Prec	Rec	F1
Baseline	61.0	-	-	-
Inter-annotator agreement	90.3	90.4	85.5	87.9
Using all features	79.5	76.8	68.0	72.1
At least 5	79.0	75.7	68.0	71.6
At least 10	79.3	75.9	68.6	72.1
Top1000	85.5	88.3	72.4	79.6
Top5000	85.5	88.3	72.4	79.6

Table 3: Performance of the request classifier with 3-gram features: We use 10-fold cross validation on the 400 emails. Baseline: label every email as *no-request*.

5 Factors influencing formality

As mentioned in Section 1, Brown and Levinson (1987) proposed three factors that influence communication choices such as formality: social distance, relative power, and the weight of an imposition. In this section, we test whether these factors indeed affect formality in emails.

We measure social distance in two ways: one is based on the nature of emails (personal vs. business), and the other is based on the number of emails sent from the sender to the recipient. While these aspects do not directly define the social distance between individuals, they are employed to illustrate

related social properties in absence of data which outlines the social distance of all Enron employees. For relative power, we use the rank difference of the positions that the sender and the recipient held in Enron. Since relative power is complex to define without more data, this definition of rank difference serves as one dimension in which we can study relative power. For the weight of imposition, we compare emails that contain requests and the ones that do not.

5.1 Social distance: Personal vs. Business

In general, friends, family and other such personal contacts are presumably closer in social distance than business colleagues. Therefore, it is possible that email messages of a personal nature will be more likely to be informal than those of a business nature. To test the hypothesis, we compare the degree of formality in business vs. personal emails. We use the Sheffield dataset, which contains 13,822 non-empty emails that have been manually labeled as "personal" or "business". We ran the formality classifier on the data, and the results are in Table 4. The first and second columns show the number of emails that are labeled as *formal* or *informal* by our formality classifier, and the last column shows the percentage of emails in that row that are labeled *informal* (a.k.a. *the rate of informality*).

The table demonstrates that the rate of informality in personal emails (56.0%) is indeed much higher than that of business emails (21.3%). We have run the Chi-square test and G test with the counts in the table, and both tests indicate that formality (formal vs. informal) is not independent from the business nature of an email message (business vs. personal) at p=0.0001. The same is true for formality and other social factors that we tested in this section (see Tables 5, 7, 8, and 9).[9]

[9]There are two caveats for using these statistical tests to determine whether two random variables (formality and a social factor) are independent. First, the counts in the tables are based on the output of the two classifiers, which could be different from the real counts. Second, the data points in some experiments were not chosen randomly from the whole email corpus; for instance, the emails in Table 7 were from a small set of people whose ranks at Enron were known.

	Formal	Informal	Inf %
Personal	1410	1793	56.0%
Business	8409	2270	21.3%
Total	9819	4063	29.3%

Table 4: Formality in personal vs business emails, $p < 0.0001$

5.2 Social distance: Amount of contact

Besides the difference in personal and business matters, another way to measure social distance is the amount of contact that two individuals have with each other. People with more email exchange are likely to be closer in social distance than those with less email exchange, and are therefore likely to have a higher rate of informality. To test this hypothesis, we started with the ISI dataset and looked at the subset of emails where an email has exactly one recipient, and both the sender and the recipient are in the enron.com domain. The emails were then grouped into several buckets based on the number of emails from a sender to a recipient.

The results are in Table 5. The first column is the range of the numbers of emails from a sender to a recipient, and the last column is the number of (sender, recipient) pairs where the number of emails that the sender sent to the recipient is in the range specified in the first column. The second column is the total number of *formal* emails from the senders to the recipients in those pairs. The third column is defined similarly, and the 4th column is the rate of informality. Note that the rates of informality in the first two rows are about the same; it might be due to the fact that the Enron corpus contains emails only in a 3.5-year period. The rate of formality does go up in the third and fourth rows.

