Informality Judgment at Sentence Level and Experiments with Formality Score

Shibamouli Lahiri, Prasenjit Mitra, and Xiaofei Lu

The Pennsylvania State University, University Park PA 16802, USA
shibamouli@cse.psu.edu, pmitra@ist.psu.edu, xxl13@psu.edu

Abstract. Formality and its converse, informality, are important dimensions of authorial style that serve to determine the social background a particular document is coming from, and the potential audience it is targeted to. In this paper we explored the concept of formality at the sentence level from two different perspectives. One was the Formality Score (F-score) and its distribution across different datasets, how they compared with each other and how F-score could be linked to human-annotated sentences. The other was to measure the inherent agreement between two independent judges on a sentence annotation task. It gave us an idea how subjective the concept of formality was at the sentence level. Finally, we looked into the related issue of document readability and measured its correlation with document formality.

1 Introduction

Writing style is an important dimension of human languages. Two documents can provide the same content, but they may have been written using very different styles [9]. Authors from different social, educational and cultural backgrounds tend to use different writing styles [4]. With the evolution of Web 2.0, user-generated content has given rise to a variety of writing styles. Blog posts, for example, are written differently from the way academic papers are written. Twitter chats manifest yet another kind of writing style. Wikipedia articles use their own style guide[1].

One prominent dimension of writing style is the formality of a document. Academic papers are usually considered more formal than online forum posts. The notions of formality and contextuality at the document level have been illustrated by Heylighen and Dewaele [7]. They proposed a frequentist statistic known as the Formality Score (F-score) of a document, based on the number of deictic and non-deictic words (cf. Section 2). F-score is a coarse-grain measure, but it works well when used to classify documents according to their authorial style [15].

Classifying sub-document units such as sentences as formal or informal is more difficult because they are typically much smaller than a document and provide much less information. For example, the sentence “She doesn’t like the piano” may be considered informal because it contains the colloquial usage “doesn’t”. But some native English speakers may think that the usage of “doesn’t” is quite appropriate and formal. So we note that the notion of formality at the sentence level is subjective. On the other

[1] http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style

A. Gelbukh (Ed.): CICLing 2011, Part II, LNCS 6609, pp. 446–457, 2011.

hand, the sentence “She does not like the piano” is more formal than the sentence “She doesn’t like the piano”. So instead of classifying a sentence as formal or informal, we might actually be better off by assigning a formality score to a sentence, which would then reflect its *degree of formality*. A question that immediately arises is whether we can use the F-score of a sentence for this purpose.

As pointed out in [7], a frequentist statistic such as F-score should not be applied directly to measure the formality of a small text sample, e.g., a sentence. In this paper we look into the F-score distribution at the sentence level for four independent corpora and observe that these distributions broadly follow the corpus-level F-score trend. Moreover, the sentence-level F-score distribution on a human-annotated dataset shows a clear distinction between sentences labeled formal and sentences labeled informal. These observations indicate that the sentence-level F-score may be used as a feature in designing a formality score for sentences.

The second experiment reported in this paper is an inter-annotator agreement study for constructing a gold-standard dataset for the binary sentence classification task. Two independent annotators, both native speakers of English, judged sentences as formal or informal according to their own perception and intuition. Annotation judgments on two different datasets show poor agreement. We reason that this negative result is because of the arbitrariness of the notion of informality in two different judges’ minds. A take-home message from this study is to either carefully design an annotation guideline or to adopt a Likert-style labeling scheme instead of a binary one, and let the judges discuss their results among themselves to improve agreement.

Apart from the formality of a document, we also consider the related issue of its readability. Traditional readability tests like the Flesch Reading Ease Score measure how difficult it is to read a piece of text. As a document becomes more formal, it starts introducing more context (cf. Section 2). So the document usually becomes longer, with more intricate sentence structure. Intuition suggests that such context insertion would typically mean a corresponding increase in reading time, i.e., reading difficulty. Document-level correlation between F-score and readability tests justifies our intuition. We found moderate correlation in all cases.

This paper is organized as follows. Section 2 discusses the background on F-score. Section 3 describes our experiments. Section 4 gives related work and Section 5 concludes the paper. The complementary code and data are available at
http://www.CICLing.org/2011/software/251.

