An Empirical Analysis of Formality in Online Communication

Ellie Pavlick
University of Pennsylvania*
epavlick@seas.upenn.edu

Joel Tetreault
Yahoo Labs
tetreaul@yahoo-inc.com

Abstract

This paper presents an empirical study of linguistic formality. We perform an analysis of humans' perceptions of formality in four different genres. These findings are used to develop a statistical model for predicting formality, which is evaluated under different feature settings and genres. We apply our model to an investigation of formality in online discussion forums, and present findings consistent with theories of formality and linguistic coordination.

1 Introduction

Language consists of much more than just content. Consider the following two sentences:

1. *Those recommendations were unsolicited and undesirable.*

2. *that's the stupidest suggestion EVER.*

Both sentences communicate the same idea, but the first is substantially more formal. Such stylistic differences often have a larger impact on how the hearer understands the sentence than the literal meaning does (Hovy, 1987).

Full natural language understanding requires comprehending this stylistic aspect of meaning. To enable real advancements in dialog systems, information extraction, and human-computer interaction, computers need to understand the entirety of what humans say, both the literal and the non-literal. In this paper, we focus on the particular stylistic dimension illustrated above: formality.

Formality has long been of interest to linguists and sociolinguists, who have observed that it subsumes a range of dimensions of style including serious-trivial, polite-casual, and level of shared knowledge (Irvine, 1979; Brown and Fraser, 1979). The formal-informal dimension has even been called the "most important dimension of variation between styles" (Heylighen and Dewaele, 1999). A speaker's level of formality can reveal information about their familiarity with a person, opinions of a topic, and goals for an interaction (Hovy, 1987; Endrass et al., 2011). As a result, the ability to recognize formality is an integral part of dialogue systems (Mairesse, 2008; Mairesse and Walker, 2011; Battaglino and Bickmore, 2015), sociolinguistic analyses (Danescu-Niculescu-Mizil et al., 2012; Justo et al., 2014; Krishnan and Eisenstein, 2015), human-computer interaction (Johnson et al., 2005; Khosmood and Walker, 2010), summarization (Sidhaye and Cheung, 2015), and automatic writing assessment (Felice and Deane, 2012). Formality can also indicate context-independent, universal statements (Heylighen and Dewaele, 1999), making formality detection relevant for tasks such as knowledge base population (Suh et al., 2006; Reiter and Frank, 2010) and textual entailment (Dagan et al., 2006).

This paper investigates formality in online written communication. The contributions are as follows: 1) We provide an analysis of humans' subjective perceptions of formality in four different genres. We highlight areas of high and low agreement and extract patterns that consis-

* Research performed while at Yahoo Labs.

Transactions of the Association for Computational Linguistics, vol. 4, pp. 61–74, 2016. Action Editor: Janyce Wiebe and Kristina Toutanova.
Submission batch: 10/2015; Revision batch: 12/2015; Published 3/2016.

tently differentiate formal from informal text. 2) We develop a state-of-the-art statistical model for predicting formality at the sentence level, evaluate the model's performance against human judgments, and compare differences in the effectiveness of features across genres. 3) We apply our model to analyze language use in online debate forums. Our results provide new evidence in support of theories of linguistic coordination, underlining the importance of formality for language generation systems. 4) We release our new dataset of 6,574 sentences annotated for formality level.

2 Related Work

There is no generally agreed upon definition as to what constitutes formal language. Some define formality in terms of situational factors, such as social distance and shared knowledge (Sigley, 1997; Hovy, 1987; Lahiri et al., 2011). Other recent work adopts a less abstract definition which is similar to the notion of "noisy text"– e.g. use of slang and poor grammar (Mosquera and Moreda, 2012a; Peterson et al., 2011). As a result, many rules have been explored for recognizing and generating informal language. Some of these rules are abstract, such as the level of implicature (Heylighen and Dewaele, 1999; Lahiri, 2015) or the degree of subjectivity (Mosquera and Moreda, 2012a), while others are much more concrete, such as the number of adjectives (Fang and Cao, 2009) or use of contractions (Abu Sheikha and Inkpen, 2011).

Much prior work on detecting formality has focused on the lexical level (Brooke et al., 2010; Brooke and Hirst, 2014; Pavlick and Nenkova, 2015). For larger units of text, perhaps the best-known method for measuring formality is the $\mathcal{F}$-score[1] (Heylighen and Dewaele, 1999), which is based on relative part-of-speech frequencies. $\mathcal{F}$-score and its more recent variants (Li et al., 2013) provide a coarse measure of formality, but are designed to work at the genre-level, making them less reliable for shorter units of text such as sentences (Lahiri, 2015). Existing statistical approaches to detecting formality (Abu Sheikha and Inkpen, 2010; Peterson et al., 2011; Mosquera and Moreda, 2012b) have treated the problem as a binary classification task and relied heavily on word lists to differentiate the two classes. Linguistics literature supports treating formality as a continuum (Irvine, 1979; Heylighen and Dewaele, 1999), as has been done in studies of other pragmatic dimensions such as politeness (Danescu-Niculescu-Mizil et al., 2013) and emotiveness (Walker et al., 2012). Lahiri et al. (2011) provided a preliminary investigation of annotating formality on an ordinal scale and released a dataset of sentence-level formality annotations (Lahiri, 2015), but did not use their data in any computational tasks. This paper extends prior work by (i) introducing a statistical regression model of formality which is based on an empirical analysis of human perceptions rather than on heuristics and (ii) by applying that model to a linguistic analysis of online discussions.