Emails sent from A to B	Formal	Inf	Inf %	# of pairs
1 to 10	23,423	7,566	24.4%	14,877
11 to 50	11,484	3,558	23.7%	737
51 to 100	3,236	1,363	29.6%	66
101 or more	2,114	1,271	37.5%	21
Total	40,257	13,758	25.5%	15,701

Table 5: Formality and the number of emails from the sender to the recipient, $p < 0.0001$

5.3 Relative power: Rank difference

Another factor that could affect the sender's choice of formality is the relative difference in power or rank between sender and recipient. For example, if a manager sends an email to the CEO of an organization, the email is more likely to be formal than if the recipient has a lower rank than the sender.

To investigate this, we started with the emails in the ISI dataset whose senders are employees appearing in the ISI Enron employee position table and recipients are in the enron.com domain. We grouped the emails by the sender's position and calculated the rate of informality in each group. The results are in Table 6: the first two columns are the title and the rank of the positions in Enron; the third column is the number of employees with that position; the fourth column is the total number of emails sent by these employees; the fifth column is the rate of informality; the last column is the percentage of emails that contain requests according to our request classifier. It is interesting to see that the rates of informality and request vary a lot for different positions; for instance, lawyers are more formal and make more requests than traders.

Position	Rank	# of emp	Emails sent	Inf %	Req %
CEO	6	4	836	19.4%	21.7%
President	5	4	2,680	34.3%	19.3%
VP	4	28	11,425	22.2%	18.1%
Manag Dir	3	6	4,953	14.0%	14.7%
Director	2	22	1,879	29.4%	15.2%
Manager	1	13	6,563	12.4%	25.3%
In-house lawyer	0	3	1,548	7.0%	26.9%
Trader	0	12	1,743	33.1%	13.4%
Employee	0	38	11,770	19.1%	19.1%
Total	-	130	43,397	22.0%	19.2%

Table 6: The set of Enron employees used in the formality vs. rank study

To study the effect of rank difference on formality, we used the first six rows in Table 6 as the relative ranks of the next three rows are not so clearly defined (Diesner et al., 2005). In total, there are 77 employees with rank 1-6, and we call this set of people *RankSet*. We then extracted from the ISI dataset only those emails that have exactly one recipient and both sender and recipient are members of *RankSet*. We grouped this small set, 3999 emails in total, according to the rank difference (which is defined to be the rank of the recipient minus the rank of the sender). The results are in Table 7: the last column is the number of (sender, recipient) pairs with that rank difference. For instance, the -2 row indicates that, among those messages addressed two ranks lower in the organizational hierarchy, 24.7% are informal.

In general, Table 7 shows a lower rate of informality when an email is addressed to a recipient of superior rank. For example, the informality rate of an email addressed to someone 4 or more ranks higher than the sender (15.6%) is less than half that of an email addressed to someone 4 or more ranks lower (31.6%). We do not know what causes the increase of informality from +1 to +2; nevertheless, from +2 to +4 (in emails addressing someone 2-4 ranks higher), there is another decrease in informality rate.

Rank diff	Formal	Inf	Inf %	# of pairs
-4 or less	39	18	31.6%	16
-3	84	32	27.6%	32
-2	226	74	24.7%	56
-1	499	141	22.0%	82
0	989	275	21.8%	190
+1	784	175	18.2%	95
+2	270	121	30.9%	58
+3	125	38	23.3%	46
+4 or more	92	17	15.6%	29
Total	3108	891	22.3%	604

Table 7: Formality and rank difference, $p < 0.0001$. *Rank diff* is equal to recipient rank minus sender rank.

5.4 Weight of imposition: Requests

According to Brown and Levinson's model of politeness, the greater weight of an imposition, the greater the usage of polite speech acts including formality. In this model, a request is one of the most imposing speech acts. Therefore, when a request is made, we would expect a lower rate of informality.

To investigate this, we used the ISI dataset and the results of our request classifier to determine the

rate of informality for request and no-request emails. Table 8 shows that there is indeed a lower rate of informality when a request is being made.

