Noname manuscript No.
(will be inserted by the editor)

SQUINKY! A Corpus of Sentence-level Formality, Informativeness, and Implicature

Shibamouli Lahiri

Received: date / Accepted: date

Abstract We introduce a corpus of 7,032 sentences rated by human annotators for formality, informativeness, and implicature on a 1-7 scale. The corpus was annotated using Amazon Mechanical Turk,[1] and reliability in the obtained judgments was examined by comparing mean ratings across two Mechanical Turk experiments, and correlation with pilot annotations (on sentence formality) conducted in a more controlled setting. Despite the subjectivity and inherent difficulty of the annotation task, correlations between mean ratings were quite encouraging, especially on formality and informativeness. We further explored correlation between the three linguistic variables, genre-wise variation of ratings and correlations within genres, compatibility with automatic stylistic scoring and grammatical aspects of a sentence, and sentential make-up of a document in terms of style. We gave examples of low and high-variance annotations, and annotations that do and do not conform to established formality literature. We analyzed the comments Turkers provided, and investigated how comment length varied with stylistic properties of a sentence. To date, our corpus is the second largest sentence-level annotated corpus released for formality, and the largest released for informativeness and implicature. We also released more than 2,000 anonymized Turker comments as a separate corpus, in the hope that it would spur interesting future research in this domain.

Keywords Formality · Informativeness · Implicature · Dataset · Corpus · Mechanical Turk · sentences · Annotation · Stylistics · Genre · Lexical Density · Likert Scale

A pre-print version of this paper appears on arXiv.

S. Lahiri
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109
Tel.: +1-814-321-7351
E-mail: lahiri@umich.edu

[1] https://www.mturk.com/mturk/welcome.

1 Introduction

Consider the two following utterances:[2]

1. `This is to inform you that your book has been rejected by our publishing company as it was not up to the required standard. In case you would like us to reconsider it, we would suggest that you go over it and make some necessary changes.`
2. `You know that book I wrote? Well, the publishing company rejected it. They thought it was awful. But hey, I did the best I could, and I think it was great. I'm not gonna redo it the way they said I should.`

Not only are the styles of the two utterances different (first one is formal, second one is informal), but they are also targeted at different audiences. While the first one is meant to address a person (*interlocutor*) whose *distance* from the speaker(s) is somehow large (either physically, or conceptually), the second one is meant to address a person in close proximity (again, either physically, or conceptually) – perhaps a friend or a colleague. It shows that we choose and use different *styles* of language depending on *who* we are talking to. This dichotomy of (in)formal expressions was examined in great detail by Heylighen and Dewaele [19]. As they observed, *formality* is the most important dimension of speaking and writing styles (cf. [5,21]),[3] has deep connotations in terms of the social and psychological *situation* of language, and shows close connections to *informativeness* and *implicature*. They argued, in particular, that formality emerges out of a communicative objective – to maximize the amount of information being conveyed to the listener while at the same time maintaining (or at least appearing to maintain) Grice's communicative maxims of Quality, Quantity, Relevance and Manner as much as possible [17].

Heylighen and Dewaele introduced the notion of *deep formality* – "avoidance of ambiguity by minimizing the context-dependence and fuzziness of expression", and reasoned that the other type of formality (*surface formality*; formalizing language for stylistic effects) is a corruption of the language's original deep purpose. Deep formality was characterized by a lack of *contextuality*, evidenced in particular by decreased levels of *deixis* and *implicature* in linguistic realizations, and increased levels of *informativeness*. Deixis indicates a set of words that *point to* another set of words for their meaning [28]. Four types of deixis have been recognized – time, place, person, and discourse. Word correlation studies show that some categories of words are *deictic* (pronouns, verbs, adverbs, interjections), some are *non-deictic* (nouns, adjectives, prepositions, articles), and the rest are *deixis-neutral* (conjunctions) [19]. Heylighen and Dewaele combined the frequencies of deictic and non-deictic words to come up with a measure of formality, known as the "$\mathscr{F}$-score" (Section 2.1).

While several of the arguments Heylighen and Dewaele made are open to question (we will examine some of them in this paper), an important take-home message

[2] Courtesy: `http://www.word-mart.com/html/formal_and_informal_writing.html`

[3] For a general discussion on the linguistic theory of *registers*, see [27] and [28].

from their theory is a so-called *continuum of formality*, arising out of a process where a document (or a piece of text) can be "formalized" *ad infinitum*, simply by adding more and more context. This precludes us from labeling a document or a sentence binarily as "formal" or "informal". Instead, we *rate* sentences on a scale of formality – from very informal to very formal. This Likert scale approach [31] to sentence formality annotation has been shown to work well by Lahiri and Lu [24] on 600 sentences. In this paper, we extend their work to a much larger corpus, and also take into account *informativeness* and *implicature* ratings – in order to probe deeper into some of the original questions raised by Heylighen and Dewaele. We rate 7,032 sentences – each one on a scale from 1 to 7 – for formality, informativeness, and implicature. In some sense, our work is similar to the Stanford politeness corpus [13], as both corpora are at the sentence/utterance level, and both measure pragmatic variables on an ordinal scale (politeness vs formality, informativeness and implicature).

The rest of this paper is organized as follows. In Section 2, we introduce the concepts of Formality, Informativeness and Implicature, and discuss why a thorough understanding of these concepts is necessary for an adequate analysis of this paper.[4] We provide background material, relevant related work, and the definition of the $\mathscr{F}$-score. Section 3 deals with corpus creation, examples of high and low-variance sentences, and sentences that do and do not conform to Heylighen and Dewaele's theory. Detailed exploratory analysis of the data is reported in Section 4, followed by the analysis of Turker comments in Section 5. We conclude in Section 6 with our contributions, limitations of the study, and future research directions. Relevant terminology is introduced as and when they first appear in the paper.

2 Background and Related Work

2.1 Formality

Heylighen and Dewaele's study, while seminal in the field of formality scoring, had its limitations. They stressed the relationship between contextuality (missing information) and implicature, but the relationship was never quantified. They also refrained from quantifying implicature itself – a major component in their theory – to avoid intricacies at the level of phonetics, syntax, semantics and pragmatics, citing that the "recognition of phonetic patterns, syntactical parsing, and even more semantic and pragmatic interpretation of natural language are still extremely difficult... to perform automatically." Further, we suspect that the relation between deep formality and implicature might have been over-emphasized (cf. Section 4.2).

In the end, they quantified formality using deixis only (percentage difference between deictic and non-deictic parts-of-speech), which we will henceforth refer to as the "$\mathscr{F}$-score":[5]

[4] For a more elementary treatment of the issues involved, please see Sections 1 and 2 of [25].

[5] Not to be confused with the harmonic mean of precision and recall.

$$\begin{aligned}\mathscr{F} = (&noun\ frequency\ +\ adjective\ freq. +\ preposition\ freq.\\ &+ article\ freq.\ -\ pronoun\ freq.\ -\ verb\ freq.\\ &- adverb\ freq.\ -\ interjection\ freq.\ +\ 100)/2\end{aligned} \tag{1}$$

where the frequencies are taken as percentages with respect to the total number of words in the document.[6] $\mathscr{F}$-score was used in genre analysis by Nowson et al. [40], and shown to be quite effective in discriminating between the 17 genres used in their study. Further, systematic variation in $\mathscr{F}$-score was observed across gender and personality traits. Teddiman noted in particular that $\mathscr{F}$-score can successfully differentiate between genres, but it cannot explain why the genres are different [52]. $\mathscr{F}$-score was found to be the same for diary entries, and comments on those entries.[7] In follow-up work, Li et al. proposed a version of $\mathscr{F}$-score (called "$\mathscr{CF}$-score") based on Coh-Metrix [16] dimensions of narrativity, referential and deep cohesion, syntactic simplicity and word concreteness [30]. $\mathscr{CF}$-score was better able to discriminate between genres than $\mathscr{F}$-score. A very recent study by Biyani et al. showed that rather surprisingly, Heylighen and Dewaele's $\mathscr{F}$-score ranks among the top-scoring features in *clickbait identification* [6].

In a separate strand of work, Brooke and Hirst [8] identified formality as a continuous *lexical* attribute, and assigned a formality score to a *word* based on its co-occurrece frequency with a hand-picked seed set of formal and informal words, smoothed by Latent Semantic Analysis [9]. Formality of words was further shown to be correlated with other stylistic dimensions such as *concreteness* and *subjectivity* [7].

While all the above studies are very important, they looked at formality from *document* and *word* levels, not from the sentence level. Abu Sheikha and Inkpen [2] equated formality of a sentence with the formality of its corresponding document, and Brooke and Hirst [8] predicted formality of sentences using word-level features. Machili [32] and Peterson et al. [44] looked into formality of emails at workplace, the latter exploring the Enron corpus and how formality varies with social distance, relative power, and the weight of imposition, and the former conducting similar analyses among workplace emails from Greek multinational companies.

As Lahiri et al. [25] showed in their work, sentence formality is *not* the same as document formality. While it is true that sentences do follow document-level trends (academic paper sentences are more formal *on average* than blog and news article sentences, which in turn are more formal *on average* than online forum sentences), it was observed that there is a wide spread among sentences in terms of formality – not all sentences from a document are equally formal (cf. Lahiri and Lu [24], and Section 4.4 of this paper). Lahiri and Lu further showed that there are cases where the words in a sentence are formal, but the sentence as a whole is not (*"For all the stars in the sky, I do not care."*) – thus raising questions regarding a straightforward application of lexical formality to explain sentence formality.[8]

[6] Note that $\mathscr{F}$-score was defined at the *document*-level. We will work with $\mathscr{F}$-score later in this paper at the *sentence*-level.

[7] This could be due to linguistic style co-ordination [12].

[8] Also see the examples given by Potts [47].

The only studies we are aware of that looked into formality annotation of sentences (*utterances*, to be more accurate), are Lahiri and Lu [24], Dethlefs et al. [15], Pavlick and Nenkova [42], and Pavlick and Tetreault [43]. Lahiri and Lu annotated 600 sentences by two undergraduate linguistics students on a Likert scale of 1-5. Inter-rater agreement was shown to improve substantially from binary annotations, which could be attributed to the *continuum of formality* phenomenon described in Section 1. Dethlefs et al., on the other hand, were interested in formality from a natural language generation (NLG) perspective.[9] They annotated utterances using Amazon Mechanical Turk on three dimensions of style – colloquialism (opposite of formality), politeness, and naturalness. A 1-5 Likert scale was used. The problem with this study is that the number of annotated sentences was quite limited, and they came from a restricted class of documents talking about restaurant reviews in a single city. This makes Dethlefs et al.'s corpus unsuitable for our purpose. We wanted a generic corpus of sentences annotated with formality ratings that could help build a sentence formality predictor, so we extended the work of Lahiri and Lu [24] instead.

Pavlick and Nenkova [42] created a formality-annotated corpus of 900 sentences. This corpus was substantially smaller than ours, and did not involve informativeness and implicature ratings. The corpus was created at a time when we were creating our own corpus. Thus, this study happened parallelly to ours, and should be construed as simultaneous. Later, Pavlick took our dataset to expand upon [42] to create a larger dataset of 11,274 sentences in [43]. This study appeared *after* the creation of our dataset, and they cited our pre-print version on arXiv [23]. Currently, Pavlick and Tetreault's dataset of 11,274 sentences is the largest formality-annotated dataset of its kind. However, this dataset does not have informativeness and implicature ratings, because they did not follow up on the lead of Heylighen and Dewaele to explore the connections between formality, informativeness, and implicature.

