Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference

A Measure of Text Formality as a Human Construct

Haiying Li, Qinyu Cheng, & Arthur C. Graesser
202 Psychology Building; University of Memphis; Memphis, TN, USA, 38152
{hli5; qcheng; graesser}@memphis.edu

Abstract

Formality has long been of interest in the study of language and discourse, and different measures of formality have been developed to predict genre variation. However, it is unclear to what extent these formality metrics are similar to the formality construct perceived by humans. This study first investigated what linguistic features predicted the text formality as humans constructed, then developed a weighted formality model, and finally tested this measure in different approaches. The corpus of this study consisted of 390 excerpts in TASA corpus with three genres: language arts, social studies and science. The five Coh-Metrix dimensions were used to develop the weighted formality model. Results showed the weighted model perceived by humans was constructed by five dimensions as theories constructed, but each dimension contributed differently to formality construct. This formality model was evaluated through comparisons between human construct of formality, Flesch-Kincaid Grade Level, and genres. All results showed the weighted formality scores had much higher correlations with human judgments of formality than non-weighted formality scores.

Introduction

Stylistic variation of language was originally proposed to identify the distinctive language style among different speakers/writers or within a single speaker/writer in different social contexts (Bell 1984; Hymes 1974; Labov 1973). Formal vs. informal language, or careful (formal) vs. casual (informal) speech (Graesser et al. 2014; Heylighen and Dewaele 2002), can be distinguished by both linguistic features such as choices of words, syntactic structures, semantic and pragmatic meanings at the multi-textual levels of discourse (Biber 1988; Carroll 1966; Graesser et al. 2014; Graesser, McNamara, and Kulikowich, 2011; Sardinha and Pinto 2014).

The present study investigated a measure of formality that attempt to account for human perceptions of formality as a gold standard. We first constructed the weighted formality model from the multiple discourse levels, namely, five primary Coh-Metrix components: narrativity, word concreteness, syntactic simplicity, referential cohesion, and deep cohesion. We then evaluated this formality model by comparing this model with an unweighted formality model from three aspects: (1) the model's reliability in human judgments of formality, (2) the model's sensitivity in accounting for variations in the Flesch-Kincaid Grade Level (FKGL and Klare 1974), and (3) in text genres.

This paper is organized into four sections. First, the paper briefly describes the definitions and measures of formality in previous studies. Then the automated text analysis tool−Coh-Metrix is introduced as it is used to generate the composite formality scores and to construct the model. Second section describes the methodology of the study and human judgment of formality. Results and discussion are reported in section three. The paper ends with conclusions and future work.

Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Definitions of Formality

Formality has been defined from different perspectives. Some definitions consider everyday oral conversations as a standard of informality (Atkinson 1982) and institutional or organizational conventions as a standard of formality (Andren, Sanne, and Linell 2010). Atkinson (1982) defined "formal" as "non-conversational," based on noticeable features that differed from everyday conversations in unfamiliar settings (e.g., court hearing), persons involved (e.g., attorney), or status of readers (e.g., professionals). The scope of this definition applies primarily to spoken registers, including an attempt to account for silence, repair, and turn allocation. Andren et al. (2010) further defined formality as achieved "according to officially standardized and recognized institutional conventions or prescriptions" (p. 224). They proposed four categories of informal linguistic features to reduce formality in conversations: informal lexical embedding (e.g., "hi there"), colloquial style (Hovy 1990) or jargons (e.g., "what do you say to that?" instead of "how do you plead, guilty or not?''; Linell, Alemyr and Jönsson 1993), omissions of formally required parts (e.g., abbreviations), and addition of non-task talks (e.g., phatic talk).

These definitions claim some features that distinguish formality, but they fail to provide an adequate construct of formality that is accepted by colleagues for a broad landscape of discourse, including registers designed for print.

Heylighen and Dewaele (2002) proposed the concepts of high-context versus low-context (Hall 1976) and defined formality as being progressively more prevalent when “a maximum of meaning is carried by the explicit, objective form of the expression, that is to say, the actual sequence of linguistic symbols used, rather than by the cluster of implicit, poorly delimited, and subjective factors that constitute a context.” Low-context increases in formality, whereas high-context decreases in formality. However, this definition stresses on the context that may impact the choice of formality, but fails to point out the important linguistic features that predict formality.