2 Background

The seminal study on measuring text formality by Heylighen and Dewaele [7] considers two different variants of formal expressions - *surface formality* and *deep formality*. Surface formality is the case when language is formalized for its own sake, e.g., a marital vow. Deep formality on the other hand represents the case when language is formalized so that the meaning is communicated clearly and as completely as possible. Complete communication of meaning involves putting in more background information so that no question regarding a document may go unanswered. This background information is known as “context”. So we observe that as more context is inserted into a document, the language tends to become more (deeply) formal. Conversely, as a document is gradually robbed of its context, the language tends to become more *contextual*. Heylighen and Dewaele also argued that surface formality emerges from deep formality, so the latter is sufficient to characterize both.

As an example of deep formality, consider the sentence “She likes the piano”. This sentence can be made more formal by saying “Ms Muffet likes the piano”. Here “Ms Muffet” is a part of the context of the first sentence. However, we can make the sentence even more (deeply) formal by saying “Ms Muffet likes the piano beside the door”. Note that in the last sentence we added more context than there was in the second sentence. This context-addition and resulting formalization process can be continued ad infinitum, because it is impossible to fully specify the meaning of a text in itself without some unsaid background assumptions. Since context-addition is always possible, we cannot make a hard judgment that one document is strictly formal and another one is strictly informal. We can say that document A is more formal than document B. This is known as the *continuum of formality*.

Informality is introduced by deixis and implicature. Deixis indicates a set of words that anchors to another set of words for contextual information [11]. For example, in the sentence “She likes the piano”, the word “she” anchors to “Ms Muffet”. Four types of deixis have been recognized - time, place, person and discourse [11]. Time deixis can be seen in the words “today”, “now”, “then”, etc. These words anchor to specific time points. Place deixis is exemplified in the place-anchoring words “here”, “there”, “around”, etc; person (or object) deixis gives us words like “this”, “that”, “he”, “she”, etc; and discourse deixis engenders words like “therefore”, “hence”, “notwithstanding”, etc. Detailed word correlation studies indicate some categories of words are *deictic* (pronouns, verbs, adverbs, interjections), some others are *non-deictic* (nouns, adjectives, prepositions, articles) and the rest are *deixis-neutral* (conjunctions) [7].

In deixis, there are some anchor words that explicitly relate to the context information. In implicature, the context information must be inferred from background knowledge. As an example, consider the sentence “Einstein rocks!” In this sentence the context information - why Einstein rocks - is absent. Only when we couple this sentence with the background knowledge that Einstein was a great scientist, do we come to appreciate the full meaning. But quantifying the impact of implicature is more difficult because we need to call upon the background information - something which is not present in the document. Therefore only deictic and non-deictic words were considered in the definition of F-score:

F = (noun frequency + adjective freq. + preposition freq. + article freq. - pronoun freq. - verb freq. - adverb freq. - interjection freq. + 100)/2

where the frequencies are taken as percentages with respect to the total number of words in the document [7]. Note that as the number of deictic words increases and non-deictic words decreases, F-score becomes lower, indicating a more contextual (informal) document. The reverse happens in the case of a more formal document. F-score of a document can range from zero to 100.

Note that the definition of F-score is valid for sentences as well. But sentences are much smaller than documents, so we cannot directly use F-score for measuring sentence-level formality. However, we would like to observe if F-score can be used as a feature for designing a sentence-level formality score. To address this question, we look into the sentence-level F-score distributions on unlabeled as well as labeled corpora. In the next section we describe the results of our exploratory analysis.

3 Experiments

3.1 Datasets

We compiled four different datasets - blog posts, news articles, academic papers and online forum threads. Each dataset has 100 documents. For the blog dataset, we collected most recent posts from the top 100 blogs listed by Technorati[2] on October 31, 2009. For the news article dataset, we collected 100 news articles from 20 news sites (five from each). These articles are mostly from "Breaking News", "Recent News" and "Local News" categories, with no specific preference to any of the categories. The news sites we used 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. For the academic paper dataset, we randomly sampled 100 papers from the CiteSeerX[3] digital library. For the online forum dataset, we sampled 50 random documents crawled from Ubuntu Forums[4] and 50 random documents crawled from TripAdvisor New York forum[5]. The blog, news, paper and forum datasets have 2110, 3009, 161406 and 2569 sentences respectively. The overall F-scores of these datasets are 65.24, 66.51, 68.62 and 58.52 respectively[6].