2.2 Implicature

Other than the lack of quantification of implicature, a second issue with Heylighen and Dewaele's $\mathscr{F}$-score is that it is unreliable on small documents, such as sentences and utterances (cf. [25]). It is therefore of interest to examine if the $\mathscr{F}$-score still correlates with human notion of formality at sentence level (cf. Section 4.2). But perhaps more importantly, it shows a big problem in the conceptualiztion of $\mathscr{F}$-score: it is based on large documents, and a lot of *context*. Large documents make it easy to measure the absence of deixis (on which $\mathscr{F}$-score is based), but they do not ease the measurement of implicature – which is much subtler. Moreover, longer documents may actually have *less* implicature (and therefore be more formal – according to Heylighen and Dewaele), simply because they have more context. Hence, it becomes crucial to measure the *amount of implicature* present in a document (or a sentence, for that matter) – a feat that $\mathscr{F}$-score clearly does not achieve.

Note that in general, it is true that as we add more context to a document (or a sentence), it tends to become longer. The opposite is also true: as we rob a document

[9] Note that the importance of formality in language generation has long been recognized [1,20].

(or sentence) of context, it tends to become shorter (*contextual*). So it could be reasoned that sentences by themselves have a lot of un-stated context as compared to a document (which is usually resolved by looking at *neighboring* sentences – much like resolving the meaning of a word by looking at neighboring words). So if we could somehow estimate the amount of "missing" context in a sentence, we would be one more step ahead in assessing its true formality.

Quantifying the missing context is complicated by the fact that it depends on both deixis and implicature. Deixis uses words that *anchor to* other words for their meaning, thereby suppressing unnecessary repetitions (and at the same time, generating some "missing" context). Implicature, on the other hand, suppresses *whole expressions* – thus missing a *lot* of context – in the hope that the listener will infer them from *background information*. While $\mathscr{F}$-score gives a reasonable estimate of the amount of *relative deixis* present in a sentence, it does not give any estimate of the amount of implicature. This forced us to rate sentences for the amount of implicature they carry (on Likert scale, because implicature is a *continuous attribute* [14] – just like formality). This annotation process not only gave us implicature ratings, but also allowed us to look into how subjective the concept of implicature is (cf. Section 3.2).

Note that Degen [14] had already conducted a similar study on implicature annotation using Mechanical Turk. However, the focus of her study was on one particular type of implicature (*some* but not *all*), and the annotation process was not tied to formality or any other stylistic attribute. Also to be noted is the fact that our annotated corpus of 7,032 sentences is much larger than Degen's corpus of 1,363 utterances.

A general discussion of the vast literature on implicature (starting with Grice [17], and expanded by Harnish [18], among others) is beyond the scope of this paper. Interested readers are referred to the book by Potts [46] for a gentle introduction to the theory of *conventional implicatures* (CIs), and to [3,4,29] for a discussion on *causal implicatures*. Grice also introduced *scalar implicatures* – arguably the most prominent class of implicatures – that equate "some" with "not all" for the sake of politeness. For example, "John likes *some of* the restaurants on Main Street" has the scalar implicature that "John *does not like all* restaurants on Main Street". Papafragou and Musolino [41] discussed the acquisition of scalar implicatures by children, and Carston [10] related scalar implicatures with *relevance* and *informativeness* – a topic we will briefly visit in the next section. Note that the concept of *informativeness* is germane to an adequate treatment of formality, and *relevance* and *informativeness* both have their origins in Grice's maxims [17].

Potts [47] noted that syntax interacts with logical forms in non-trivial ways to offer two complementary hypotheses for conversational implicature – the interactional (game-theoretic) hypothesis, and the grammar-driven hypothesis. It was shown that both hypotheses can explain embedded as well as uncancelable implicatures. Potts [48] further argued that words and phrases can contribute multiple independent pieces of meaning simultaneously, and while the meanings involved are semantically independent, they interact pragmatically to reduce underspecification – much of that pragmatics being driven by conventional implicatures. Vogel et al. [55] designed a multi-agent decentralized POMDP[10] as a computational model of implicature generation. They observed that the model came up with implicature-laden interpretations of the surrounding world in the process of reasoning to maximize joint utility. Tatu and Moldovan [51] extracted implicatures from English and Arabic tweet conversations using logical forms, common-sense knowledge, and Grice's maxims. Their system achieved 70% F-score on English tweets, and 51% F-score on Arabic.

Apart from Degen [14], we are not aware of any work that specifically looked into implicature rating at sentence/utterance level. Degen's work, as we already pointed out, is not tied to formality scoring, so we used our own dataset of 7,032 sentences to rate for both formality and implicature.

2.3 Informativeness

We also rated sentences for informativeness – a trait Heylighen and Dewaele [19] identified with *deep formality*, where language is formalized to communicate meaning more clearly and directly. Heylighen and Dewaele hypothesized that (Gricean) informativeness is the driving force behind deep formality. We will test this hypothesis by checking if the formality of a sentence positively correlates with its informativeness (Section 4.2). Interestingly, Carston [10] independently arrived at a similar conclusion, where she related *informativeness* with *meaning*: "informativeness principles... give rise to... a strengthening or narrowing down of the encoded meaning of the utterance." While Carston's specific argument was tied to scalar implicatures, it is not very far-fetched to see that the same argument would, in effect, also apply to *deep formality* as evinced by Heylighen and Dewaele. It is to be noted that the word *informativeness* has different connotations in different settings. In the machine translation community, for example, the word *informativeness* denotes a type of *fidelity* measure to be applied to the translated text – in order to verify how much content of the original text is preserved under the translation. Informativeness under this setting is evaluated by having human subjects answer multiple-choice reading comprehension questions on the translated text [49]. Informativeness of *words and phrases* is an important parameter in problems ranging from named entity detection [50] to keyword extraction [53]. Under this setting, informativeness is known as *term informativeness* [22,56]. Rennie and Jaakkola [50] showed that term informativeness can be modeled as mixture distributions, and estimated via expectation maximization. A variant of IDF[11] – *Residual IDF* – was shown to be competitive with mixture models. Interestingly, Rennie and Jaakkola pointed out that their term informativeness estimation approach would be especially helpful in "extracting information from *informal*, written communication" (emphasis ours).

Kireyev [22] modeled term informativeness using Latent Semantic Analysis (LSA; cf. Landauer and Dumais [26]), and correlated the predicted informativeness scores with the number of WordNet senses [36], hypernymy, and text genres. Wu and Giles [56] used *citation context* in assessing the informativeness of a term. They applied the resulting scores in keyword extraction, and back-of-the-book index generation

[10] Partially Observable Markov Decision Process.

[11] Inverse Document Frequency.

(cf. Csomai and Mihalcea [11]). Timonen et al. [53] extracted keywords from short and semi-formal documents using term informativeness, and hierarchical agglomerative clustering (HAC). They evaluated word informativeness at three levels – corpus, cluster, and document.

Informativeness is especially important in *extractive* and *abstractive summarization*, where we want the machine-generated summaries to retain as much relevant information of the original document as possible [37,39]. In the context of sentiment summarization, Nishikawa et al. [39] jointly optimized the *sentiment informativeness* and readability of a summary, where the former essentially reduces to packing as many sentiment-bearing words in the summary as possible, and the latter is loosely approximated by the natural ordering of sentences in the original text. A very different way of utilizing informativeness appeared in Molina et al. [37], where the authors divided sentences first into *elementary discourse units* (EDUs), and then deleted the less informative segments to come up with an abstractive (compressed) summary. Informativeness of discourse segments was modeled using *textual energy*, and their grammaticality was estimated using language models.

Last but not the least, the concept of informativeness has been applied to evaluate the connected speech of aphasic adults [38]. In this context, informativeness is measured by Correct Information Unit (CIU) analysis, a method that quantifies how well the patients are able to communicate their thoughts and ideas with others, without focusing on the details of presentation style.

While all the above studies are important in their own right, and ground-breaking in some cases, we found none that specifically looked into informativeness rating of sentences in the context of formality, and there is no publicly available annotated dataset for *sentence informativeness*. In this work, we bridged the gap.

3 Corpus Creation

3.1 Data

Our data comes from the pioneering study of Lahiri et al. [25]. They compiled four different datasets – blog posts, news articles, academic papers, and online forum threads – each consisting of 100 documents. For the blog dataset, they collected most recent posts from the top 100 blogs listed by Technorati[12] on October 31, 2009. For the news article dataset, they collected 100 news articles from 20 news sites (five from each). The articles were mostly from “Breaking News”, “Recent News”, and “Local News” categories, with no specific preference attached to any particular category.[13] For the academic paper dataset, they randomly sampled 100 papers from the CiteSeerX[14] digital library. For the online forum dataset, they sampled 50 random

[12] http://technorati.com/.

[13] The news sites were CNN, CBS News, ABC News, Reuters, BBC News Online, New York Times, Los Angeles Times, The Guardian (U.K.), Voice of America, Boston Globe, Chicago Tribune, San Francisco Chronicle, Times Online (U.K.), news.com.au, Xinhua, The Times of India, Seattle Post Intelligencer, Daily Mail, and Bloomberg L.P.

[14] http://citeseerx.ist.psu.edu/.

Table 1: Mechanical Turk HIT (Human Intelligence Task) details.

HIT Title	How formal is this sentence? English as first language required.
HIT Description	This is a formality survey HIT, where we have three stylistic questions on an English sentence. Please do not enter if you do not have English as first language.
Question 1	How **formal** do you think is the above sentence?
Question 2	How much **information** do you think the above sentence carries?
Question 3	How much do you think the above sentence **implies/suggests**, or leaves to possible interpretations?
Comment Box (with each question)	Please explain the reason behind your choice (optional):

Table 2: Spearman's ρ between the mean ratings obtained from our Mechanical Turk experiments. All results are statistically significantly different from zero, with p-value < 0.0001.

	Overall	**Blog**	**News**	**Forum**
Formality	0.68	0.60	0.35	0.48
Informativeness	0.64	0.63	0.42	0.63
Implicature	0.14	0.19	0.09	0.11

documents crawled from the Ubuntu Forums,[15] and 50 random documents crawled from the TripAdvisor New York forum.[16] The blog, news, paper, and forum datasets had 2110, 3009, 161406 and 2569 sentences respectively.

We manually cleaned and sentence-segmented the blog, news, and forum datasets to come up with 7,032 sentences. The manual cleaning was necessary for our annotation process, because we cannot expect our Mechanical Turk annotators to deal with corrupt/incomplete/inaccurate sentences. The much larger and more complex *paper dataset* was discarded, because manual cleansing and sentence segmentation of text data extracted from PDF was prohibitively time-consuming, and often unsuccessful because of spurious characters, words, and corrupted/missing segments of text. Having said that, the paper dataset is relatively un-interesting from a formality perspective, because as Lahiri et al. showed, paper sentences (automatically segmented) have the lowest standard deviation in terms of formality – they are all highly formal, and the formality does not vary much across sentences [25]. This is unlike other corpora, where reasonable variations exist.

3.2 Annotation

With the 7,032 sentences thus obtained, we conducted two Mechanical Turk annotation experiments. In our first experiment, Turkers were requested to rate sentences on a 1-7 scale for formality, informativeness, and implicature, as follows:

[15] `http://ubuntuforums.org/`.

[16] `http://www.tripadvisor.com/ShowForum-g60763-i5-New_York_City_New_York.html`.

Table 3: Spearman's ρ between the mean formality ratings from Mechanical Turk, and mean formality ratings from [24]. All results are statistically significantly different from zero, with p-value < 0.0001. For the results marked with a *, their p-values are < 0.01.

	Overall	**Blog**	**News**	**Forum**
MTurk Experiment 1	0.78	0.73	0.32*	0.49
MTurk Experiment 2	0.73	0.61	0.30*	0.53

Table 4: Example sentences with high and low mean Mechanical Turk ratings for formality, informativeness, and implicature.