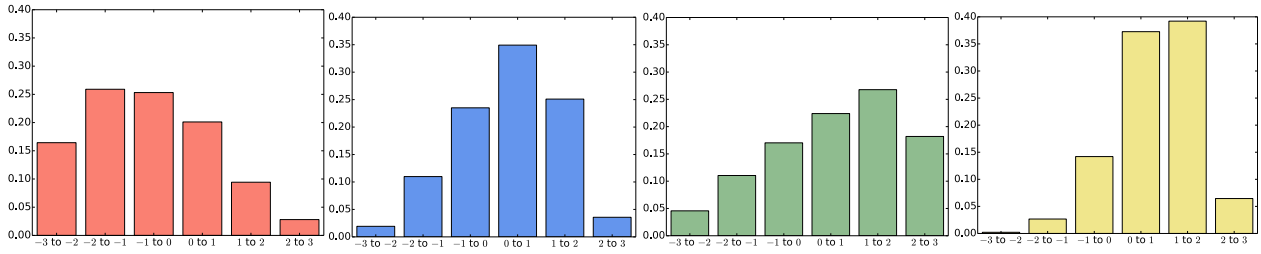
3 Human perceptions of formality

Before we can automatically recognize formality, we need an understanding of what it means for language to be formal or informal. As we discussed in Section 2, a number of theories exist with no clear consensus. In this work, we do not attempt to develop a concrete definition of formality, but instead take a bottom-up approach in which we assume that each individual has their own definition of formality. This approach of using unguided human judgments has been suggested by Sigley (1997) as one of the most reliable ways to get a gold-standard measure of formality, and has been applied in prior computational linguistics studies of pragmatics (Danescu-Niculescu-Mizil et al., 2013; Lahiri, 2015). We aim to answer: do humans' individual intuitions collectively provide a coherent notion of formality (§3.2)? And, if so, which linguistic factors contribute to this notion (§3.3)?

3.1 Data and Annotation

Since formality varies substantially across genres (Li et al., 2013), we look at text from four different genres: News, Blogs, Emails, and com-

[1]We use special font to denote Heylighen and Dewaele's $\mathcal{F}$-score to avoid confusion with F1 measure.



(a) Answers (μ=-0.7,σ=1.3) (b) Blogs (μ=0.2,σ=1.1) (c) Emails (μ=0.5,σ=1.4) (d) News (μ=0.7,σ=0.86)

Figure 1: Distribution of sentence-level formality scores by genre.

Answers	2.8	That is in addition to any customs duties that may be assessed.
Answers	-3.0	(LOL) jus kidding...the answer to your question is GAS PRICES!!!
News	2.6	Baghdad is a city of surprising topiary sculptures: leafy ficus trees are carved in geometric spirals, balls, arches and squares, as if to impose order on a chaotic sprawl.
News	-2.2	He bought and bought and never stopped.

Table 1: Examples of formal (positive) and informal (negative) sentences in different genres. Scores are taken as the mean of 5 human judgments on a scale from -3 to 3.

munity question answering forums (henceforth "Answers"). Lahiri (2015) released a corpus of sentence-level formality annotations, which contains 2,775 news and 1,821 blog sentences. In addition we take a random sample of 1,701 sentences from professional emails[2] and 4,977 sentences from Yahoo Answers.[3] We follow the protocol used in Lahiri (2015) in order to gather judgments on Amazon Mechanical Turk for the Email and Answers data. Specifically, we use a 7-point Likert scale, with labels from -3 (Very Informal) to 3 (Very Formal). So as not to bias the annotators with our own notions of formality, we provide only a brief description of formal language and encourage annotators to follow their instincts when making their judgments. We use the mean of 5 annotators' scores as the overall formality score for each sentence.[4] Our newly collected annotations have been made public.[5] For more information on the annotation, please refer to the supplementary material to this paper.[6]

3.2 Analysis

Figure 1 shows the distribution of mean formality scores for the sentences in each of our genres. We see that News is the most formal of our domains and Answers is the least. However, we can see anecdotally (Table 1) that the standard of what constitutes "informal" depends heavily on the genre: an informal sentence from News is much more formal than one from Answers. We can also see clear differences in the variance of sentence formalities within each genre. In general, the interactive genres (Email and Answers) show a much flatter distribution than do the informational genres (News and Blogs).

Inter-annotator agreement. We want to know whether individuals' intuitions about formal language result in a coherent collective notion of formality. To quantify this, we measure whether annotators' ordinal ratings of formality are well correlated and whether their categorical judgments are in agreement. For the former, we use intraclass correlation[7] (ICC) which

[2] `http://americanbridgepac.org/jeb-bushs-gubernatorial-email-archive/`

[3] `https://answers.yahoo.com/`

[4] In total, we had 301 annotators, meaning each annotator labeled 22 sentences on average.

[5] `http://www.seas.upenn.edu/~nlp/resources/formality-corpus.tgz`

[6] `http://www.seas.upenn.edu/~epavlick/papers/formality_supplement.pdf`

[7] We report the average raters absolute agreement (ICC1k) using the psych package in R: `https://cran.`

3,3,3,3,3	Formal	I would trust the social workers to make the appropriate case by case determination .
-3,-3,-3,-3,-3	Informal	* what the world needs is only more of U & UR smile ! !
-3,-2,0,-1,3	Mixed	Governor, if this was intentionally done, whoever did it has at least one vote to go to hell.
-1,0,0,0,1	Neutral	You should try herbal peppermint tea.