	Formal	Informal	Inf %
Request	42,313	9,928	19.0%
No-request	128,958	51,616	28.6%
Total	171,271	61,544	26.4%

Table 8: Formality and request, $p < 0.0001$

# of recipients	Formal	Inf	Inf %
1	70,361	33,115	32.0%
2	5,807	1,914	24.8%
3-5	22,139	4,383	16.5%
6-10	12,903	2,626	16.9%
11 or greater	22,080	5,429	19.7%
Total	133,290	47,467	26.3%

Table 9: Formality and the number of recipients, $p < 0.0001$

5.5 Number of recipients

Another hypothesis we considered is the assumption that a sender is less likely to be informal when there are more recipients on an email since he does not want to broadcast a style which is more personal and could be perceived as unprofessional. To test this hypothesis, we started with the ISI dataset and looked at the subset of emails where an email has at least one recipient.[10] The emails were then grouped based on the number of recipients in the emails.

Table 9 shows the rate of informality with different numbers of recipients. For the most part in these results, a greater number of recipients results in a lower rate of informality. For instance, the rate of informality is nearly cut in half when there are 3 to 5 recipients as opposed to a single recipient. However, at the upper end of this scale, the rate of informality rises again slightly. One possible explanation is that when an email is addressed to a very large number of recipients, the strategies employed (e.g., the model of saving face) might differ from those employed in an email addressed to a small audience.

6 Discussion

In this study, we explored the relation between formality and five factors: (1) personal vs. business, (2) amount of contact, (3) rank difference, (4) request, and (5) number of recipients. The experiments show that the general patterns between the rate of informality and the five factors are consistent with Brown and Levinson's model and our intuition; for instance, an email tends to be more formal if it is about business matter, it is sent to someone with a higher rank, or it contains a request. But the experiments did produce some unexpected results; for instance, the rate of informality increased slightly when the number of recipients is more than 10.

There are several possible reasons for the unexpected results. One is due to the limitation of our dataset. For instance, the social interaction between two people could easily go beyond the 3.5 years covered by the Enron corpus, and people could choose other ways of communication besides email. Therefore, the Enron corpus alone may not be sufficient to capture the social distance between two people in the corpus. Another possible reason is that the errors made by our classifiers could contribute to some of the unexpected results.

The third possible reason, the one that is most interesting to us, is that there are indeed some interesting phenomena which can explain away the unexpectedness of the results. For instance, an email sent to a large number of strangers (e.g., an advertisement sent to a large mailing list) may choose to use an informal and entertaining style in order to catch the recipients' attention. Therefore, a theory that intends to account for people's email behavior may need to distinguish emails sent to a large number of strangers from those sent to a small group of friends. The benefit of a study like ours is that it allows researchers to test a linguistic or social theory on a large data set in a real setting. The study can either provide supporting evidence for a theory or reveal certain discrepancies between the prediction made by the theory and the statistics in the real data, which could lead to revision or refinement of the theory.

While this case study has concentrated on email

[10] Some emails in the ISI dataset do not contain any recipient information. We suspect that the recipient information has been somehow removed before the data was released to the public. With the at-least-one-recipient requirement, the number of non-empty emails in the ISI dataset is reduced from 232,815 to 180,757.

communication, it would be interesting to study formality behavior in other communication media such as Facebook and Twitter. By applying our methodology to other media, it would be possible to determine whether there are other social factors that influence formality on these media. For example, it would be useful to determine whether there is a difference in formality with respect to the number of 'friends' or 'followers' that a person has. Similarly, it would be interesting to examine correlations on the basis of whether a Facebook profile is configured as 'public' or 'private' since the potential viewing audience would be reduced in the case of 'private' profiles. Since Facebook also contains profiles which are associated with both individuals and businesses, it would be interesting to compare these as we did with personal and business emails. Finally, it remains to be seen whether requests could be examined in these media but other social factors (including whether posts related to personal matters, social causes, or event promotion) could be explored to examine formality behavior.

7 Conclusions and Future Work

We believe that NLP techniques can be used to test linguistic or social theories. As a case study, we choose Brown and Levinson's model of politeness (1987), which states that three factors are considered in a decision whether or when to use communication techniques such as formality. We test the theory on the Enron email corpus, and our experimental results are largely consistent with the theory and human intuition.