	High	**Low**
Formality	And in its middle-class neighborhoods, 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.	Thanx!
Informativeness	According to the Shanghai Jiao Tong University Press, the press is currently compiling a picture album of Qian and a collection of his writings based on 800-plus-page documents retrieved from the U.S. National Archives, which include details about his encounters with the U.S. government and his trip back home.	Any recommendations?
Implicature	Who will join?	Most mornings they rise before their rooster crows, bolting down a meager breakfast of coconut and chile-spiced vegetables over rice before venturing out on their journey: rowing to school aboard a hand-carved 15-foot sampan.

- Formality scale:
  1. Very informal
  2. Informal
  3. Somewhat informal
  4. In-between
  5. Somewhat formal
  6. Formal
  7. Very formal
- Informativeness scale:
  1. Very uninformative sentence
  2. Uninformative sentence
  3. Somewhat uninformative sentence

4. In-between
5. Somewhat informative sentence
6. Informative sentence
7. Very informative sentence

- Implicature scale:
  1. Very non-implicative
  2. Non-implicative
  3. Somewhat non-implicative
  4. In-between
  5. Somewhat implicative
  6. Implicative
  7. Very implicative

Each sentence was a HIT (Human Intelligence Task), and we requested five *assignments* per HIT so that we could get five independent ratings for each sentence. We requested Turkers with English as first language in our HIT title and description (cf. Table 1). To guard against possible abuse, we required *Turkers from US* as qualification, and took the *average* across five independent ratings to control the variation in individual measurements. Our instructions were minimal (reason explained later in this section) – we started with the two examples given at the beginning of Section 1 to prime the Turkers with the notion of *formality*, and gave them four more links to explore the concept on their own.[17] Then we told them to rate sentences on how formal, informative, and implicative they are (cf. Table 1). The three questions were presented in this order (on a single screen):

- Formality
- Informativeness
- Implicature

Each question involved seven radio buttons (corresponding to seven choices) – only one of which could be selected at a time. Turkers were requested to be *consistent* in their ratings across sentences, and rate sentences independently of each other. The order of presentation of the sentences was scrambled so as to remove any potential sequence effect. We also had optional comment boxes so that Turkers could leave us their thoughts on the annotation process (cf. Table 1). In total, 527 Turkers participated in our first experiment.

Then we conducted a second experiment, which was similar to the first, except that now we added two more requirements – at least 1,000 HITs completed with at least 99% approval rate – on top of the US-based requirement. This resulted in 187 Turkers participating in our second experiment.

Correlations between the mean ratings obtained from these two experiments are shown in Table 2. Note that even without extensive quality control and with relatively weak enforcement of the English-as-first-language policy, Turkers' *mean ratings* correlated pretty well for both formality as well as informativeness, echoing previous

[17] http://www.engvid.com/english-resource/formal-informal-english/, http://dictionary.cambridge.org/us/grammar/british-grammar/formal-and-informal-language, http://www.englishspark.com/informal-language/, http://www.antimoon.com/how/formal-informal-english.htm.

findings by Lahiri and Lu [24]. Note further that even without detailed instructions, Turkers were able to rate subjective concepts like "formality" and "informativeness" quite well, again echoing the findings summarized by Lahiri and Lu. The reasons we did not provide Turkers with extensive and detailed instructions, are:

a We did not want to bias them with our view of the English language (removing *experimenter bias*). Style is a fairly subjective issue, and different people have different opinions. If we force annotators to adopt a particular definition of formality, informativeness, and implicature, then we are effectively tuning them to *our judgments*, rather than observing *what they think*.
b We wanted to see if Likert scale annotations were good enough (as claimed by Lahiri and Lu [24]) to instil sufficient reliability and agreement in formality and related annotation processes, especially between *mean ratings*.
c We wanted to see if mean ratings across multiple raters could effectively eliminate the idiosyncrasies of individual Turkers in a subjective annotation task like this.

Note from Table 2 that the correlation values for implicature are rather low – across all genres (albeit positive). This shows that implicature is arguably the most subjective (and therefore, the *hardest*) among the three pragmatic variables we investigated, echoing Heylighen and Dewaele's suspicion that quantifying implicature may not be feasible by any straightforward syntactic, lexical, or semantic approach. Having said that, Degen has recently conducted implicature annotations at sentence level [14]. Her approach was successful, but the success depended on four crucial assumptions:

1. Degen focused on the *some, but not all* class of implicatures, which is the most prominent among scalar implicatures and hence, (arguably) the *easiest*. We are considering *all* possible implicatures.
2. Degen's approach was *contrastive*, where annotators were asked to *compare* two utterances that only differed in "some" vs "not all". This approach – while it makes the task much more focused and easy for the annotators – is not feasible in our (much broader) setting. We cannot reasonably ask for a contrastive measurement across all combinations of possible variations.
3. Degen essentially focused only on *phrases* that contain "some" vs "not all", not whole sentences.
4. Degen relied on *context information* in the annotation. Sentences before and after the actual utterance were presented to the annotators to make the task easier. This is problematic from our perspective, because we would like to have individual sentences rated for the amount of implicature they carry, *without regard to* context sentences. If we are relying on context, then effectively we are not judging a single sentence, we are judging the context *and* the sentence together.

We believe that if these assumptions were to be relaxed, Degen's approach would prove much less tenable, and may in fact degenerate to something very close to what we did. It will, however, be an interesting idea to *combine* Degen's approach with ours. We leave this line of research to future work.

We further compared our mean *formality ratings* from Mechanical Turk to the mean formality ratings reported by Lahiri and Lu [24] in their "actual" annotation

Table 5: Low and High-variance examples of Formality ratings.

Low Variance	**High Variance**
High Formality: According to a 2009 survey by the Center for American Progress, a think tank, the majority of millennials focus less than their parents on battles over sexual orientation and race.	Making matters far worse is the proliferation of qat trees, which have replaced other crops across much of the country, taking up a vast and growing share of water, according to studies by the World Bank.
Low Formality: 1. Is Levi Johnston screwing with all of us?	What a treat we found.

phase. Results are shown in Table 3. Note that the mean Turker ratings are highly positively correlated with the mean ratings from Lahiri and Lu's *quality-controlled* study – except the *news* genre, where correlations are weaker (also see Table 2). We plan to investigate the news genre in future work. But the overall patterns are strongly encouraging, and validate the idea that an annotated corpus can indeed be built reliably with Likert-scale-style annotations – at least for formality and informativeness.

We show some example high- and low- formality, informativeness and implicature sentences in Table 4.[18] Note that they follow the usual intuitions about formality, informativeness, and implicature quite well; for example, sentences that are high in formality and informativeness, but low in implicature, are longer and more difficult to read. The opposite is also true; informal and uninformative sentences are much shorter, and are often laden with a lot of implicature.[19] For the rest of the paper, we only consider the mean ratings from our *second Mechanical Turk experiment*, which comprises better-qualified Turkers. For notational convenience, *mean ratings* will henceforth be referred to as *Formality*, *Informativeness* and *Implicature*, as appropriate.

3.3 Low and High Variance Sentences

It is to be noted that in our annotation study, different sentences had different *rating variance* for formality, informativeness, and implicature. In this section, we will look into example sentences with high and low rating variance for formality, informativeness, and implicature. This exercise is instructive, because it allows us to examine *sure* sentences that are agreed upon by most of the annotators, and *confusing* or *diffi-*

[18] The full dataset is available at `https://drive.google.com/file/d/0B2Mzhc7popBgdXZmRlg2RUdqdDA/view?usp=sharing`. Examples in Table 4 are from our second Mechanical Turk experiment, which comprises better-qualified Turkers.

[19] Interesting trivia: the title of this paper derives from a sentence in our corpus that is very low in formality and informativeness, and medium in implicature.

Table 6: Low and High-variance examples of Informativeness ratings.

Low Variance	**High Variance**
High Informativeness: CIT dropped 23 cents, or 24 percent, to 72 cents in New York Stock Exchange composite trading yesterday.	Matt Nichols threw for 413 yards and four touchdowns and tied the record for most TD passes in school history as Eastern Washington rolled past Portland State 47-10 on Saturday.
Low Informativeness: Below.	Mr. Azenberg declined to comment on Friday night.

Table 7: Low and High-variance examples of Implicature ratings.

Low Variance	**High Variance**
High Implicature: A Few Buck the Trend	HTH A
Low Implicature: At 6:48 a.m. Friday, the Panamanian tanker Dubai Star spilled bunker oil into the bay as the ship was being refueled about 2 1/2 miles south of the Bay Bridge.	HELP

cult sentences that have a much lower agreement. Low variance in this aspect accords with high agreement.

Table 5 shows high and low-variance examples of formality ratings. Note that the low-variance examples are intuitive. For instance, the long sentence – *"According to a 2009 survey by the Center for American Progress, a think tank, the majority of millennials focus less than their parents on battles over sexual orientation and race."* – indeed shows high formality, whereas the short sentence *"1. Is Levi Johnston screwing with all of us?"* indeed shows low formality. On the other hand, the sentence – *"Making matters far worse is the proliferation of qat trees, which have replaced other crops across much of the country, taking up a vast and growing share of water, according to studies by the World Bank."* – is long and complex, but its high formality is somewhat debatable, because of informal constructions such as *"making matters far worse"* and *"qat trees"* (no capitalization).

An interesting example is the sentence *"What a treat we found."* Here, all the words are formal, but the sentence as a whole is informal. This is similar to the example we discussed in Section 2.1 (*"For all the stars in the sky, I do not care."*)

These examples show that the phenomenon of formality is rather complex, and unless we are aware of pitfalls like these, an effective modeling will be out of reach.

Similar observations are repeated for low and high variance examples of informativeness ratings (Table 6). For instance, the sentence *"CIT dropped 23 cents, or 24 percent, to 72 cents in New York Stock Exchange composite trading yesterday."* is indeed highly informative, with exact numerical quantities, proper nouns such as *"New York Stock Exchange"*, and technical terms such as *"composite trading"*. On the other hand, the one-word sentence *"Below."* carries almost no information.

The high-variance examples are interesting. The sentence *"Matt Nichols threw for 413 yards and four touchdowns and tied the record for most TD passes in school history as Eastern Washington rolled past Portland State 47-10 on Saturday."* is informative, but it does require an understanding of American football jargon, such as *"touchdowns"* and *"TD passes"*. On the other hand, the sentence *"Mr. Azenberg declined to comment on Friday night."* does not have much information, but it at least mentions the person involved (*"Mr. Azenberg"*) and the date/time (*"Friday night"*).

The most interesting examples are perhaps from Table 7, where we show low and high-variance sentences for the implicature rating. Note that our implicature rating is somewhat less reliable (cf. Section 3.2), so crisp examples like these are rather difficult to obtain.

First of all, the sentence *"A Few Buck the Trend"* has high implicature, because it says nothing about *who* buck the trend, and what trend is being bucked. It leaves a lot to guess and possibly infer from unsaid background assumptions. The sentence *"At 6:48 a.m. Friday, the Panamanian tanker Dubai Star spilled bunker oil into the bay as the ship was being refueled about 2 1/2 miles south of the Bay Bridge."*, on the other hand, has much lower implicature, because it leaves very little to guessing or imagination.

The high-variance examples of Table 7 are somewhat trickier to analyze. First, note that the sentence *"HELP"* is a *plea for help*, and hence has low implicature (although it says nothing about the actual context). It is also an example of an *imperative sentence*. On the other hand, the sentence *"HTH A"* is clearly high-implicature, because it requires us to know the fact that *"HTH"* means *"hope that helps"*, and *"A"* possibly refers to the first name of a person. But one might argue that this is a *subjunctive sentence*, which leaves little reason to infer context information – something that would have been necessary were it to be a *descriptive sentence*.