Table 2: Examples of sentences with different patterns of agreement. Numbers show the list of scores assigned by the 5 annotators. Some sentences exhibit "mixed" formality, i.e. workers were split on whether to call the sentence generally informal or generally formal, while others are "neutral," i.e. workers agreed the sentence was neither formal nor informal.

is similar to Pearson ρ but accounts for the fact that we have different groups of annotators for each sentence. For the latter, we use a quadratic weighted κ, which is a variation of Cohen's κ better fit for measuring agreement on ordinal scales.[8] When using crowdsourced labels, computing reliable measures of κ is difficult since, for a given pair of annotators, the number of items for which both provided a label is likely small. We therefore simulate two annotators as follows. For each sentence, we randomly choose one annotator's label to be the label of Annotator 1 and we take the mean label of the other 4 annotators, rounded to the nearest integer, to be the label of Annotator 2. We then compute κ for these two simulated annotators. We repeat this process 1,000 times, and report the median and 95% confidence interval (Table 3).

	N	ICC	Weighted κ
Answers	4,977	0.79 $\pm$0.01	0.54$\pm$0.05
Blog	1,821	0.58 $\pm$0.03	0.31$\pm$0.05
Email	1,701	0.83 $\pm$0.02	0.59$\pm$0.04
News	2,775	0.39$\pm$0.05	0.17$\pm$0.06

Table 3: Annotator agreement measured by intraclass correlation (ICC) and categorical agreement (quadratic weighted κ) for each genre.

Agreement is reasonably strong across genres, with the exception of News, which appears to be the most difficult to judge. Table 2 sheds light on the types of sentences that receive high and low levels of agreement. At the extreme ends of the spectrum where agreement is very high (mean scores near -3 and $+3$), we see sentences which are unambiguously formal or informal. However, in the middle (mean scores near 0) we see both high and low agreement sentences. High agreement sentences tend to be "neutral," i.e. annotators agree they are neither formal nor informal, while the low-agreement sentences tend to exhibit "mixed" formality, i.e. they contain both formal and informal sub-sentential elements. We leave the topic of sub-sentential formality for future work, and instead allow our use of the mean score to conflate mixed formality with neutral formality. This fits naturally into our treatment of formality as a continuous as opposed to a binary attribute.

3.3 Factors affecting formality

From the above analysis, we conclude that humans have a reasonably coherent concept of formality. However, it is difficult to tease apart perceived formality differences that arise from the literal meaning of the text (e.g. whether the topic is serious or trivial) as opposed to arising from the style in which those ideas are expressed. To get a better understanding of the stylistic choices that differentiate formal from informal, we ran a second experiment in which we asked annotators to rewrite informal sentences in order to make them more formal. The goal is to isolate some of the linguistic factors that contribute to perceived formality while constraining the literal content of the text to be the same. We use this data for analysis in this section, as well as for evaluation in Section 4.2.

For this task, we chose 1,000 sentences from the Answers dataset, since it displays the widest variety of topics and styles. We attempt to

r-project.org/web/packages/psych/psych.pdf

[8] Weighted κ penalizes large disagreements more than small disagreements. E.g. if Annotator 1 labels a sentence as -2 and Annotator 2 labels it -3, this is penalized less than if Annotator 1 chooses -2 and Annotator 2 chooses $+3$.

Capitalization	50%	**i** do not like **w**almart.	**I** do not like **W**almart.
Punctuation	39%	She's 40, but she seems more like a 30**!!!!!**	She is 40, but she seems more like 30**!**
Paraphrase	33%	Lexus cars are **awesome**!	Lexus brand cars are **very nice**.
Delete fillers	19%	**well** it depends on that girl.	It depends on the girl.
Completion	17%	looks good on your record.	**It** looks good on your record.
Add context	16%	alive - i have seen **that guy** working at a 7-11 behnd the counter	My opinion is that **Osama Bin Laden** is alive as I have encountered him working at a 7-11 store .
Contractions	16%	I really **don't** like them.	I really **do not** like them.
Spelling	10%	i love dancing **iwth** my chick friends.	I enjoy dancing **with** my girlfriends.
Normalization	8%	**juz** try to put **ur** heart in to it.	**Just** try to put **your** heart into it.
Slang/idioms	8%	that's a big no.	I do not agree.
Politeness	7%	uh, more details?	**Could you provide** more details, **please**?
Split sentences	4%	[. . .] not as tough**...** like high school	[. . .] not as tough**. It's** like high school.
Relativizers	3%	sorry i ' m not much help heh	Sorry **that** I am not much help.

Table 4: Frequency of types of edits/changes made in rewriting experiment, and examples of each. Note the categories are not mutually exclusive.

choose sentences that are informal enough to permit formalizing, while covering all ranges of informality, from highly informal ("yep...love the pic lol") to only slightly informal ("As long as you feel good."). Each sentence is shown in the context of the question and the full answer post in which it appeared. We collect one rewrite per sentence, and manually remove spammers.

People make a large variety of edits, which cover the "noisy text" sense of formality (e.g. punctuation fixes, lexical normalization) as well as the more situational sense (e.g. inserting politeness, providing context). To characterize these different edit types, we manually reviewed a sample of 100 rewrites and categorized the types of changes that were made. Table 4 gives the results of this analysis. Over half of the rewrites involved changes to capitalization and punctuation. A quarter involved some sort of lexical or phrasal paraphrase (e.g "awesome" → "very nice"). In 16% of cases, the rewritten sentence incorporated additional information that was apparent from the larger context, but not present in the original sentence. This accords with Heylighen and Dewaele (1999)'s definition of "deep" formality, which says that formal language strives to be less context-dependent.

4 Recognizing formality automatically

In the above section, we asked whether humans can recognize formality and what contributes to their perception of formal or informal. We now ask: how well can computers automatically distinguish formal from informal and which linguistic triggers are important for doing so?

4.1 Setup

We use the data described in Section 3.1 for training, using the mean of the annotators' scores as the gold standard labels. We train a ridge regression[9] model with the model parameters tuned using cross validation on the training data. Unless otherwise specified, we keep genres separate, so that models are trained only on data from the genre in which they are tested.