3.2 Sentence Level F-score Distributions

We recall from Section 2 that the F-score of a document uses deixis information as a measure of formality. Since a sentence can be thought of as a small document, deixis is present at the sentence level as well. It is therefore of interest to explore how the sentence-level F-score distributions compare with each other, and whether they bear any consistent form across different datasets. Apart from shedding light on the variation of sentence-level deixis and its types in various corpora, such an exploratory analysis would also allow us to observe if the sentence-level F-score distributions follow a specific trend. Figure 1(a) gives us the histogram of sentence-level distributions on four datasets (cf. Section 3.1) and Table 1 outlines some of their key properties. We note from

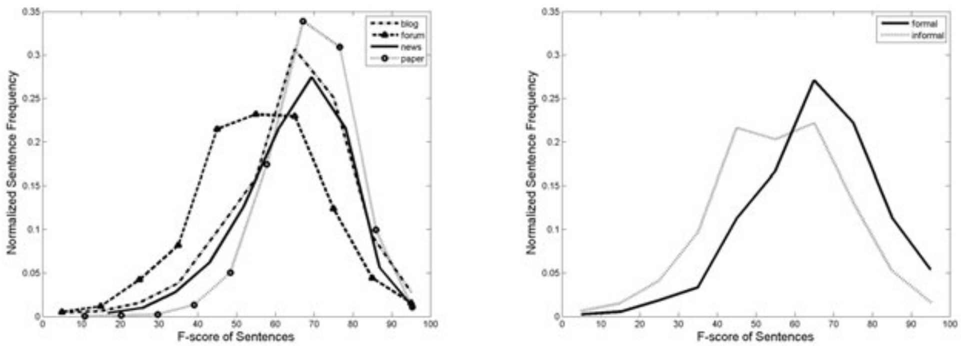
[2] http://www.technorati.com

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

[4] http://ubuntuforums.org/

[5] http://www.tripadvisor.com/ShowForum-g28953-i4-New York.html

[6] F-score computation involves part-of-speech tagging. We used CRFTagger [16] in all our experiments.



(a) Unlabeled data

(b) Labeled data

Fig. 1. Histogram of sentence-level F-score distributions on different datasets

Figure 1(a) that the sentence-level F-score distribution of a higher-F-scored dataset is shifted towards the high formality zone and the sentence-level F-score distribution of a lower-F-scored dataset is shifted towards the low formality zone. Moreover, as the corpus-level F-score increases more and more, the sentence-level F-score distributions shift more and more to the higher formality zone.

Table 1. Properties of Sentence-level F-score Distributions

Dataset	Mean	SD	Median	QD	Skewness	Kurtosis
Forum	56.74	15.82	57.14	9.58	-0.12	3.37
Blog	65.02	15.01	66.67	9.38	-0.64	4.27
News	65.18	13.34	66.67	8.93	-0.59	3.57
Paper	69.29	10.44	70	6.70	-0.53	3.96

Table 2. Multiple Comparison Test between all groups with Tukey-Kramer's HSD correction

Group 1-Group 2	F^{Group}_{mean} F^{Group2}_{mean}	Confidence Interval	Conclusion
Blog-Forum	8.28	[7.48, 9.09]	$F^{blog}_{mean} > F^{forum}_{mean}$
Blog-News	-0.16	[-0.93, 0.62]	$NOT(F^{news}_{mean} > F^{blog}_{mean})$
Blog-Paper	-4.27	[-4.87, -3.67]	$F^{paper}_{mean} > F^{blog}_{mean}$
Forum-News	-8.44	[-9.18, -7.70]	$F^{news}_{mean} > F^{forum}_{mean}$
Forum-Paper	-12.55	[-13.10, -12.01]	$F^{paper}_{mean} > F^{forum}_{mean}$
News-Paper	-4.11	[-4.62, -3.61]	$F^{paper}_{mean} > F^{news}_{mean}$

Table 1 gives the Mean, Median, Standard Deviation (SD), Quartile Deviation (QD), Skewness and Kurtosis of the distributions. Note that the standard and quartile deviations for paper sentences are the smallest, so these sentences vary least in terms of F-score, while those from the forum dataset vary the most. One possible reason for such a high variation in forum sentences (along with low kurtosis) is that they come from