3.4 Positive and Negative Examples

It follows from Heylighen and Dewaele's theory that as the (deep) formality of a piece of text increases, its informativeness increases and implicature decreases [19]. Hence, we should see a positive correlation between formality and informativeness, and a negative correlation between formality and implicature. Whether that holds in practice, will be examined in detail in Section 4.2. In this section, we explore *examples* of sentences where the theoretical predictions hold true (a majority of the formality-informativeness cases in our data), and where they do not. We call the first

Table 8: Positive and Negative examples between Formality and Informativeness.

Positive Examples	**Negative Examples**
High Formality, High Informativeness: As Maoists menace continued to be unabated, the government is all set to launch the much-awaited full-fledged anti-Naxal operations at three different areas, considered trijunctions of worst Naxal-affected states.	High Formality, Low Informativeness: 4) "We find no clear relation between income inequality and class-based voting."
Low Formality, Low Informativeness: A BIG THANKYOU GOES TO holli!	Low Formality, High Informativeness: 2) Just wipe the Mac OS X partition when u install the dapper.

Table 9: Positive and Negative examples between Formality and Implicature.

Positive Examples	**Negative Examples**
High Formality, Low Implicature: Maoists sabotaged Essar's 166-mile underground pipeline, which transfers slurry from one of India's most coveted iron ore deposits to the Bay of Bengal.	High Formality, High Implicature: All seven aboard the Coast Guard plane are stationed at the Coast Guard Air Station in Sacramento, Calif., where their aircraft was based.
Low Formality, High Implicature: alright, well, i guess i just made a newbie mistake.	Low Formality, Low Implicature: Wait.

type of sentences *positive examples* (predictions hold true), and the second type of sentences *negative examples* (predictions do not hold).

Table 8 shows positive and negative examples between formality and informativeness. The sentence *"As Maoists menace continued to be unabated, the government is all set to launch the much-awaited full-fledged anti-Naxal operations at three different areas, considered trijunctions of worst Naxal-affected states."* is a positive example, because it is both highly formal (formality rating 6.8), and very informative (informativeness rating 6.2). On the other hand, the sentence *"A BIG THANKYOU GOES TO holli!"* is also a positive example, because it is highly informal (formality rating 1.0), and does not carry much information (informativeness rating 2.2).

The negative examples in Table 8 require more deliberation. The sentence *"4) "We find no clear relation between income inequality and class-based voting.""* is highly formal (formality rating 6.4), but not as informative (informativeness rating 5.2), because the pronoun *"We"* has not been resolved to any named entity, measures of *"income inequality"* and *"class-based voting"* have been left unspecified, and no numbers have been given for *"no clear relation"*. It seems like the sentence glosses over several important details to make a very general and rather strong statement. On the other hand, the sentence *"2) Just wipe the Mac OS X partition when u install the dapper."* is very informal (formality rating 1.4), but it provides a clear and precise instruction, and does so very concisely and cogently. It gives most of the relevant information (if not all), sounds thorough and authoritative, and leaves relatively little reason to ask clarifying questions. Hence, its informativeness rating (4.8) is much higher than its formality rating.

Table 9 shows positive and negative examples between formality and implicature. Note that according to Heylighen and Dewaele, high formality should correlate with low implicature (and *vice versa*), hence the positive examples in Table 9 refer to either high formality and low implicature, or low formality and high implicature.

The sentence *"Maoists sabotaged Essar's 166-mile underground pipeline, which transfers slurry from one of India's most coveted iron ore deposits to the Bay of Bengal."* is high in formality (formality rating 6.4), and leaves relatively little to infer from unsaid background assumptions (implicature rating 3.0). Hence, it is a positive example. On the other hand, the sentence *"alright, well, i guess i just made a newbie mistake."* has low formality (formality rating 1.2), and leaves several aspects to be inferred (Who is *"i"*? What is a *"newbie mistake"*? What *"mistake"* was made?), leading to an implicature rating of 5.2. Hence, it is also a positive example.

The negative examples in Table 9 are more interesting, and also more complex. The sentence *"All seven aboard the Coast Guard plane are stationed at the Coast Guard Air Station in Sacramento, Calif., where their aircraft was based."* is high in formality (formality rating 6.4), but it leaves two important questions unanswered:

1. Who are *"All seven"*?
2. What type of *"plane"* or *"aircraft"* is being discussed?

which leads to a relatively high implicature rating of 4.8. On the other hand, the sentence *"Wait."* is very low in formality (formality rating 1.4), but it also has relatively low implicature (implicature rating 3.0), because it is an *instruction* – clear and cogent. There is not much to be inferred from background assumptions. Incidentally, it is also an *imperative sentence*.

4 Exploratory Analysis

We performed four separate experiments on the 7,032 annotated sentences to identify different aspects of the annotations. In our first experiment, we explored how sentence-level formality, implicature, and informativeness vary across three different online genres – news, blog, and forums (Section 4.1). In the second experiment, we investigated the correlation among these three variables, and correlation with stylistic scores (Section 4.2). Our third experiment investigated the correlations between stylistic variables and grammatical complexity measures (Section 4.3). Finally, in Section 4.4, we examined how documents varied in terms of sentential formality, informativeness, and implicature – on average.

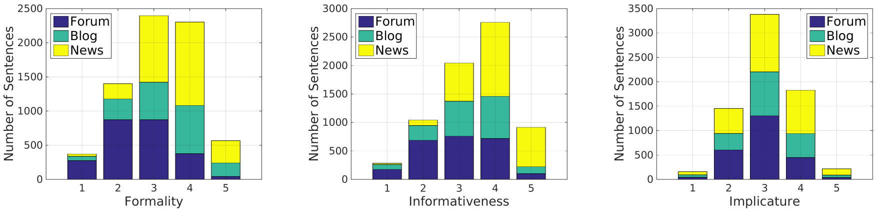

Fig. 1: Genre-wise variation of formality, informativeness, and implicature (can be viewed in grayscale).

4.1 Genre-wise Variation

We plot five-bin histograms of formality, informativeness, and implicature in Figure 1. Note from Figure 1 that *overall*, our corpus is dominated by high-informativeness, mid-to-high-formality, and mid-implicature sentences. Since our implicature rating is less reliable than the other two ratings (cf. Section 3.2), this *mid-implicature* trend should be considered with a grain of salt. It could either be a real phenomenon, or it could be a reflection of *central tendency bias* affecting the annotators – who, lacking a better choice and a better interpretation – chose middling values for the implicature rating. Central tendency in implicature is observed for all three individual genres – news, blog, forums.

The news genre is dominated by high-informativeness, and mid-to-high-formality sentences; blogs, too, are mostly high-formality and mid-to-high-informativeness sentences; forums, on the other hand, are dominated by mid-to-low-formality sentences, and are spread out almost evenly when it comes to informativeness. The general trends corroborate earlier studies [24, 25].

The fact that forums are spread out in terms of (sentential) informativeness shows that there are all kinds of sentences in forums – some are very informative, some are somewhat informative, and some are uninformative (e.g., help-eliciting setences such as “help please!”, sentences expressing gratitude such as “Thanks everybody!”, and suggestive sentences such as “give it a shot.”). Filtering forum sentences by informativeness may be a useful first step towards effective mining of forum data.

4.2 Relationship with Others

We wanted to see how human ratings of three pragmatic variables (formality, informativeness, implicature) correlated among themselves, and also with automatic measures of stylistic scoring. While Lahiri et al. [25] showed that $\mathscr{F}$-score correlates positively with *reading difficulty*, it will be instructive to look into what other variables (if any) the human ratings correlated with. We experimented with eight different sentential stylistic variables, as follows:

1. **Fo:** Formality of the sentence, i.e., the mean formality rating assigned by Turkers in our second Mechanical Turk experiment.

Table 10: Spearman's ρ between stylistic variables, as explained in text. Most of the results are statistically significantly different from zero, with p-value < 0.0001. For the results marked with a *, p-values are < 0.01; for those marked with a **, p-values are < 0.05. Results in *italics* are statistically insignificant.

Overall

	Fo	**In**	**Im**	**Lw**	**Lc**	**F**	**I**	**LD**
Fo	1.00	0.73	0.07	0.55	0.59	0.34	0.03*	*0.01*
In		1.00	0.05	0.62	0.65	0.31	0.05	*-0.02*
Im			1.00	0.10	0.10	-0.06	0.03**	*0.00*
Lw				1.00	0.98	0.23	0.12	-0.18
Lc					1.00	0.28	0.07	-0.08
F						1.00	-0.14	0.04*
I							1.00	*-0.02*
LD								1.00

Blog

	Fo	**In**	**Im**	**Lw**	**Lc**	**F**	**I**	**LD**
Fo	1.00	0.73	-0.10	0.51	0.54	0.33	0.07*	*-0.04*
In		1.00	-0.08*	0.62	0.65	0.29	0.06**	-0.06*
Im			1.00	*0.02*	*0.01*	-0.18	*0.04*	*-0.02*
Lw				1.00	0.98	0.18	0.13	-0.23
Lc					1.00	0.23	0.08*	-0.15
F						1.00	-0.12	0.06*
I							1.00	-0.06**
LD								1.00

News

	Fo	**In**	**Im**	**Lw**	**Lc**	**F**	**I**	**LD**
Fo	1.00	0.63	-0.08	0.34	0.38	0.27	*-0.01*	*0.00*
In		1.00	-0.10	0.43	0.45	0.28	*-0.01*	*-0.02*
Im			1.00	*-0.01*	*-0.02*	-0.12	*0.02*	*0.00*
Lw				1.00	0.98	0.21	0.08	-0.17
Lc					1.00	0.27	*0.03*	-0.08
F						1.00	-0.15	*-0.03*
I							1.00	0.05*
LD								1.00

Forum

	Fo	**In**	**Im**	**Lw**	**Lc**	**F**	**I**	**LD**
Fo	1.00	0.57	*0.04*	0.42	0.43	0.07*	0.16	-0.07*
In		1.00	0.08	0.58	0.60	0.09	0.16	-0.08*
Im			1.00	0.06*	0.05*	-0.08*	0.05**	*-0.02*
Lw				1.00	0.97	*0.02*	0.23	-0.26
Lc					1.00	0.06*	0.19	-0.15
F						1.00	-0.12	*0.01*
I							1.00	*-0.03*
LD								1.00

2. **In:** Informativeness of the sentence, i.e., the mean informativeness rating assigned by Turkers in our second Mechanical Turk experiment.
3. **Im:** Implicature of the sentence, i.e., the mean implicature rating assigned by Turkers in our second Mechanical Turk experiment.
4. **Lw:** Length of the sentence in words.
5. **Lc:** Length of the sentence in characters.
6. **F:** Formality score of the sentence, as proposed by Heylighen and Dewaele [19].
7. **I:** Ambiguity score of the sentence.
8. **LD:** Lexical density of the sentence (Ure [54]).

Among these variables, Heylighen and Dewaele's formality score is given by:

$$\begin{aligned} \mathscr{F} = (noun\,frequency\; +\; adjective\,freq. + \;preposition\,freq. \\ +\,article\,freq.\; -\; pronoun\,freq.\; -\; verb\,freq. \\ -\,adverb\,freq.\; -\; interjection\,freq.\; +\; 100)/2 \end{aligned} \tag{2}$$

where the frequencies are taken as percentages with respect to the total number of words in the sentence. The inspiration for this score comes from the fact that nouns, adjectives, prepositions, and articles are found to be *non-deictic* in word correlation studies, whereas pronouns, verbs, adverbs, and interjections are found to be *deictic*.[20] $\mathscr{F}$-score measures formality as the amount of *relative non-deixis* present in a sentence (cf. Section 2.1).