Features. We explore 11 different feature groups, described in Table 5. To the best of our knowledge, 5 of these feature groups (ngrams, word embeddings, parse tree productions, dependency tuples, and named entities) have not been explored in prior work on formality recognition. The remaining features (e.g. length, POS tags, case, punctuation, formal/informal lexicons, and subjectivity/emotiveness) largely subsume the features explored by previously published classifiers. We use Stanford CoreNLP[10] for all of our linguistic processing, except for subjectivity features, for which we use TextBlob.[11]

[9] `http://scikit-learn.org/`

[10] `http://nlp.stanford.edu/software/corenlp`

[11] `https://textblob.readthedocs.org`

case	Number of entirely-capitalized words; binary indicator for whether sentence is lowercase; binary indicator for whether the first word is capitalized.
*dependency	One-hot features for the following dependency tuples, with lexical items backed off to POS tag: (gov, typ, dep), (gov, typ), (typ, dep), (gov, dep).
*entity	One-hot features for entity types (e.g. PERSON, LOCATION) occurring in the sentence; average length, in characters, of PERSON mentions.
lexical	Number of contractions in the sentence, normalized by length; average word length; average word log-frequency according to Google Ngram corpus; average formality score as computed by Pavlick and Nenkova (2015).
*ngram	One-hot features for the unigrams, bigrams, and trigrams appearing in the sentence.
*parse	Depth of constituency parse tree, normalized by sentence length; number of times each production rule appears in the sentence, normalized by sentence length, and not including productions with terminal symbols (i.e. lexical items).
POS	Number of occurrences of each POS tag, normalized by the sentence length.
punctuation	Number of '?', '...', and '!' in the sentence.
readability	Length of the sentence, in words and characters; Flesch-Kincaid Grade Level score.
subjectivity	Number of passive constructions; number of hedge words, according to a word list; number of 1st person pronouns; number of 3rd person pronouns; subjectivity according to the TextBlob sentiment module; binary indicator for whether the sentiment is positive or negative, according to the TextBlob sentiment module. All of the counts are normalized by the sentence length.
*word2vec	Average of word vectors using pre-trained word2vec embeddings, skipping OOV words.

Table 5: Summary of feature groups used in our model. To the best of our knowledge, those marked with (*) have not been previously studied in the context of detecting linguistic formality.

Baselines. We measure the performance of our model using Spearman ρ with human labels. We compare against the following baselines:

- **Sentence length**: We measure length in characters, as this performed slightly better than length in words.
- **Flesch-Kincaid grade level**: FK grade level (Kincaid et al., 1975) is a function of word count and syllable count, designed to measure readability. We expect higher grade levels to correspond to more formal text.
- **$\mathcal{F}$-score**: Heylighen and Dewaele (1999)'s formality score ($\mathcal{F}$-score) is a function of POS tag frequency which is designed to measure formality at the document- and genre-level. We expect higher $\mathcal{F}$-score to correspond to more formal text.
- **LM perplexity**: We report the perplexity according to a 3-gram language model trained on the English Gigaword with a vocabulary of 64K words. We hypothesize that sentences with lower perplexity (i.e. sentences which look more similar to edited news text) will tend to be more formal. We also explored using the ratio of the perplexity according to an "informal" language model over the perplexity according to a "formal" language model as a baseline, but the results of this baseline were not competitive, and so, for brevity, we do not include them here.
- **Formality lexicons**: We compare against the average word formality score according to the formality lexicon released by Brooke and Hirst (2014). We compute this score in the same way as Sidhaye and Cheung (2015), who used it to measure the formality of tweets.
- **Ngram classifier**: As our final baseline, we train a ridge regression model which uses only ngrams (unigrams, bigrams, and trigrams) as features.

Comparison against previously published models. Note that we are not able to make a meaningful comparison against against any of the previously published statistical models for formality detection. To our knowledge, there are three relevant previous publications that produced statistical models for detecting formality: Abu Sheikha and Inkpen (2010), Peterson et al. (2011), and Mosquera and Moreda (2012b). All three of these models performed a binary classification (as opposed to regression) and oper-

ated at the document (as opposed to sentence level). We were able to closely reimplement the model of Peterson et al. (2011), but we choose not to include the results here since their model was designed for binary email-level classification and thus relies on domain-specific features (e.g. casing in the subject line), that are not available in our real-valued, sentence-level datasets. The other models and the data/lexicons on which they relied are not readily available. For this reason, we do not compare directly against the previously published statistical models, but acknowledge that several of our features overlap with prior work (see Section 4.1 and Table 5).

4.2 Performance

Table 6 reports our results on 10-fold cross validation. Using our full suite of features, we are able to achieve significant performance gains in all genres, improving by as much as 11 points over our strongest baseline (the ngram model).

	Answers	Blogs	Email	News
LM ppl	0.00	-0.01	0.14	-0.08
$\mathcal{F}$-score	0.16	0.35	0.21	0.27
Length	0.23	0.51	0.53	0.34
F-K level	0.45	0.54	0.63	0.41
B&H lexicon	0.47	0.41	0.55	0.30
Ngram model	0.60	0.55	0.65	0.43
Classifier	**0.70**	**0.66**	**0.75**	**0.48**

Table 6: Spearman ρ with human judgments for our model and several baselines.

Note that, while the basic LM perplexity correlates very weakly with formality overall, the Email genre actually exhibits a trend opposite of that which we expected: in Email, sentences which look *less* like Gigaword text (higher perplexity) tend to be *more* formal. On inspection, we see that many of the sentences which have low perplexity but which humans label as informal include sentences containing names and greeting/signature lines, as well as sentences which are entirely capitalized (capitalization is not considered by the LM).