Table 3. Confidence Intervals obtained using different multiple comparison tests

Group 1	Group 2	Confidence Intervals			
		LSD	**Bonferroni**	**Dunn-Šidák**	**Scheffé**
Blog	Forum	[7.67, 8.90]	[7.46, 9.11]	[7.46, 9.11]	[7.41, 9.16]
Blog	News	[-0.75, 0.44]	[-0.96, 0.64]	[-0.95, 0.64]	[-1.00, 0.69]
Blog	Paper	[-4.73, -3.81]	[-4.89, -3.65]	[-4.88, -3.65]	[-4.92, -3.62]
Forum	News	[-9.00, -7.88]	[-9.20, -7.68]	[-9.20, -7.69]	[-9.24, -7.64]
Forum	Paper	[-12.97, -12.14]	[-13.11, -11.99]	[-13.11, -11.996]	[-13.15, -11.96]
News	Paper	[-4.50, -3.73]	[-4.63, -3.59]	[-4.63, -3.60]	[-4.66, -3.56]

different types of users - some are information seekers, typically issuing sentences with less context (lower F-score), while others are information providers, issuing sentences with more context (higher F-score). On the other hand the paper sentences are somewhat "homogenized" and "compressed" into the higher end of formality continuum, because they all tend to follow the strict norms of written English.

To test whether these distributions are significantly different from each other, we performed a two-sample Kolmogorov-Smirnov test on each pair of distributions. At significance level $\alpha = 0.001$, all pairs (except the blog-news pair) were found to be significantly different from each other. Similar results were obtained in the pairwise comparison between distribution means (F_{mean}). We first performed a one-way ANOVA[7] on the null hypothesis:

$$F^{paper}_{mean} = F^{news}_{mean} = F^{blog}_{mean} = F^{forum}_{mean}$$

where F^{i}_{mean} denotes the mean sentence-level F-score of dataset i. The ANOVA results reject this null hypothesis at significance level $\alpha = 0.001$, which indicates that at least two of the group means are significantly different from each other. Pairwise comparison between the group means were performed next with multiple testing correction. The results are shown in Tables 2 and 3. Each pairwise test is equivalent to an unpaired two-sample one-tailed t-test for comparing the means of two groups, with the addition of correction and adjustments for multiple comparison problem. Table 2 has six rows. Each row gives the groups of one pair, the difference in F_{mean}s between the two groups and the confidence interval of this difference using Tukey-Kramer's HSD correction. Note that if this confidence interval contains zero, then we conclude that the group means are not significantly different from each other. Otherwise, the sign of the group mean difference indicates whether group 1 has larger F_{mean} than group 2, or vice versa.

In Table 3, we report the confidence intervals obtained using other multiple comparison tests, e.g., Fisher's least significant difference (LSD) method, Bonferroni's method, Dunn-Šidák's method and Scheffé's method, respectively. The confidence intervals follow the same trend as in Table 2, and they lead to the same conclusions - all group means are significantly different from each other (except the Blog-News pair) and the group means satisfy

[7] We used MATLAB for all our significance tests.

1. $F_{mean}^{paper} > F_{mean}^{news}$
2. $F_{mean}^{news} > F_{mean}^{forum}$
3. $F_{mean}^{paper} > F_{mean}^{blog}$
4. $F_{mean}^{blog} > F_{mean}^{forum}$

The reason why F_{mean}^{news} was not significantly different from F_{mean}^{blog} is that the blog posts were collected from the top 100 list of Technorati. Since blog is a bridging genre [6], many blog posts may actually be modified news articles. It is especially true with a generic blog search engine like Technorati, which indexes all kinds of blogs. This is also the reason why sentence-level F-score distribution for blogs was not found to be significantly different from that for news articles in the Kolmogorov-Smirnov Test. Note that the sentence-level F-score distribution for two very similar corpora may not be significantly different from each other. For example, if we modify a large dataset by introducing a few non-deictic words here and there, then the overall F-score will slightly increase, but the sentence-level F-score distribution will remain virtually the same.

3.3 Sentence Level F-score on Annotated Data

The results of Section 3.2 indicate that unless two corpora are very similar in their deixis content, their sentence-level F-score distributions will be different. But this observation in itself is not sufficient for declaring F-score as a sentence-level feature. We would also need to link F-score with the human notion of formality at the sentence level. In this experiment we labeled a 50-document dataset (7488 sentences) from the Splog Blog Collection[8]. A graduate student labeled each sentence as formal or otherwise according to whether or not the sentence contains informal/slang words and expressions, grammatical inconsistencies, visual cues like smileys and character repetition, etc. This student was not given any background on F-score at the time of the annotation, thereby eliminating bias. Among 7488 sentences, 4185 were labeled formal, and 3303 were labeled informal.