Ure's lexical density takes the form:

$$\mathscr{L}\mathscr{D} = 100\left(\frac{N_{lex}}{N}\right) \tag{3}$$

where N_{lex} is the number of *lexical tokens* (nouns, adjectives, verbs, adverbs) in the sentence, and N is the total number of words in the sentence.

The *ambiguity score* (**I**) is a scoring formula we propose in this paper. The idea is as follows. Recall from Section 1 that *contextuality* – the opposite of *deep formality* – is affected by both deixis as well as implicature. Although implicature is hard to quantify, a measure of "ambiguity" in a given piece of text can be formulated by counting how many WordNet senses [36] the words in that text carry on average. The more senses words have, the more semantically ambiguous the text is. The *ambiguity score* (**I**) of a sentence is thus given by the *average number of WordNet senses per word in the sentence*.

Correlations between the eight variables are given in Table 10. Note from Table 10 that formality and informativeness are highly correlated in all cases, thereby validating Heylighen and Dewaele's hypothesis that the purpose of formality (*deep formality* in particular) is *more informative communication*. Note, however, that in most cases, there is very little correlation between formality and implicature (small positive/negative values). There are two possible reasons for this: (a) implicature is a *broad* phenomenon, and maybe formality and implicature are not as antagonistically related as argued by Heylighen and Dewaele; (b) our implicature annotation by Turkers showed a *central tendency bias* and low agreement between two Mechanical Turk experiments, so maybe the mean implicature ratings we obtained are not truly reflective of the actual amount of implicature present in a sentence. Implicature at any rate is a *hard* pragmatic variable to track, so validating which of these two (or maybe both) is the correct reason behind sub-optimal correlation values, constitutes a part of our future work.

Note further from Table 10 that formality and informativeness are positively correlated (moderate-to-good correlation) with length of the sentence – in words and characters. This corroborates the earlier finding by Lahiri et al. [25] that as a piece of

[20] Conjunctions are deixis-neutral. We used CRFTagger [45] to part-of-speech-tag our sentences.

text gets more formal, it tends to become longer and more intricate, leading to higher *reading difficulty*. Formality and informativeness also correlate positively (moderate correlation) with Heylighen and Dewaele's $\mathscr{F}$-score, except in the Forum genre. On the other hand, they do not have significant correlations with the ambiguity (**I**) score except the Forum genre. Implicature has a significant, but small negative correlation with $\mathscr{F}$-score in all cases. Lexical density negatively correlates with length of the sentence (both in words and characters). Ambiguity score correlates positively with length, but negatively with Heylighen and Dewaele's $\mathscr{F}$-score, as expected. Implicature also correlates negatively with $\mathscr{F}$-score in all cases. This is an important finding, because it indicates that humans tend to rate formal sentences *low* in terms of implicature (and *vice versa*) – a key hypothesis in Heylighen and Dewaele's formulation. The two length scores have an almost perfect positive correlation among them, which is unsurprising.

The surprising part, however, is that formality and informativeness (as rated by humans) are not very highly correlated (either positively or negatively) with Heylighen and Dewaele's $\mathscr{F}$-score or our ambiguity (**I**) score. Maybe these two scores are measuring complementary aspects of the phenomenon of formality, and are not individually able to explain all the variations. Automated scoring/prediction of formality by modeling it on top of scores like these (perhaps as features) is our future plan. We would also like to investigate how to predict informativeness, and how to get a better handle on implicature scoring – both by humans as well as automated.

4.3 Relationship with Grammatical Variables

In this section, we investigate the relationship between eight sentential stylistic variables (cf. Section 4.2), and six measures of *grammatical complexity* of a sentence. The reason we would like to investigate this relationship, is because Lahiri et al. [25] showed that formality has a positive linear correlation with *reading difficulty*, and reading difficulty often indicates syntactic complexity. We chose six different grammatical complexity measures, as follows:

1. **PO:** Total number of unique part-of-speech tags in the sentence.
2. **PD:** Depth of the constituency parse tree of the sentence.
3. **NP:** Total number of non-terminal production rules in the constituency parse tree of the sentence.
4. **UP:** Total number of *unique* non-terminal production rules in the constituency parse tree of the sentence.
5. **ND:** Total number of dependency types in a dependency parse of the sentence.
6. **UD:** Total number of *unique* dependency types in a dependency parse of the sentence.

We used CRFTagger [45] to part-of-speech-tag the sentences, and Stanford CoreNLP Pipeline [33] to obtain the constituency and dependency parses. By *dependency type*, we mean the type of dependency relation (*nsubj*, *dobj*, *nmod*, *amod*, *advmod*, *xcomp*, etc) as described in the Universal Stanford Dependencies [34]. We did not take words and other tokens into account while obtaining the dependency types.

Table 11: Spearman's ρ between stylistic variables and grammatical variables, as explained in text. Most of the results are statistically significantly different from zero, with p-value < 0.0001. For the results marked with a *, p-values are < 0.01; for those marked with a **, p-values are < 0.05. Results in *italics* are statistically insignificant.

Overall

	PO	**PD**	**NP**	**UP**	**ND**	**UD**
Fo	0.43	0.39	0.46	0.45	0.54	0.5
In	0.52	0.45	0.54	0.53	0.62	0.58
Im	0.12	0.11	0.1	0.1	0.1	0.11
Lw	0.88	0.8	0.95	0.93	1.0	0.95
Lc	0.86	0.77	0.92	0.9	0.98	0.93
F	*0.0*	*-0.0*	0.06	0.03*	0.22	0.15
I	0.11	0.19	0.13	0.14	0.11	0.14
LD	-0.14	-0.18	-0.2	-0.15	-0.16	-0.17

Blog

	PO	**PD**	**NP**	**UP**	**ND**	**UD**
Fo	0.37	0.36	0.43	0.41	0.5	0.45
In	0.51	0.45	0.55	0.54	0.62	0.57
Im	0.07*	0.07*	0.05**	0.06*	*0.02*	*0.04*
Lw	0.87	0.8	0.95	0.93	1.0	0.95
Lc	0.85	0.77	0.92	0.91	0.98	0.93
F	-0.06*	*-0.03*	*0.03*	*-0.0*	0.17	0.08*
I	0.12	0.2	0.14	0.16	0.12	0.15
LD	-0.21	-0.23	-0.26	-0.21	-0.22	-0.23

News

	PO	**PD**	**NP**	**UP**	**ND**	**UD**
Fo	0.19	0.23	0.26	0.26	0.33	0.31
In	0.27	0.26	0.35	0.34	0.43	0.39
Im	0.05*	0.05*	*0.02*	*0.03*	*-0.01*	*0.01*
Lw	0.78	0.72	0.93	0.9	1.0	0.92
Lc	0.74	0.69	0.89	0.87	0.98	0.89
F	-0.18	-0.07*	*-0.0*	-0.06*	0.2	0.08
I	0.09	0.16	0.1	0.12	0.08	0.12
LD	-0.12	-0.16	-0.2	-0.12	-0.16	-0.16

Forum

	PO	**PD**	**NP**	**UP**	**ND**	**UD**
Fo	0.38	0.33	0.36	0.35	0.41	0.41
In	0.54	0.45	0.52	0.51	0.58	0.57
Im	0.05**	0.06*	0.06*	0.05**	0.07*	0.07*
Lw	0.93	0.84	0.95	0.94	1.0	0.97
Lc	0.92	0.81	0.93	0.91	0.97	0.95
F	-0.05**	-0.12	-0.1	-0.1	*0.01*	*-0.01*
I	0.19	0.26	0.23	0.23	0.23	0.23
LD	-0.19	-0.22	-0.24	-0.22	-0.23	-0.23

The correlations between eight stylistic variables and six grammatical variables across three genres and overall are shown in Table 11. Qualitatively, patterns and trends are similar across genres, so we focus on the "Overall" part of the table. First of all, note that the *length of the sentence* (**Lw** and **Lc**) correlates almost perfectly with all the syntactic complexity measures. This shows that as a sentence gets longer, it invariably attracts greater levels of grammatical complexity. This is not too surprising, because any long text requires *hierarchical organization* to convey ideas clearly and effectively. Syntactic complexity measures gauge the amount of hierarchical organization present in a sentence (or any given piece of text).

Table 11 further indicates that human ratings of formality (**Fo**) and informativeness (**In**) strongly correlate with syntactic complexity measures (*news* genre shows somewhat weaker correlations). Highest correlations are found in the **ND** column, which implies that from a formality and informativeness perspective, dependency types are important measures of syntactic complexity. Table 11 validates our earlier speculation that with increases in formality and informativeness, a sentence becomes more and more difficult to read, at least partly owing to additional grammatical complexity. It is also to be noted that the human rating of implicature (**Im**) has very low and/or statistically insignificant correlations with all grammatical variables, which accords with our earlier observation about the relative unreliability of the implicature ratings we obtained.

Interestingly, both $\mathscr{F}$-score (**F**) and ambiguity score (**I**) have low and/or statistically insignificant correlations with grammatical variables, which means that they do not directly measure the grammatical complexity of a sentence. The reason why this happens is not very clear to us, and it will be an interesting research direction to pursue in future. The most surprising part of the table is the fact that Ure's Lexical Density (**LD**) correlated negatively with all grammatical variables. An investigation of the possible reasons behind this phenomenon would constitute a very exciting line of research.

4.4 Sentential Make-up of Documents

In our final experiment, we investigated how the sentences in a document vary in terms of formality, implicature, and informativeness – starting from the beginning sentences, then the middle ones, and finally the last ones. We divided the sentences into ten successive bins (*deciles*) based on their position in the document, and measured the mean formality, informativeness, and implicature *per decile*. The results – averaged across all documents in a particular genre (blog, forums, news, overall) – are shown in Figure 2. Figure 2 also shows the standard errors for each decile.

Note from Figure 2 that news sentences are most formal and most informative, followed by blog sentences, followed by forum sentences. This observation corroborates Lahiri et al.'s findings [25]. In terms of formality and informativeness trends, news sentences start with high formality and informativeness, then gradually diminish in both – perhaps reflecting the fact that in journalistic writing, first few sentences carry the most information (to catch the readers' attention), and the information/interesting-ness content decreases substantially thereafter. Forum sentences, on the other hand, maintain a low level of formality and informativeness throughout – with a few small peaks and valleys in-between. For blogs, the trend is first decreasing, then increasing, and then decreasing again – indicating that the most informative (and formal) sentences in blogs are in the middle. All three genres taken together, both formality and informativeness show a decreasing trend. There is no clear trend in the implicature rating of sentences – it is mostly an assortment of peaks and valleys.

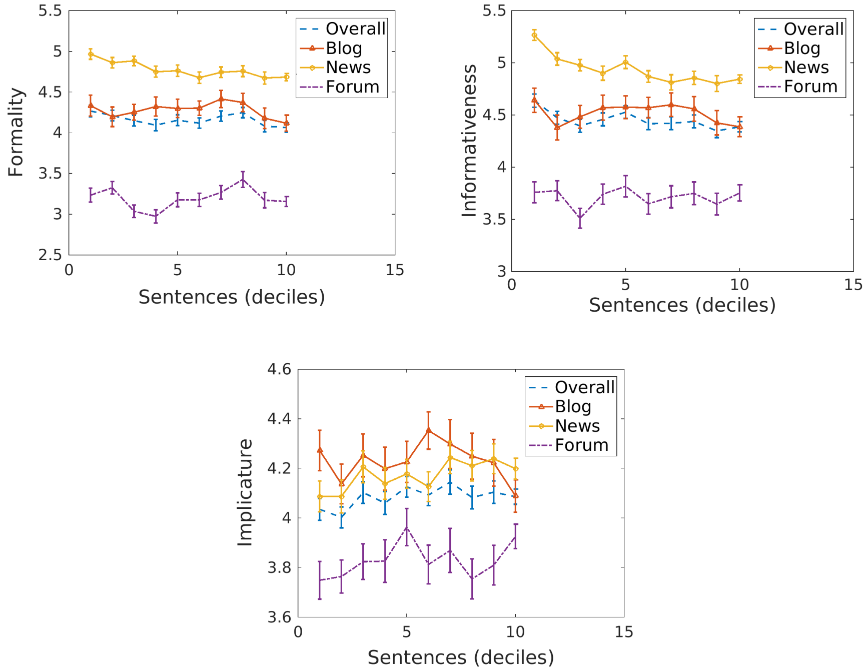

Fig. 2: Sentential make-up of formality, informativeness, and implicature.