Contributions of feature groups. In order to gain better insight into how formality differs across genres, we look more closely at the performance of each feature group in isolation. Table 7 shows the performance of each feature group relative to the performance of the full classifier, for each genre. A few interesting results stand out. Ngram and word embedding features perform well across the board, achieving over 80% of the performance of the full classifier in all cases. Casing and punctuation features are significantly more important in the Answers domain than in the other domains. Constituency parse features and entity features carry notably more signal in the Blog and News domains than in the Email and Answers domains.

	Answers	Blogs	Email	News
ngram	0.84	0.85	0.84	0.91
word2vec	0.83	0.83	0.84	0.87
parse	0.70	0.89	0.74	0.89
readability	0.69	0.75	0.84	0.83
dependency	0.64	0.89	0.84	0.85
lexical	0.56	0.55	0.59	0.70
case	0.50	0.28	0.24	0.37
POS	0.49	0.74	0.67	0.74
punctuation	0.47	0.38	0.37	0.20
subjectivity	0.29	0.31	0.25	0.37
entity	0.14	0.63	0.34	0.72

Table 7: Relative performance of each feature group across genres. Numbers reflect the performance (Spearman ρ) of the classifier when using only the specified feature group, relative to the performance when using all feature groups.

train\test	Answers	Blogs	Email	News
Answers	0	-5	-5	-6
Blogs	-17	0	-9	-2
Email	-13	-4	0	-4
News	-23	-4	-13	0

Table 8: Drop in performance (Spearman $\rho \times 100$) when model is trained on sentences from one domain (row) and tested on sentences from another (column). Changes are relative to the performance when trained only on sentences from the test domain (represented by zeros along the diagonal). All models were trained on an equal amount of data.

Observing these differences between data sets raises the question: how well does knowledge of formality transfer across domains? To answer this, we measure classifier performance when trained in one domain[12] and tested in another (Table 8). In our experiments, the model trained

[12] All models were trained on an equal amount of data.

on Answers consistently provided the best performance out of domain, resulting in performance degradations of roughly 5 points (Spearman ρ) compared to models trained on target domain data. Training on News and testing on Answers caused the largest drop (23 points compared to training on Answers).

Pairwise classification. As a final evaluation, we use the 1,000 rewritten sentences from Section 3.3 as a held-out test set. This allows us to test that our classifier is learning real style differences, not just topic differences. We assume that workers' rewrites indeed resulted in more formal sentences, and we frame the task as a pairwise classification in which the goal is to determine which of the two sentences (the original or the rewrite) is more formal. A random baseline achieves 50% accuracy. If we use the F-K readability measure, and assume the sentence with the higher grade level is the more formal of the two, we achieve only 57% accuracy. By running our supervised regression model and choosing the sentence with the higher predicted formality score as the more formal sentence, we achieve 88% accuracy, providing evidence that the model picks up subtle stylistic, not just topic, differences.

5 Formality in online discussions

So far we have focused on building a model that can automatically distinguish between formal and informal sentences. We now use that model to analyze formality in practice, in the context of online discussion forums. We look to existing theories of formality and of linguistic style matching to guide our analysis. In particular:

- Formality is higher when the amount of shared context between speakers is low (Heylighen and Dewaele, 1999).
- Formality is higher when speakers dislike one another (Fielding and Fraser, 1978).
- Speakers adapt their language in order to match the linguistic style of those with whom they are interacting (Danescu-Niculescu-Mizil et al., 2011).

Ladywolf I was checking out this website for Exodus International and I understand their mission is to provide an alternative for people who choose to be heterosexual. [...] I just find it hard to believe that they don't somehow manipulate the situation in a less than fair way.

joebrummer I started a thread earlier about just this! These groups are dangerous Ladywolf, There is so much evidence to support that [...]

Ladywolf I thought so [...] I also see that they are running major newspaper ads...hmmm, how unbiased can a newspaper ad like this be? [...] I'm so glad I wasn't raised a Christian because from the tone of some of the replies, some members of this cult can be pretty mean huh?

joebrummer Yes, The are mean funny enough in the name of god. I was raised christian, catholic no less. I studied the bible, I was raised believing I would go to hell. That was tough.

Ladywolf I bet that was tough [...] I was raised Jewish [...] It's like so wierd because I've never had to deal with these types of people before.

Figure 2: Example of a thread from our data. [...] indicates text has been left out to save space.

With these hypotheses in mind, we explore how formality changes across topics and users (§5.2), how it relates to other pragmatic dimensions (§5.3), and how it changes over the lifetime of a thread (§5.4). Understanding these patterns is an important first step toward building systems that can interact with people in a pragmatically competent way.

5.1 Discussion Data

Our data comes from the Internet Argument Corpus (IAC) dataset (Walker et al., 2012), a corpus of threaded discussions from online debate forums. The dataset consists of 388K posts covering 64 different topics, from Economics to Entertainment. We focus on threads in our analysis, defined as chains of posts in which each is an explicit reply to the previous post (Figure 2). When the same user makes multiple consecutive posts in a thread (i.e. replies to their own post), we collapse these and treat them as a single post. In total, our data covers 104,625 threads.

Automatic Classification. First, we assign a formality score to each post in our data using the Answers model in Section 4. Since this model is designed for sentence-level prediction, we define the score of a post to be the mean score of the sentences in that post. We acknowledge that this approximation is not ideal; to confirm that it will be sufficient for our analyses, we collect human judgments for 1,000 random posts using the same task setup as we used for the sentence-level judgments in Section 3.1. The

correlation of our predicted score with the mean human score is 0.58, which is within the range of inter-annotator agreement for labeling post formality ($0.34 \leq \rho \leq 0.64$).[13] We take this as confirmation that the mean predicted sentence score is a decent approximation of human formality judgments for our purposes.