Table 4. Properties of Sentence-level F-score Distributions - Labeled Data

Dataset	Mean	SD	Median	QD	Skewness	Kurtosis
Formal	65.82	15.85	66.67	9.92	-0.31	3.55
Informal	56.65	16.58	57.14	10.99	-0.19	3.18

The sentence-level F-score histogram of these sentences is shown in Figure 1(b) and the distribution properties are given in Table 4. Figure 1(b) and Table 4 show that the two distributions are different from each other with formal sentences shifted towards relatively higher F-score zones and informal sentences shifted towards relatively lower F-score zones. This is an important finding, because it indicates that the human-labeled sentences form a clear split in terms of F-score distribution.

[8] http://ebiquity.umbc.edu/resource/html/id/212/Splog-Blog-Dataset

A two-sample Kolmogorov-Smirnov test showed that at significance level $\alpha = 0.001$, the two distributions were different from each other. A one-tailed two-sample unpaired t-test for comparing the group means led to the same conclusion, where the confidence interval of the group mean difference was found to be [7.93,10.41]. Note that this interval does not contain zero, so the two group means are significantly different. This observation allows us to reason that F-score can be useful as one of the sentence-level features for capturing formality.

3.4 Inter-annotator Agreement Study

Designing a sentence-level formality score is complicated by the fact that different people have different notions regarding what should be considered formal or what should be considered informal. The concept of formality as native speakers perceive, is fairly subjective. It is therefore of importance to measure by how much two independent judges differ on a set of sentences, when no specific instructions are given as to what constitutes a formal or an informal sentence. If this "inherent" agreement is high, then we are able to establish a reliable gold-standard sentence-annotated dataset. If on the other hand this agreement is low, then we get an idea of how subjective the idea of sentence-level formality really is. In that case we can either employ a set of annotation instructions for improving agreement, or we can change the labeling scheme and let the annotators discuss among themselves to minimize disagreement. Note that the issue of constructing a gold-standard sentence-annotated dataset assumes importance because a sentence-level formality score can only be evaluated on such a hand-crafted corpus.

Table 5. Confusion Matrix and Inter-annotator Agreement

Blog Posts				
	C	NC	Raw Agreement	0.692
C	168	172	Kappa	0.164
NC	480	1300	Jaccard	0.205
News Articles				
	C	NC	Raw Agreement	0.756
C	71	383	Kappa	0.019
NC	352	2204	Jaccard	0.088

In this section we describe the results of an inter-annotator agreement study aimed at measuring the inherent agreement between two native English speakers regarding the concept of sentence-level formality. We enlisted help from four undergraduate students, who independently labeled each sentence of the blog and news datasets (cf. Section 3.1) as formal or informal.[9] Two students worked on the blog dataset and the other two worked on the news dataset. Students were requested to mark each sentence they considered informal as "C" and each sentence they considered formal as "NC" (Table 5).

[9] Students were remunerated with extra course credit at the end of the annotation.

They were not allowed to discuss among themselves or see each other's annotations. Since the purpose of this study was to measure "inherent" agreement between two native speakers of English, we did not specify what constitutes a formal sentence or an informal sentence. In other words, we did not have an annotation guideline or a rubric.

After the annotation process was over, we computed Cohen's Kappa and Jaccard Similarity along with raw agreement scores based on the confusion matrices (Table 5). Jaccard Similarity was computed as:

$$Jaccard = \frac{\#CC}{\#CC + \#CNC + \#NCC}$$

where $\#CC$, $\#CNC$ and $\#NCC$ denote the number of sentences in the top left, top right and bottom left cells of the confusion matrix, respectively. The agreement results are shown in Table 5. The raw agreement values are moderately high, but both Cohen's Kappa and Jaccard Coefficient indicate poor agreement. The reason behind this apparent paradox lies in the fact that the number of NCNC sentences - sentences both annotators considered formal, is very high (Table 5, NC row and NC column).