Table 12: Mechanical Turk comment statistics.

	Formality	Informativeness	Implicature	Total
Number of comments	908	677	677	2262
Number of unique sentences with comments	861	650	657	2168

5 Analysis of Turker Comments

Recall from Section 3.2 that we had an optional comment box along with each of the three questions we asked on Amazon Mechanical Turk (cf. Table 1). This allowed us to collect Turkers' thoughts on the annotation process in general, and sentences that they were working on in particular. Analysis of these comments is an intriguing and exciting research study in itself, as it allows us to peer into the strengths and weaknesses of each annotator, lets us infer whether they were doing a good job or not, and in general reveals to us the reasoning and thought processing going behind each of the commented ratings. Our analysis in this section is necessarily preliminary, but we release the comments as a separate corpus for future research.[21] What we did observe, however, is as follows:

[21] The comments are available from `http://web.eecs.umich.edu/~lahiri/turker_comments.zip`.

Table 13: Comments Example 1.

Original sentence (from *news* corpus)	At least seven people were killed and several were missing.
Formality comment 1	Statement of fact.
Informativeness comment 1	Tells you how many people and what they were affected by.
Informativeness comment 2	Several is a vague statement
Informativeness comment 3	I think this gives you a good amount of information. I feel it would be very informative if they added context to the missing. How much estimated?
Implicature comment 1	That there was a severe event that occurred.
Implicature comment 2	It leaves what "several" means open to interpretation

Table 14: Comments Example 2.

Original sentence (from *news* corpus)	Although Peter has a shelf of city league soccer trophies, he and the others are mediocre pupils at best, sources said.
Formality comment 1	has some formal language structure, but not totally.
Formality comment 2	I don't think it is a concise sentence. I'd like to know what the source is in reference to Peter and the other pupils. I think the best thing to do would either change that, or the word "said" to "say."
Formality comment 3	by adding sources it makes the statement seems to have an air of credibility and formaility.
Formality comment 4	reads like a newspaper line
Informativeness comment 1	It states information, but not important information.
Informativeness comment 2	Even thought Peter has a shelf of city trophies, he's still a mediocre player, in a much wider field of similar athletes.
Implicature comment 1	The first part states a fact but the other is not factual as it doesn't offer any proof.
Implicature comment 2	Peter isn't as good outside of the city he plays in.
Implicature comment 3	It implies to me that Peter really enjoys soccer, and doesn't like school. It also implies that he's been playing awhile, as a shelf full of trophies would take some time to accumulate

Table 15: Comments Example 3.

Original sentence (from *blog* corpus)	Apple has created a media consumption experience that has reduced friction to such a point that soon the consumer will not know if he is buying music, a movie or a game.
Formality comment 1	Basic concept given-media consumption saturating the market, get rid of the word friction (too misleading), blurs the ability of someone recognizing it as such, when purchasing music, movies and games.
Formality comment 2	This is writing you would find reviewing a product on a website.
Informativeness comment 1	Basically stating, that Apple has found electronic solutions in the genres of music, movies and gaming.
Informativeness comment 2	This is stating an opinion, not offering information.
Informativeness comment 3	The sentence makes a claim, but provides provides no evidence, and the claim seems suspect.
Implicature comment 1	Apples taking over...look out!! Ho HO HA!!
Implicature comment 2	Again, the sentence makes a claim about consumers. Not sure if the claim is valid, or even how it could be.

Table 16: Comments Example 4.

Original sentence (from *blog* corpus)	A $1999 option upgrades the CPU to a “Nehalem” quad core 2.66GHz Core i5 750 processor, or for $200 more, a Core i7 860, both of which include 8MB of L3 cache.
Formality comment 1	Most of the words are technical terms.
Formality comment 2	This sounds like an attempt to be formal and professional, but comes across a bit muddled and rushed.
Informativeness comment 1	The sentence gives information, but also requires one to ask further questions to find out more.
Informativeness comment 2	Informative, but almost to a fault. Trying to cram too much information together, without adequate transitions between the offerings. And we still aren’t being given information about what makes these upgrades valuable.
Informativeness comment 3	Very detailed.
Implicature comment 1	Suggesting to purchase the upgrade options, but not providing reasons for consumers wanting/needing the upgrades.

Table 17: Spearman's ρ between comment length and stylistic variables, as explained in text. Most of the results are too low, and/or statistically insignificant, with p-value ≥ 0.0001. For the results marked with a *, p-values are < 0.01; for those marked with a **, p-values are < 0.05. Results in *italics* are statistically insignificant.

Overall

	FO_C	**FO_W**	**IN_C**	**IN_W**	**IM_C**	**IM_W**
Fo	0.13	0.12*	*0.0*	*0.0*	*0.05*	*0.04*
In	0.2	0.18	0.14*	0.13*	0.13*	0.12*
Im	*0.06*	0.07**	*0.01*	*0.01*	0.13*	0.12*
Lw	0.16	0.14	0.14*	0.12*	0.12*	0.11*
Lc	0.17	0.14	0.13*	0.11*	0.1*	0.09**
F	*-0.01*	*-0.01*	*-0.06*	*-0.06*	-0.08**	-0.09**
I	*0.05*	*0.05*	*0.01*	*0.01*	*0.05*	*0.05*
LD	*-0.07*	-0.07**	-0.09**	*-0.07*	-0.11*	-0.11*

Blog

	FO_C	**FO_W**	**IN_C**	**IN_W**	**IM_C**	**IM_W**
Fo	*0.06*	*0.03*	*0.08*	*0.06*	*0.02*	*-0.01*
In	0.14**	0.14**	*0.11*	*0.07*	*0.06*	*0.03*
Im	0.14**	0.16**	*-0.01*	*-0.02*	0.19**	0.18**
Lw	0.15**	0.15**	0.17**	*0.14*	0.16**	*0.13*
Lc	0.14**	*0.13*	0.17**	*0.15*	*0.13*	*0.1*
F	*-0.02*	*-0.02*	*-0.04*	*-0.05*	*-0.11*	*-0.12*
I	*0.1*	*0.11*	*-0.09*	*-0.1*	*0.04*	*0.05*
LD	*-0.06*	*-0.08*	*-0.02*	*-0.02*	*-0.12*	*-0.11*

News

	FO_C	**FO_W**	**IN_C**	**IN_W**	**IM_C**	**IM_W**
Fo	*0.06*	*0.03*	*-0.02*	*-0.02*	*-0.02*	*-0.06*
In	0.13**	*0.1*	*0.11*	*0.12*	0.17*	0.16**
Im	*0.08*	*0.11*	*0.1*	*0.09*	*0.08*	*0.07*
Lw	*0.11*	*0.08*	*0.05*	*0.04*	*0.03*	*0.01*
Lc	*0.11*	*0.08*	*0.06*	*0.04*	*0.02*	*-0.0*
F	*0.01*	*0.0*	*-0.07*	*-0.06*	*-0.12*	-0.13**
I	*-0.07*	*-0.08*	*0.02*	*0.02*	*0.03*	*0.02*
LD	*-0.1*	*-0.1*	*-0.11*	*-0.1*	-0.13**	*-0.13*

Forum

	FO_C	**FO_W**	**IN_C**	**IN_W**	**IM_C**	**IM_W**
Fo	0.13**	0.13**	*0.08*	*0.08*	0.17*	0.17*
In	0.24	0.2	0.27	0.25	0.17*	0.17*
Im	*-0.07*	*-0.07*	*-0.04*	*-0.02*	0.14**	0.13**
Lw	0.14*	0.11**	0.27	0.23*	0.2*	0.21*
Lc	0.15*	0.11**	0.25	0.21*	0.19*	0.19*
F	*-0.07*	*-0.07*	*-0.06*	*-0.05*	*-0.04*	*-0.05*
I	*0.09*	*0.08*	*0.06*	*0.07*	*0.06*	*0.07*
LD	*-0.06*	*-0.06*	*-0.11*	*-0.09*	*-0.09*	*-0.09*

- Turkers (those who commented) understood the annotation task well. In fact, they understood it very well, presumably at expert levels (examples later).
- We can have very important insights and design classification/regression features based on Turker comments.
- The *implicature* task was hard.
- Turkers understood nuances, fine shades of meaning, and even inferred which genre of documents a sentence was from (examples later).
- Turkers opined on best writing practices, linguistic trivia, and different shades of opinion.

Table 18: Spearman's ρ between comment length and sentence readability, as explained in text. Most of the results are too low, and/or statistically insignificant, with p-value $\geq$ 0.0001. For the results marked with a *, p-values are $<$ 0.01; for those marked with a **, p-values are $<$ 0.05. Results in *italics* are statistically insignificant.

Overall

	FO_C	FO_W	IN_C	IN_W	IM_C	IM_W
FRE	*-0.05*	*-0.03*	*-0.03*	*-0.01*	*-0.0*	*0.02*
ARI	0.1*	0.07**	*0.06*	*0.04*	*0.02*	*-0.0*
FKR	0.1*	0.07**	*0.07*	*0.05*	*0.05*	*0.03*
CLI	0.09**	*0.05*	*0.05*	*0.03*	*-0.01*	*-0.04*
GFI	0.15	0.12*	0.12*	0.09**	0.1**	0.09**
SMOG	0.09*	0.07**	*0.03*	*0.01*	*0.05*	*0.03*

Blog

	FO_C	FO_W	IN_C	IN_W	IM_C	IM_W
FRE	*-0.08*	*-0.05*	*-0.1*	*-0.09*	*-0.08*	*-0.07*
ARI	*0.09*	*0.07*	0.16**	*0.15*	*0.07*	*0.05*
FKR	*0.11*	*0.09*	*0.14*	*0.12*	*0.12*	*0.1*
CLI	*0.04*	*-0.01*	*0.12*	*0.12*	*-0.03*	*-0.04*
GFI	*0.13*	*0.12*	*0.13*	*0.11*	*0.13*	*0.1*
SMOG	*0.1*	*0.07*	*0.04*	*0.04*	*0.1*	*0.08*

News

	FO_C	FO_W	IN_C	IN_W	IM_C	IM_W
FRE	*-0.09*	*-0.05*	*-0.01*	*-0.01*	*0.05*	*0.07*
ARI	*0.1*	*0.06*	*0.04*	*0.03*	*-0.01*	*-0.04*
FKR	*0.11*	*0.07*	*0.03*	*0.03*	*-0.02*	*-0.05*
CLI	*0.08*	*0.03*	*0.06*	*0.05*	*-0.0*	*-0.03*
GFI	0.12**	*0.1*	*0.09*	*0.08*	*0.03*	*0.01*
SMOG	0.12**	*0.11*	*0.1*	*0.08*	*0.03*	*0.01*

Forum

	FO_C	FO_W	IN_C	IN_W	IM_C	IM_W
FRE	*0.05*	*0.06*	*-0.04*	*-0.01*	*0.0*	*0.03*
ARI	*0.04*	*0.02*	*0.07*	*0.03*	*0.03*	*0.01*
FKR	*0.01*	*-0.01*	0.13**	*0.1*	*0.08*	*0.06*
CLI	*0.04*	*0.02*	*0.06*	*0.04*	*0.02*	*-0.02*
GFI	0.11**	*0.08*	0.22*	0.18*	0.18*	0.17*
SMOG	*-0.03*	*-0.04*	*0.0*	*-0.03*	*0.06*	*0.02*

The observations above are rather encouraging, and allow us to infer that the Mechanical Turk annotation task was successful in general, with many Turkers operating at expert or close to expert levels.