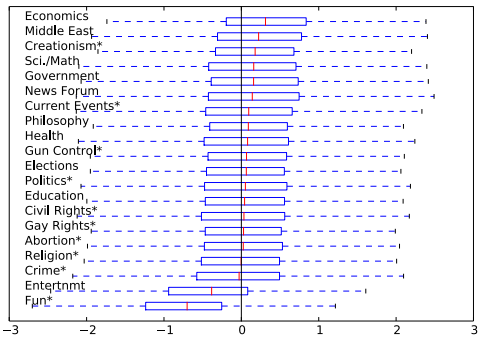


Figure 3: Formality distribution of posts in 20 most popular topics in discussion data. The 10 most popular topics (*) are used in our other analyses.

5.2 How do topic and user affect formality?

As formality is intertwined with many content-specific style dimensions such as "serious-trivial" (Irvine, 1979), we expect the overall formality level to differ across topics. Figure 3 confirms this– many topics are clearly skewed toward being formal[14] (e.g. Economics) while others are skewed toward informal (e.g. Fun). However, every topic includes both formal and informal posts: there are informal posts in Economics ("Oh my! A poor person....how could this have happened!") and formal posts in Fun ("Difficult to consider either one, or their variations, as a viable beverage when beer is available.").

We see a similar pattern when we look at post formality levels by user: while most people speak generally formally or generally informally (84% of users have a mean formality level that is significantly different from 0 at $p < 0.01$), nearly every user (91%) produces both formal and informal posts.[15] This is true even when we look at users within one topic. These results are interesting: they suggest that while the formality of a post is related to the topic of discussion and to the individual speaker, these alone do not explain formality entirely. Rather, as the aforementioned theories suggest, the same person discussing the same topic may become more or less formal in response to pragmatic factors.

5.3 How does formality relate to other pragmatic styles?

Formality is often considered to be highly related with, and even to subsume, several other stylistic dimensions including politeness, impartiality, and intimacy. Heylighen and Dewaele (1999) suggest that formality is higher when shared social context is lower, and thus language should be more formal when directed at larger audiences or speaking about abstract concepts. Fielding and Fraser (1978) further suggest that *in*formality is an important way of expressing closeness with someone, and thus formality should be higher when speakers dislike one another.

To investigate these ideas further, we look at how formality correlates with human judgments of several other pragmatic dimensions. We use the manual style annotations that are released for a subset of post-reply pairs (3K total) in the IAC dataset (Walker et al., 2012). These annotations include, for example, the extent to which the reply agrees/disagrees with the post and the extent to which the reply is insulting/respectful of the post. Each of these dimensions has been rated by human annotators on a Likert scale, similar to our own formality annotations. Additionally, to investigate how formality correlates

[13] This range matches the agreement range observed for post-level politeness annotations (Danescu-Niculescu-Mizil et al., 2013). Note agreement is more varied at the post level than at the sentence level. This makes sense given the "mixed formality" phenomenon: i.e. for long posts, a range of formality can be expressed, making the choice of a single score more subjective.

[14] The range of post formalities is generally narrower than was the range of sentence formalities. While sentence-level scores range between -3 and 3, we find that 80% of post scores fall between -1 and 1.

[15] We consider posts with scores > 0.25 as "formal" and those with scores < -0.25 as "informal."

Emotional	The main cause of so much hate and disrespect is the phony war we're fighting and our tactics in violation of international law, our attitude of superiority in the world, and our bullying of others.
Impolite	As a former administrator, and therefore a veteran editor who knows how wikipedia really works, I am actually surprised you would even ask such a question with such an obvious answer.
Insulting	And here ladies and gentlemen we have the evidence of why I am justified in calling the likes of 'stormboy' an idiot.
Sarcastic	Thank you for bringing to my attention that atoms, neutrons and protons are merely scientific assumptions. Now as I gaze at the night sky with all its bits and pieces spinning around each other I can sleep happily knowing that our solar system is not part of a housebrick afterall.

Table 9: Formal posts exhibiting style properties often thought not to co-occur with formality.

with politeness, we use the the Stanford Politeness Corpus (Danescu-Niculescu-Mizil et al., 2013), which consists of 11K short posts from Wikipedia discussion forums which again have been manually annotated on an ordinal scale.

Our results are generally consistent with what theories suggest. We find that posts which are targeted toward more general audiences (as opposed to specific people) and which make fact-based (as opposed to emotion-based) arguments are generally more formal ($\rho = 0.32$ and 0.17, respectively), and that formality is significantly positively correlated with politeness ($\rho = 0.14$). We find significant negative correlations between formality and the extent to which the post is seen as sarcastic ($\rho = -0.25$) or insulting ($\rho = -0.22$). Interestingly, we do not find a significant correlation between formality and the degree of expressed agreement/disagreement.

While the directions of these relationships match prior theories and our intuitions, the strength of the correlation in many cases is weaker than we expected to see. Table 9 provides examples of some of the less intuitive co-occurences of style, e.g. impolite but formal posts. These examples illustrate the complexity of the notion of formality, and how formal language can be used to give the impression of social distance while still allowing the speaker's emotions and personality to be very apparent.

5.4 How does formality change throughout a discussion?

Prior work has revealed that speakers often adapt their language to match the language of those with whom they are interacting (Danescu-Niculescu-Mizil et al., 2011). We therefore investigate how formality changes over the lifetime of a thread. Do discussions become more or less formal over time? Do speakers' levels of formality interact with one another?