For future work, we plan to improve the performance of our formality and request classifier by adding additional features such as the ones that look at the layout and zoning of an email (e.g., greetings and signoffs). We also plan to apply our methodology to other genres of data (e.g., blogs, Facebook, Twitter) or to test other theories.

Another direction for future work is to explore what communication techniques such as formality can reveal about the *culture* of a particular social network. For instance, among all the positions listed in the ISI Enron employee position table, lawyers have the lowest rate of informality (7.0%), compared to other positions (e.g., 33.1% for traders). This implies that the workplace behavior of lawyers (at least with respect to emails) is very different from that of traders. It will be interesting to compare the behaviors of people from different occupations or from different social networks. Furthermore, if we could define the norm of behavior within a social group, we could then identify the outliers who might deserve special attention for various reasons.

Acknowledgment We would like to thank Todd Lingren, Chris Rogers and three anonymous reviewers for helpful comments and Katherine Coleman, Carmen Harris and David Horton for providing email annotation. Special thanks are extended to Drew Marrè for his insight in application of this data.

References

Nor Azni Abdullah. 2006. Constructing Business Email Messages: A Model of Writer's Choice. *ESP Malaysia*, 12:53–63.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Vitor R. Carvalho and William W. Cohen. 2006. Improving "email speech acts" analysis via n-gram selection. In *Proceedings of the Analyzing Conversations in Text and Speech (ACTS) Workshop at HLT-NAACL 2006*, pages 35–41, New York.

Vitor R. Carvalho, Wen Wu, and William W. Cohen. 2007. Discovering leadership roles in email workgroups. In *Proc. of the 4th Conference on Email and Anti-Spam (CEAS 2007)*.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to classify email into "speech acts". In *Proceedings of the EMNLP-2004*, Barcelona, Spain.

A. Culotta, R. Bekkerman, and A. McCallum. 2004. Extracting social networks and contact information from email and the web. In *Proc. of the Conference on Email and Anti-Spam (CEAS 2004)*.

Richard L. Daft and Robert H. Lengel. 1986. Organisational Information Requirements, Media Richness, and Structural Determinants. *Management Science*, 32:554–571.

Jana Diesner, Terill Frantz, and Kathleen Carley. 2005. Communication networks from the enron email corpus "it's always about the people. enron is no different".

Computational & Mathematical Organization Theory, 11(3):201–228.

Sanaz Jabbari, Ben Allison, David Guthrie, and Louise Guthrie. 2006. Towards the Orwellian nightmare: separation of business and personal emails. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 407–411.

Parambir S. Keila and David B. Skillcorn. 2005. Detecting unusual and deceptive communication in email. Technical report, Queens University, Ontario, Canada.

Brian Klimt and Yiming Yang. 2004. Enron corpus: A new data set for email classification research. Technical report, Carnegie Mellon University.

Anton Leuski. 2004. Email is a stage: Discovering people roles from email archives. In *Proc. of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 502–503.

Andrew McCallum, Xuerui Wang, and Andres Corrada-Emmanuel. 2007. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30:249–272.

Andrew McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Einat Minkov, Richard C. Wang, and WilliamW. Cohen. 2005. Extracting personal names from email: Applying named entity recognition to informal text. In *Proc. of EMNLP-2005*.

Wanda Orlikowski and JoAnne Yates. 1994. Genre repertoire: The structuring of communicative practices in organizations. *Administrative Science Quarterly*, 39(4):541–574.

John R. Searle. 1975. A taxonomy of illocutionary acts. In K. Gunderson, editor, *Language, Mind, and Knowledge*, pages 344–369. Minneapolis.

Jitesh Shetty and Jafar Adibi. 2004. The enron email dataset database schema and brief statistical report. Technical report, Information Sciences Institute at University of South California.

L. Zhou, J.K. Burgoon, J.F. Nunamaker Jr, and D. Twitchel. 2004. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation*, 13:81–106.