We show the comment statistics in Table 12. The formality question attracted the highest number of comments, presumably because it was the *first* question asked (cf. Section 3.2). Following it, informativeness and implicature questions had the same number of comments. In total, there were 2,262 comments on 2,168 sentences. Most of the commented sentences had only one or two comments per question, and very few had three or more.

In the rest of this section, we will discuss four specific example sentences – first two from the *news* genre, and last two from the *blog* genre – to illustrate our observations. The sentences are as follows:

1. At least seven people were killed and several were missing. (Table 13)
   - *Formality 5.0, Informativeness 5.8, Implicature 4.6.*

2. Although Peter has a shelf of city league soccer trophies, he and the others are mediocre pupils at best, sources said. (Table 14)
   - *Formality 5.6, Informativeness 5.8, Implicature 3.6.*
3. Apple has created a media consumption experience that has reduced friction to such a point that soon the consumer will not know if he is buying music, a movie or a game. (Table 15)
   - *Formality 5.6, Informativeness 6.2, Implicature 4.8.*
4. A $1999 option upgrades the CPU to a "Nehalem" quad core 2.66GHz Core i5 750 processor, or for $200 more, a Core i7 860, both of which include 8MB of L3 cache. (Table 16)
   - *Formality 5.6, Informativeness 6.6, Implicature 4.0.*

Regarding the first example (Table 13), note that the formality comment mentioned it as a *"Statement of fact"*, which is correct, because it is indeed a *descriptive sentence*. The first informativeness comment presents evidence as to why the sentence is informative (*"Tells you how many people and what they were affected by."*). The second informativeness comment goes further (and deeper) into the nuances: *"Several is a vague statement"*, which is seconded by the third informativeness comment: *"I feel it would be very informative if they added context to the missing. How much estimated?"* The first implicature comment reveals *"there was a severe event that occurred"* – a crucial piece of background information, and the second implicature comment goes back to the debate on what *"several"* meant in this particular case. Indeed, the meaning of *"several"* is under-specified, leading us to infer the actual number from background assumptions – which may be wrong.

The second example (Table 14) is also interesting. The first formality comment correctly identified that it *"has some formal language structure, but not totally."* The fourth formality comment went further: *"reads like a newspaper line"*, which is indeed the case. The second formality comment demanded more information, and the third one opined that the sentence *"seems to have an air of"* credibility and formality. The first informativeness comment opined that the sentence did not convey important information, while the second informativeness comment declared that Peter still was a mediocre player, *"in a much wider field of similar athletes"*.

The implicature comments in Table 14 are rather engaging. The third implicature comment infers that *"Peter really enjoys soccer"*, *"doesn't like school"*, and *"he's been playing awhile"*. The second one infers: *"Peter isn't as good outside of the city he plays in."* The first implicature comment mentions that the second part of the sentence (*"mediocre pupils at best"*) does not offer any proof. All these comments show some form of out-of-the-box thinking on the part of the Turkers, which is very encouraging.

In the third example (Table 15), the first formality comment went on to *suggest* best writing practices and how to make the sentence more formal, while the second one correctly predicted the *genre* of such writing: *"reviewing a product on a website."* The first informativeness comment was not very valuable, but the second and third ones maintained that the sentence represented an *"opinion"* or a *"claim"*, without providing *"information"* or *"evidence"*. The first implicature comment went on to infer that Apple was taking over the Electronics world, while the second implicature comment questioned the validity of the *"claim about consumers"*.

The fourth example (Table 16) is on computer hardware, where the first formality comment identifies it as such: *"Most of the words are technical terms."* The second formality comment goes deeper and fishes out subtle aspects: *"This sounds like an attempt to be formal and professional, but comes across a bit muddled and rushed."* The third informativeness comment merely says that the sentence is *"Very detailed"*, but the first informativeness comment mentions that the sentence does require one to *"ask further questions to find out more"*, while the second informativeness comment opines that the sentence tries to *"cram too much information together, without adequate transitions between the offerings."* It further points out that the relative *utility* of the hardware upgrades – *"what makes these upgrades valuable"*, and why we might need them – has not been discussed. The implicature comment lends support to this idea – the *"reasons for consumers wanting/needing the upgrades"* have not been discussed.

What the above discussion shows at a high level, is that the Turkers did understand the annotation task very well, and Mechanical Turk appears to be a suitable platform for conducting pragmatic annotation studies like this.

We went further to analyze if Turkers were *more likely* to comment depending on the *stylistic properties* of certain sentences than others. Note that most of the sentences received only one or two comments per question (if at all), so the *number of comments* was not a very good proxy for Turker commenting activity. Instead, we chose the *average comment length* (averaged across Turkers) for a particular sentence as the measure of commenting activity for that sentence. The average was taken at both word and character levels. We investigated if the average correlated with the eight stylistic variables from Section 4.2. The averages were as follows:

1. **FO_C:** Average length (in characters) of the formality comments received by a sentence.
2. **FO_W:** Average length (in words) of the formality comments received by a sentence.
3. **IN_C:** Average length (in characters) of the informativeness comments received by a sentence.
4. **IN_W:** Average length (in words) of the informativeness comments received by a sentence.
5. **IM_C:** Average length (in characters) of the implicature comments received by a sentence.
6. **IM_W:** Average length (in words) of the implicature comments received by a sentence.

Results are shown in Table 17. Note that most of the correlations are too low, and/or statistically insignificant. This shows that in our corpus, comment length did not vary significantly with any of the stylistic variables we investigated.

Lastly, we checked if the average comment length had any relationship with the *readability* of a sentence. We experimented with six standard readability tests (cf. Lahiri et al. [25]):

1. **FRE:** Flesch Reading Ease

2. **ARI:** Automated Readability Index
3. **FKR:** Flesch-Kincaid Readability Test
4. **CLI:** Coleman-Liau Index
5. **GFI:** Gunning fog Index
6. **SMOG:** Simple Measure of Gobbledygook [35]

Results are shown in Table 18. Once again, most of the correlations are too low, and/or statistically insignificant. The values are even lower than those in Table 17. It shows that comment length did not vary significantly with readability of the sentences.

It is intriguing to observe this negative result, because intuitively one would expect to see some correlation between comment length and at least one of the stylistic properties of a sentence. But the data does not bear this intuition out. Perhaps the commenting process involves a more nuanced and complex judgment, a proper treatment of which is beyond the scope of this paper. We therefore leave the exploration of the possible reasons behind this phenomenon to future work.

6 Conclusion

In this paper, we introduced a dataset of 7,032 sentences rated for formality, informativeness, and implicature on a 1-7 scale by human annotators on Amazon Mechanical Turk. To the best of our knowledge, this is the first large-scale annotation effort that ties together all three pragmatic variables at the sentence level. We measured reliability of our annotations by running two independent rounds of annotation on Mechanical Turk, and inspecting the correlation among mean ratings between the two rounds. We further examined correlation of our annotations with pilot sentence formality annotations done in a more controlled setting [24]. It was observed that while formality and informativeness can be reliably annotated on a 1-7 scale, implicature poses a much more difficult challenge. We analyzed the distribution of formality, informativeness, and implicature across three genres (news, blogs, and forums), and found significant differences – both in terms of overall distribution, and also in terms of the documents' sentential make-up. Correlations between the human ratings and five other stylistic variables were carefully examined. We further examined correlations between six measures of *grammatical complexity*, and the above eight variables. We gave examples of high and low-variance sentences, as well as sentences that do and do not conform to established formality literature. We analyzed the comments Turkers provided as part of the annotation task, to arrive at a set of important insights regarding Turker behavior. We analyzed comment examples, and investigated if the average comment length correlated with any of the eight sentential stylistic variables, or with any of the six readability measures.

Our future plans include an automatic sentence-level formality and informativeness predictor, in the same spirit as Danescu-Niculescu-Mizil et al. [13], and Pavlick and Tetreault [43]. We also plan to investigate the implicature rating more thoroughly, and figure out a good way to improve reliability in implicature annotation.

Our (intentional) lack of stringent control on the Mechanical Turk experiments can potentially be considered a limitation. Note, however, that in rating pragmatic variables, stringent control can do more harm than good. We wanted to observe what people think/feel as formal, informative, and implicative – in a *control-free* environment. We obtained quite good annotations – on both formality as well as informativeness. Implicature posed a more difficult challenge, but we suspect that the challenge would have persisted even after controlling Turker behavior – simply because implicature is inherently more complex. To tackle implicature properly, we would like to combine our approach with Degen's [14] in future work. Future work could also use measures like background questionnaires, linguistic attentiveness surveys, and z-scoring that have been successfully used in previous studies to weed out/smooth Mechanical Turk annotation difficulties [13]. Lastly, we publicly released our annotated corpus as well as the corpus of Turker comments, which we hope will spur further studies in the nascent but growing area of research in automated scoring of formality, informativeness, and implicature.

Acknowledgements We gratefully acknowledge Rada Mihalcea for her support; Mechanical Turk annotators for their annotations; Eduard Hovy and Julian Brooke for valuable discussions; Haiying Li and Nina Dethlefs for inspiration and dataset; Francis Heylighen and Jean-Marc Dewaele for their kindness and brilliant ideas, including the ambiguity score (**I**); and lastly but most importantly, Xiaofei Lu for his continuous encouragement, warm intellectual companionship, excellent advice and camaraderie, sound ideas in the early stages of the study, and great help with thought processing. This work would not have been possible without you. All results, discussions, and comments contained herein are the sole responsibility of the author, and in no way associated with any of the above-mentioned people. The errors and omissions, if any, should be addressed to the author, and will be thankfully received.