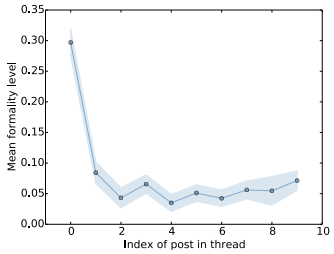
For these analyses, we focus on threads from 5 to 20 posts in length. Because threads can branch, multiple threads might share a prefix sequence of posts. To avoid double counting, we group together threads which stem from the same post and randomly chose one thread from each such group, throwing away the rest.

Following the theory that formality is determined by the level of shared context, Heylighen and Dewaele (1999) hypothesize that formality should be highest at the beginning of a conversation, when no context has been established. We observe that, in fact, the first posts have significantly higher formality levels on average than do the remaining posts in the thread (Figure 4).

Once a context is established and a discussion begins, the theory of linguistic style matching suggests that people change their language to match that of others in the conversation (Niederhoffer and Pennebaker, 2002; Danescu-Niculescu-Mizil et al., 2011). Is this phenomenon true of formality? Does a person's level of formality reflect the formality of those with whom they are speaking?

Figure 2 shows an example thread in which the speakers together move toward more informal tone as the conversation becomes more personal. To see if this kind of formality matching is the case in general, we use a linear mixed effects model.[16] Briefly, a mixed effects model is

[16]We use the mixed effects model with random intercepts provided by the statsmodels python package: `http://statsmodels.sourceforge.net/devel/mixed_`



Initial I wish to have a formal debate in the Debate Tournaments section on global warming. I propose the subject title of "Global Warming is both occuring and has been shown to be at least in part caused by human activity" I will take the afirmative position. Anyone want to argue the opposite?

Reply Global warming is a controversy. Personally I am like hundred of maybe thousands if not millions of people that think it is liberal ###. The hole in the ozone layer is false, and I am sure this is too.

Initial The US military says that Saddam Hussein's briefcase contained transcripts of meetings with terrorists, contact information for those terrorists, and information on financial transactions that he carried out. [...] I wonder what else was in the briefcase. [...]

Reply Transcripts? Strange. I would be curious too.

Figure 4: On average, first posts are significantly more formal than later posts. Left: mean formality of posts by position in thread. Right: some examples where formal initial posts are followed by less formal replies. (Note: 4forums.com replaces expletives with #s.)

a regression analysis that allows us to measure the influence of various "fixed effects" (e.g. the formality of the prior post) on a post's formality, while controlling for the "random effects" which prevent us from treating every post as an independent observation. In our case, we treat the topic and the author as random effects, i.e. we acknowledge that the formality levels of posts in the same topic by the same author are not independent, and we want to control for this when measuring the effects of other variables.

We include 7 fixed effects in our model of a post's formality: the formality of the previous post, the number of prior posts in the thread (position), the number of prior posts *by this author* in the thread (veteran level), the length of the entire thread, the total number of participants in the entire thread, and the lengths of the current and prior posts. We also include the pairwise interactions between these fixed effects. We include the topic and author as a random effect. For these analyses, we omit the first post in every thread, as prior analysis suggests that the function of the first post, and its formality, is markedly different from that of later posts.

Table 10 gives the most significant results from our regression. We observe several interesting significant effects, such as a negative relationship between the number of times an author has posted in the thread and their formality level: i.e. people are more informal the more they post. However, the single best predictor of the formality of a post is the formality of the post to which it is replying. The estimated effect size is 0.22, meaning, all else being equal, we expect an increase of 1 in the prior post's formality to correspond to an increase of 0.22 in the formality of the current post. This suggests that a person's formality does depend on the formality of others in the conversation.

	Coefficient
Previous score	0.219
Veteran level	−0.078
Thread length	0.020
Number of participants	−0.010
Previous score × position	0.009
Position	0.008

Table 10: Estimated coefficients of variables strongly related to the formality of a post, controlling for topic- and author-specific random effects. All effects are significant at $p < 0.0001$. × signifies an interaction between variables.

Perhaps more interestingly, we see a significant positive effect of the interaction between previous score and position. That is, the effect of prior post formality on current post formality becomes stronger later in a thread compared to at the beginning of a thread. Figure 5 shows how the estimated coefficient for prior post formality on current post formality changes when we look only at posts at a particular index in a thread (e.g. only second posts, only tenth posts). We can see that the coefficient is more than twice as large for the tenth post of a thread than it is for the second post in that thread. One could imagine several explanations for this: i.e. users with similar formality levels may engage in longer discussions, or users who engage in longer discus-

`linear.html`.

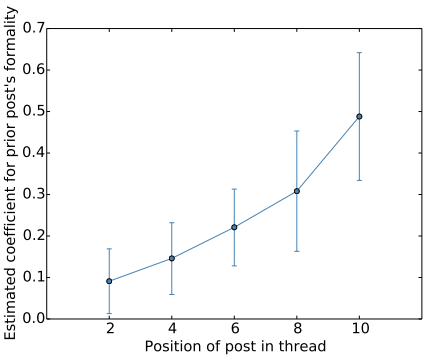


Figure 5: Effect size of prior posts's formality on current post's formality for posts at different positions in a thread. Effect size can be interpreted as the expected increase in a post's formality corresponding to an increase of 1 in the prior post's formality, all else being equal.

sions may tend to adapt better to one another as the discussion progresses. We leave further investigation for future work.

6 Conclusion

Language contains more than its literal content: stylistic variation accounts for a large part of the meaning that is communicated. Formality is one of the most basic dimensions of stylistic variation in language, and the ability to recognize and respond to differences in formality is a necessary part of full language understanding. This paper has provided an analysis of formality in written communication. We presented a study of human perceptions of formality across multiple genres, and used our findings to build a statistical model which approximates human perceptions of formality with high accuracy. This model enabled us to investigate trends in formality in online debate forums, revealing new evidence in support of existing theories about formality and about linguistic coordination. These findings provide important steps toward building pragmatically competent automated systems.