These findings imply a negative result in terms of inherent agreement at the sentence level regarding the notion of informality. The very low Kappa values obtained across two independent datasets show that there is hardly any agreement. This stance is bolstered by equally low values of Jaccard Coefficient obtained in both cases. So, coming up with a reliable gold-standard set of annotated sentences without some annotation guidelines is difficult. One way to improve agreement is to do several rounds of annotation and let the judges discuss after each round to converge into a common labeling scheme [2]. But this procedure as observed in [2], improves agreement only marginally, and that also when the initial agreement is already quite high. Another way to improve agreement is to design a detailed annotation guideline. However, design of such a guideline may entail loss of generalizability across multiple datasets and bias the study somewhat from the experimenter's perspective, so this approach needs to be carefully investigated before being put into effect.

The take-home message from this experiment is clear: formal/informal-type gold-standard sentence set construction will prove to be difficult because of the poor inter-annotator agreement. The poor agreement is not also very unexpected, because as we discussed in Section 2, the formality continuum is present at the sentence level as well. The binary annotation process forces the judges to do an arbitrary thresholding in this continuum and declare sentences "formal" when they are above this threshold and "informal" when they are below. This thresholding can be very different for two different persons and thereby yield poor agreement values. An alternative is to adopt a Likert-style labeling scheme [12], where instead of labeling sentences as formal/informal, judges provide a formality rating. Our future work includes working on this alternative. We also plan to let judges discuss among themselves for minimizing disagreement and coming up with a consistent set of annotation guidelines across multiple datasets.

3.5 F-score and Readability

An important observation with F-score is that it captures *deep formality* (cf. Section 2). As we go on adding context to a document, its deep formality increases. However,

Table 6. Overall F-score and Readability on different datasets

Dataset	F-score	FRES	ARI	FKRT	CLI	GFI	SMOG
Forum	58.52	77.71	7.90	6.05	9.43	9.83	9.38
Blog	65.24	61.04	11.63	9.47	11.43	13.83	12.03
News	66.51	56.21	13.13	10.78	12.47	15.50	13.46
Academic Paper	68.62	48.41	15.86	12.62	14.20	18.00	15.15

Table 7. Correlation of F-score with Readability measures

Readability Measure	Pearson's ρ	Spearman's ρ	Kendall's τ	Quadrant Correlation
ARI	0.45	0.57	0.41	0.48
CLI	0.46	0.61	0.44	0.52
FKRT	0.49	0.60	0.43	0.48
FRES	-0.50	-0.64	-0.46	-0.54
GFI	0.53	0.61	0.44	0.53
SMOG	0.54	0.62	0.46	0.55

adding context usually involves introducing new words, which increases the length of the document. Although in certain cases new words replace old words, so the document length remains unchanged, we expect that as more and more context information is added, a document tends to become longer. Longer documents take more time to process than shorter ones, so we expect that the overall *reading difficulty* of a document starts increasing as we go on adding more and more context. In other words, as the deep formality of a document increases, its reading difficulty also increases. Since the reading difficulty of a document is measured by *readability tests* and deep formality by F-score, we expect that there should be some correlation between F-score and readability tests.

To test the presence of such a correlation, we measured corpus-level F-score and readability scores on four datasets (cf. Section 3.1). Six standard readability tests were performed. These are Flesch Reading Ease Score (FRES), Automated Readability Index (ARI), Flesch-Kincaid Readability Test (FKRT), Coleman-Liau Index (CLI), Gunning fog Index (GFI) and SMOG (Simple Measure of Gobbledygook) [14]. Results are shown in Table 6, which indicates a clear trend in F-score and readability tests. All the readability tests (except FRES) show positive correlation with F-score. Pearson and rank correlation tests between document-level F-score and readability scores (Table 7) show moderate correlation values[10] in all cases. Pearson, Spearman and Kendall correlation values were found to be statistically highly significant with p-value < 0.0001. The negative correlation with FRES can be explained by the fact that FRES actually measures "reading ease" as opposed to "reading difficulty". This result justifies our intuition that context addition (F-score) and reading difficulty (readability tests) are correlated, but since the correlation is not very high, we believe there are factors other than readability that get into play when more context is inserted into a document, so the reading difficulty does not increase as much. This point merits further investigation.