References

1. Abu Sheikha, F., Inkpen, D.: Generation of Formal and Informal Sentences. In: Proceedings of the 13th European Workshop on Natural Language Generation, pp. 187–193. Association for Computational Linguistics, Nancy, France (2011). URL http://www.aclweb.org/anthology/W11-2826
2. Abu Sheikha, F., Inkpen, D.: Learning to Classify Documents According to Formal and Informal Style. Submitted to Linguistic Issues in Language Technology (LiLT) (2012)
3. Benotti, L.: Implicature as an Interactive Process. Ph.D. thesis, Université Henri Poincaré - Nancy I (2010). URL https://tel.archives-ouvertes.fr/tel-00541571
4. Benotti, L., Blackburn, P.: Classical planning and causal implicatures. In: M. Beigl, H. Christiansen, T.R. Roth-Berghofer, A. Kofod-Petersen, K.R. Coventry, H.R. Schmidtke (eds.) Modeling and Using Context, *Lecture Notes in Computer Science*, vol. 6967, pp. 26–39. Springer Berlin Heidelberg (2011). DOI 10.1007/978-3-642-24279-3_4. URL http://dx.doi.org/10.1007/978-3-642-24279-3_4
5. Biber, D.: Variation Across Speech and Writing. Cambridge University Press (1988)
6. Biyani, P., Tsioutsiouliklis, K., Blackmer, J.: "8 Amazing Secrets for Getting More Clicks": Detecting Clickbaits in News Streams Using Article Informality. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA., pp. 94–100 (2016). URL http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11807
7. Brooke, J., Hirst, G.: Hybrid Models for Lexical Acquisition of Correlated Styles. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, pp. 82–90. Asian Federation of Natural Language Processing, Nagoya, Japan (2013). URL http://www.aclweb.org/anthology/I13-1010
8. Brooke, J., Hirst, G.: Supervised Ranking of Co-occurrence Profiles for Acquisition of Continuous Lexical Attributes. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 2172–2183. Dublin City University and Association for Computational Linguistics, Dublin, Ireland (2014). URL http://www.aclweb.org/anthology/C14-1205

9. Brooke, J., Wang, T., Hirst, G.: Automatic Acquisition of Lexical Formality. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING '10, pp. 90–98. Association for Computational Linguistics, Stroudsburg, PA, USA (2010). URL http://dl.acm.org/citation.cfm?id=1944566.1944577
10. Carston, R.: Informativeness, Relevance, and Scalar Implicature. In: R. Carston, S. Uchida (eds.) Relevance Theory: Applications and Implications, pp. 179–236. John Benjamins Publishing Co., Amsterdam (1998)
11. Csomai, A., Mihalcea, R.: Linguistically Motivated Features for Enhanced Back-of-the-Book Indexing. In: Proceedings of ACL-08: HLT, pp. 932–940. Association for Computational Linguistics, Columbus, Ohio (2008). URL http://www.aclweb.org/anthology/P/P08/P08-1106
12. Danescu-Niculescu-Mizil, C.: A Computational Approach to Linguistic Style Coordination. Ph.D. thesis, Cornell University (2012)
13. Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., Potts, C.: A computational approach to politeness with application to social factors. In: ACL (1), pp. 250–259. The Association for Computer Linguistics (2013)
14. Degen, J.: Investigating the distribution of *some* (but not *all*) implicatures using corpora and web-based methods. Semantics and Pragmatics (In Press) (2015)
15. Dethlefs, N., Cuayáhuitl, H., Hastie, H., Rieser, V., Lemon, O.: Cluster-based Prediction of User Ratings for Stylistic Surface Realisation. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 702–711. Association for Computational Linguistics, Gothenburg, Sweden (2014). URL http://www.aclweb.org/anthology/E14-1074
16. Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z.: Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, & Computers **36**(2), 193–202 (2004). DOI 10.3758/BF03195564. URL http://dx.doi.org/10.3758/BF03195564
17. Grice, H.P.: Logic and Conversation. In: P. Cole, J.L. Morgan (eds.) Syntax and Semantics: Vol. 3: Speech Acts, pp. 41–58. Academic Press, New York (1975). URL http://www.ucl.ac.uk/ls/studypacks/Grice-Logic.pdf
18. Harnish, R.M.: Logical Form and Implicature. In: T.G. Bever, J.J. Katz, D.T. Langendoen (eds.) An Integrated Theory of Linguistic Ability, pp. 313–392. Thomas Y. Crowell, New York (1976)
19. Heylighen, F., Dewaele, J.M.: Formality of Language: definition, measurement and behavioral determinants. Tech. rep., Center "Leo Apostel", Free University of Brussels (1999)
20. Hovy, E.H.: Pragmatics and Natural Language Generation. Artificial Intelligence **43**(2), 153–197 (1990). DOI 10.1016/0004-3702(90)90084-D. URL http://dx.doi.org/10.1016/0004-3702(90)90084-D
21. Hudson, R.: About 37% of Word-Tokens are Nouns. Language **70**(2), pp. 331–339 (1994). URL http://www.jstor.org/stable/415831
22. Kireyev, K.: Semantic-based Estimation of Term Informativeness. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 530–538. Association for Computational Linguistics, Boulder, Colorado (2009). URL http://www.aclweb.org/anthology/N/N09/N09-1060
23. Lahiri, S.: SQUINKY! A Corpus of Sentence-level Formality, Informativeness, and Implicature. CoRR **abs/1506.02306** (2015). URL http://arxiv.org/abs/1506.02306
24. Lahiri, S., Lu, X.: Inter-rater Agreement on Sentence Formality. CoRR **abs/1109.0069** (2011). URL http://arxiv.org/abs/1109.0069
25. Lahiri, S., Mitra, P., Lu, X.: Informality Judgment at Sentence Level and Experiments with Formality Score. In: Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II, CICLing'11, pp. 446–457. Springer-Verlag, Berlin, Heidelberg (2011). URL http://dl.acm.org/citation.cfm?id=1964750.1964792
26. Landauer, T.K., Dumais, S.T.: Latent semantic analysis. Scholarpedia **3**(11), 4356 (2008). URL http://www.scholarpedia.org/article/Latent_semantic_analysis
27. Leckie-Tarry, H., Birch, D.: Language and Context: A Functional Linguistic Theory of Register. Pinter Publishers (1995). URL http://books.google.com/books?id=-qdrAAAAIAAJ
28. Levelt, W.J.M.: Speaking: From Intention to Articulation. MIT Press, Cambridge, MA (1989)
29. Levin, N.S., Prince, E.F.: Gapping and Causal Implicature. Paper in Linguistics **19**(3), 351–364 (1986). DOI 10.1080/08351818609389263
30. Li, H., Cai, Z., Graesser, A.C.: Comparing Two Measures for Formality. In: Proceedings of the Florida Artificial Intelligence Research Society Conference (2013). URL https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS13/paper/view/5899

31. Likert, R.: A Technique for the Measurement of Attitudes. Archives of Psychology **22**(140), 1–55 (1932)
32. Machili, I.: Writing in the workplace: Variation in the Writing Practices and Formality of Eight Multinational Companies in Greece. Ph.D. thesis, University of the West of England (2014)
33. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60. Association for Computational Linguistics, Baltimore, Maryland (2014). URL http://www.aclweb.org/anthology/P14-5010
34. de Marneffe, M.C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., Manning, C.D.: Universal stanford dependencies: A cross-linguistic typology. In: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 4585–4592. European Language Resources Association (ELRA), Reykjavik, Iceland (2014). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/1062_Paper.pdf. ACL Anthology Identifier: L14-1045
35. McLaughlin, H.G.: SMOG Grading – a New Readability Formula. Journal of Reading pp. 639–646 (1969)
36. Miller, G.A.: WordNet: A Lexical Database for English. Commun. ACM **38**(11), 39–41 (1995). DOI 10.1145/219717.219748. URL http://doi.acm.org/10.1145/219717.219748
37. Molina, A., Torres-Moreno, J.M., SanJuan, E., da Cunha, I., Sierra Martínez, G.E.: Discursive Sentence Compression. In: A. Gelbukh (ed.) Computational Linguistics and Intelligent Text Processing, *Lecture Notes in Computer Science*, vol. 7817, pp. 394–407. Springer Berlin Heidelberg (2013). DOI 10.1007/978-3-642-37256-8_33. URL http://dx.doi.org/10.1007/978-3-642-37256-8_33
38. Nicholas, L.E., Brookshire, R.H.: A System for Quantifying the Informativeness and Efficiency of the Connected Speech of Adults with Aphasia. Journal of Speech and Hearing Research **36**(2), 338–350 (1993)
39. Nishikawa, H., Hasegawa, T., Matsuo, Y., Kikui, G.: Optimizing Informativeness and Readability for Sentiment Summarization. In: Proceedings of the ACL 2010 Conference Short Papers, ACLShort '10, pp. 325–330. Association for Computational Linguistics, Stroudsburg, PA, USA (2010). URL http://dl.acm.org/citation.cfm?id=1858842.1858902
40. Nowson, S., Oberlander, J., Gill, A.J.: Weblogs, Genres, and Individual Differences. In: Proceedings of the 27th Annual Conference of the Cognitive Science Society, pp. 1666–1671 (2005)
41. Papafragou, A., Musolino, J.: Scalar implicatures: experiments at the semantics-pragmatics interface. Cognition **86**(3), 253 – 282 (2003). DOI http://dx.doi.org/10.1016/S0010-0277(02)00179-8. URL http://www.sciencedirect.com/science/article/pii/S0010027702001798
42. Pavlick, E., Nenkova, A.: 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, pp. 218–224. Association for Computational Linguistics, Denver, Colorado (2015). URL http://www.aclweb.org/anthology/N15-1023
43. Pavlick, E., Tetreault, J.: An Empirical Analysis of Formality in Online Communication. Transactions of the Association for Computational Linguistics **4**, 61–74 (2016). URL https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/732
44. Peterson, K., Hohensee, M., Xia, F.: Email Formality in the Workplace: A Case Study on the Enron Corpus. In: Proceedings of the Workshop on Languages in Social Media, LSM '11, pp. 86–95. Association for Computational Linguistics, Stroudsburg, PA, USA (2011). URL http://dl.acm.org/citation.cfm?id=2021109.2021120
45. Phan, X.H.: CRFTagger: CRF English POS Tagger (2006). URL http://crftagger.sourceforge.net/
46. Potts, C.: The Logic of Conventional Implicatures. Oxford Studies in Theoretical Linguistics. Oxford University Press, Oxford (2005)
47. Potts, C.: Conventional implicature and expressive content. In: C. Maienborn, K. von Heusinger, P. Portner (eds.) Semantics: An International Handbook of Natural Language Meaning, vol. 3, pp. 2516–2536. Mouton de Gruyter, Berlin (2012). This article was written in 2008
48. Potts, C.: Conversational implicature: interacting with grammar (2013). Paper presented at the University of Michigan 2013 Workshop in Philosophy and Linguistics

49. Rajman, M., Hartley, T.: Automatically predicting MT systems rankings compatible with Fluency, Adequacy or Informativeness scores. In: Procs. 4th ISLE Workshop on MT Evaluation, MT Summit VIII, pp. 29–34 (2001)
50. Rennie, J.D.M., Jaakkola, T.: Using Term Informativeness for Named Entity Detection. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05, pp. 353–360. ACM, New York, NY, USA (2005). DOI 10.1145/1076034.1076095. URL http://doi.acm.org/10.1145/1076034.1076095
51. Tatu, M., Moldovan, D.: A Tool for Extracting Conversational Implicatures. In: N.C.C. Chair), K. Choukri, T. Declerck, M.U. Doan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (2012)
52. Teddiman, L.: Contextuality and Beyond: Investigating an Online Diary Corpus. In: E. Adar, M. Hurst, T. Finin, N.S. Glance, N. Nicolov, B.L. Tseng (eds.) ICWSM. The AAAI Press (2009). URL http://dblp.uni-trier.de/db/conf/icwsm/icwsm2009.html#Teddiman09
53. Timonen, M., Toivanen, T., Teng, Y., Cheng, C., He, L.: Informativeness-based Keyword Extraction from Short Documents. In: A.L.N. Fred, J. Filipe, A.L.N. Fred, J. Filipe (eds.) KDIR, pp. 411–421. SciTePress (2012). URL http://dblp.uni-trier.de/db/conf/ic3k/kdir2012.html#TimonenTTCH12
54. Ure, J.: Lexical density and register differentiation. Applications of Linguistics pp. 443–452 (1971)
55. Vogel, A., Potts, C., Jurafsky, D.: Implicatures and Nested Beliefs in Approximate Decentralized-POMDPs. In: Proceedings of the 2013 Annual Conference of the Association for Computational Linguistics, pp. 74–80. Association for Computational Linguistics, Stroudsburg, PA (2013)
56. Wu, Z., Giles, C.L.: Measuring Term Informativeness in Context. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 259–269. Association for Computational Linguistics, Atlanta, Georgia (2013). URL http://www.aclweb.org/anthology/N13-1026