Acknowledgements. We would like to thank Martin Chodorow for valuable discussion and input, and Marilyn Walker, Shereen Oraby, and Shibamouli Lahiri for sharing and facilitating the use of their resources. We would also like to thank Amanda Stent, Dragomir Radev, Chris Callison-Burch, and the anonymous reviewers for their thoughtful suggestions.

References

Fadi Abu Sheikha and Diana Inkpen. 2010. Automatic classification of documents by formality. In *Interntional Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, pages 1–5. IEEE.

Fadi Abu Sheikha and Diana Inkpen. 2011. Generation of formal and informal sentences. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 187–193, Nancy, France, September. Association for Computational Linguistics.

Cristina Battaglino and Timothy Bickmore. 2015. Increasing the engagement of conversational agents through co-constructed storytelling. *Eighth Workshop on Intelligent Narrative Technologies*.

Julian Brooke and Graeme Hirst. 2014. Supervised ranking of co-occurrence profiles for acquisition of continuous lexical attributes. In *Proceedings of The 25th International Conference on Computational Linguistics*.

Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Coling 2010: Posters*, pages 90–98, Beijing, China, August. Coling 2010 Organizing Committee.

Penelope Brown and Colin Fraser. 1979. Speech as a marker of situation. In *Social Markers in Speech*, pages 33–62. Cambridge University Press.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, pages 177–190. Springer.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words!: Linguistic style accommodation in social media. In *Proceedings of the 20th International Conference on World Wide Web*, pages 745–754. ACM.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st International Conference on World Wide Web*, pages 699–708. ACM.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher

Potts. 2013. A computational approach to politeness with application to social factors. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, August.

Birgit Endrass, Matthias Rehm, and Elisabeth André. 2011. Planning small talk behavior with cultural influences for multiagent systems. *Computer Speech & Language*, 25(2):158–174.

Alex Chengyu Fang and Jing Cao. 2009. Adjective density as a text formality characteristic for automatic text classification: A study based on the british national corpus. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, pages 130–139, Hong Kong, December. City University of Hong Kong.

Rachele De Felice and Paul Deane. 2012. Identifying speech acts in e-mails: Toward automated scoring of the TOEIC® e-mail task. *ETS Research Report Series*, 2012(2):i–62.

Guy Fielding and Colin Fraser. 1978. Language and interpersonal relations. *The social context of language*, pages 217–232.

Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: Definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel*.

Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.

Judith T. Irvine. 1979. Formality and informality in communicative events. *American Anthropologist*, 81(4):773–790.

W. Lewis Johnson, Richard E. Mayer, Elisabeth André, and Matthias Rehm. 2005. Cross-cultural evaluation of politeness in tactics for pedagogical agents. In *AIED*, volume 5, pages 298–305.

Raquel Justo, Thomas Corcoran, Stephanie M. Lukin, Marilyn Walker, and M. Inés Torres. 2014. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*, 69:124–133.

Foaad Khosmood and Marilyn Walker. 2010. Grapevine: A gossip generation system. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, pages 92–99. ACM.

J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document.

Vinodh Krishnan and Jacob Eisenstein. 2015. You're Mr. Lebowski, I'm the Dude: Inducing address term formality in signed social networks. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1616–1626, May–June.

Shibamouli Lahiri, Prasenjit Mitra, and Xiaofei Lu. 2011. Informality judgment at sentence level and experiments with formality score. In *Computational Linguistics and Intelligent Text Processing*, pages 446–457. Springer.

Shibamouli Lahiri. 2015. SQUINKY! A corpus of sentence-level formality, informativeness, and implicature. *arXiv preprint arXiv:1506.02306*.

Haiying Li, Zhiqiang Cai, and Arthur C. Graesser. 2013. Comparing two measures for formality. In *The Twenty-Sixth International FLAIRS Conference*.

François Mairesse and Marilyn A. Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3):455–488.

François Mairesse. 2008. *Learning to adapt in dialogue systems: Data-driven models for personality recognition and generation*. Ph.D. thesis, University of Sheffield, United Kingdom.

Alejandro Mosquera and Paloma Moreda. 2012a. A qualitative analysis of informality levels in web 2.0 texts: The facebook case study. In *Proceedings of the LREC workshop:@ NLP can u tag #user generated content*, pages 23–29.

Alejandro Mosquera and Paloma Moreda. 2012b. Smile: An informality classification tool for helping to assess quality and credibility in web 2.0 texts. In *Proceedings of the ICWSM workshop: Real-Time Analysis and Mining of Social Streams (RAMSS)*.

Kate G. Niederhoffer and James W. Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 218–224, Denver, Colorado, May–June. Association for Computational Linguistics.

Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the Enron corpus. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*,

pages 86–95, Portland, Oregon, June. Association for Computational Linguistics.

Nils Reiter and Anette Frank. 2010. Identifying generic noun phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 40–49, Uppsala, Sweden, July. Association for Computational Linguistics.

Priya Sidhaye and Jackie Chi Kit Cheung. 2015. Indicative tweet generation: An extractive summarization problem? *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Robert J. Sigley. 1997. Text categories and where you can stick them: A crude formality index. *International Journal of Corpus Linguistics*, 2(2):199–237.

Sangweon Suh, Harry Halpin, and Ewan Klein. 2006. Extracting common sense knowledge from wikipedia. In *Proceedings of the Workshop on Web Content Mining with Human Language Technologies at ISWC*, volume 6.

Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. *The International Conference on Language Resources and Evaluation*, pages 812–817.