[10] http://pathwayscourses.samhsa.gov/eval201/eval201_4_pg9.htm

4 Related Work

In this section we give a very brief sketch of the related studies. The presence of formality as a prominent dimension of language variation was first noted by Biber [1]. Formality of a language is largely determined by four factors - time, place, context and person. The four factors have been arrived at in the study of *registers* in sociolinguistics [17]. Registers denote a form of language variation that occurs both as a result of difference in speaker identity and as a result of difference in situation (context) [5]. Zampolli [19] and Hudson [8] arrived at the dimension of formality based on their own style analyses, but they could not explain it theoretically. Heylighen and Dewaele [7] were the first to summarily assess the causes of formality and design a document-level formality score, called the F-score. F-score uses the idea of *context* [10] and is much in the same spirit as the lexical density [18]. While F-score has not yet been applied to the sub-document level, a recent study by Brooke, et al. [3] looks into the notion of formality at the word level. They used publicly available formal and informal word lists as seed sets and analyzed large corpora to evaluate the effectiveness of several different approaches for measuring word-level formality. While our goal is different in the sense that we want to measure sentence-level formality, we can still use the word-level scores as features. The sentences are somewhat more difficult to deal with, because we cannot have a seed set of sentences without human annotation. Some of the results reported in this paper constitute the first step towards the creation of such a gold standard.

5 Conclusion

We have four principal contributions in this paper:

1. Exploratory analysis and comparison of sentence-level F-score distributions of four different datasets
2. Linking F-score with the human perception of sentence-level formality using F-score distribution on an annotated dataset
3. An inter-annotator agreement study to measure the inherent agreement between two independent native speakers of English on the notion of sentence-level formality
4. Correlation between F-score and readability tests

Our future work includes the design of a sentence-level formality score. Such a score would require, among other things, syntactic, semantic and pragmatic considerations [7]. Even more challenging is the problem of formality assessment at sub-sentence level. While there has been work on local emotion detection [13], it remains open whether similar techniques can be exploited in sub-sentence level formality judgment.

References

1. Biber, D.: Variation Across Speech and Writing. Cambridge University Press, Cambridge (1988)
2. Brants, T., Skut, W., Uszkoreit, H.: Syntactic annotation of a German newspaper corpus. In: Proceedings of the ATALA Treebank Workshop, Paris, France, pp. 69–76 (1999)

3. Brooke, J., Wang, T., Hirst, G.: Automatic acquisition of lexical formality. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING) (2010)
4. Chambers, J.K., Schilling-Estes, N., Trudgill, P.: The handbook of language variation and change. Blackwell, Malden (2006)
5. Halliday, M.: Comparison and translation. In: Halliday, M., McIntosh, M., Strevens, P. (eds.) The linguistic sciences and language teaching. Longman, Harlow (1964)
6. Herring, S.C., Scheidt, L.A., Wright, E., Bonus, S.: Weblogs as a bridging genre. IT & People 18(2), 142–171 (2005)
7. Heylighen, F., Marc Dewaele, J.: Formality of language: definition, measurement and behavioral determinants. Tech. rep. (1999)
8. Hudson, R.: About 37% of word-tokens are nouns. Language 70(2), 331–339 (1994)
9. Karlgren, J.: Stylistic experiments for information retrieval (2000)
10. Leckie-Tarry, H., Birch, D.: Language and context: a functional linguistic theory of register. In: Birch, D. (ed.) Pinter Publishers, London (1995)
11. Levelt, W.J.M.: Speaking: From Intention to Articulation. MIT Press, Cambridge (1989)
12. Likert, R.: A technique for the measurement of attitudes. Archives of Psychology 22(140), 1–55 (1932)
13. Mao, Y., Lebanon, G.: Isotonic Conditional Random Fields and Local Sentiment Flow. In: Advances in Neural Information Processing Systems (2007)
14. McLaughlin, H.G.: SMOG grading - a new readability formula. Journal of Reading, 639–646 (May 1969)
15. 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)
16. Phan, X.H.: CRFTagger: CRF English POS Tagger (2006), `http://crftagger.sourceforge.net/`
17. Reid, T.B.: Linguistics, structuralism, philology. Archivum Linguisticum 8
18. Ure, J.N.: Lexical density and register differentiation. In: Perren, G.E., Trim, J.L.M. (eds.) Applications of Linguistics: Selected Papers of the 2nd International Congress of Linguistics, Cambridge 1969. Cambridge University Press, Cambridge (1971)
19. Zampolli, A.: Statistique linguistique et dépouillements automatiques. Lexicologie, 325–358