Graesser et al. (2014) claim that formal discourse, either in print or pre-planned oratory, occurs when there is a need to be precise, coherent, articulate, and convincing to an educated audience. Its opposite end of the continuum is informal discourse in oral conversation, personal letters, and narrative, which are replete with pronouns, deictic references (e.g., here, there, this, that), verbs, and reliance on common background. Formal language increases with grade level and with informational texts, but decreases with narrative texts.

In the current study, we adopted this theoretical definition in the directions for human formality rating when they scaled the text formality (see Human Rating section for details). Three reasons are elaborated why we adopted this definition in the directions: (1) it is the first time the characteristics of formality, being on a continuum rather than discrete, was explicitly elaborated in the definition, (2) it is first time formality was constructed with the consideration of many levels of language and discourse in its theoretical framework, and (3) this theoretical definition also considers important aspects of context (such as purpose, discourse planning, audience, and common background) that may impact the formality variation.

Measures of Formality

The previous computational measures of formality were primarily at the word level (Fang and Cao 2009; Heylighen and Dewaele 2002) or lexical level (Brooke and Hirst 2014; Hovy 1990; Peterson, Hohensee and Xia 2011). For example, Fang and Cao (2009) proposed that an adjective density (the ratio of frequency of adjectives to word tokens) would predict formality of text categories. Another popular measure of formality, F-score (formality-score), is sensitive to syntactic word categories (Heylighen and Dewaele 2002). Formality increases when there is a high relative frequency of nouns, adjectives, articles and prepositions, but a low frequency of pronouns, adverbs, verbs and interjections. The F-score measure has successfully accounted for the human construct of formality at the sentence level (Lahiri, Mitra, and Lu 2011) and for online diary analysis (Teddiman 2009) in English language and other European languages.

Graesser et al. (2014) proposed two measures of formality at the multi-textual levels rather than at the word level. The Coh-Metrix formality considered the levels of the multilevel theoretical framework, which includes words, syntax, text base, situation model, genre and rhetorical structure. (Graesser and McNamara 2011). The five Coh-Metrix dimensions are aligned with most of the levels of the multilevel theoretical framework (Graesser et al. 2004; McNamara et al. 2014). They developed a composite Coh-Metrix formality score, which was computed by the average of five Coh-Metrix dimensions. The Coh-Metrix formality increases with low narrativity, syntactical simplicity and word concreteness, but high referential cohesion and deep cohesion (see the Coh-Metrix section for detailed definitions of these five dimensions).

Grasser et al. (2014) also claimed the psychological and social attributes in the role of formality construct; therefore, they proposed the LIWC (Linguistic Inquiry and Word Count; Pennebaker, Booth, and Francis 2007) formality score. A LIWC formality composite score was computed from three of the dimensions. Higher LIWC formality scores were assumed to occur for texts with low narrativity (a robust component same as the Coh-Metrix narrativity) and processes-procedures-planning (action and events in procedures or processes or forecasted events, goals, or plans for the future), and high collection PC scores (words in the categories such as conjunction, inclusion, we, they).

Both Coh-Metrix and LIWC formality metrics were validated by correlating the scores with Flesch-Kincaid grade level (Klare 1974), Degrees of Reading Power (Koslin, Zeno, and Koslin 1987), Lexile scores (Stenner 1996) of text difficulty, and text genres. Li, Graesser, and Cai (2013) also confirmed the Coh-Metrix formality measure was more reliable and valid in predicting genres when compared with F-score. They found the Coh-Metrix formality scores could better predict the genre variation than F-score measure. Moreover, Li et al. (in press) conducted the study on measures of formality in Chinese language. They built the Chinese LIWC formality score as perceived by humans with the five LIWC components. The Chinese LIWC formality increased with low narrativity (the same function as the Coh-Metrix narrativity) and embodiment (e.g., body, ingest, health), but high cohesion (e.g., referential cohesion), positive emotions (e.g., happy, social, family), and negative emotions (e.g., anger, anxiety, death). This measure better explained text categories and had higher correlation with human formality than F-score and adjective density.

These findings from the studies on the Coh-Metrix and LIWC formality scores illustrate that formality measured at the multiple-textual levels better predicted genres than formality models measured at the word level. Moreover, the Chinese LIWC formality was constructed based on the human perception of formality and results showed this model was more reliable than the model constructed by the averaged dimension score (Li et al. in press). This raises a suspicion whether the five Coh-Metrix dimensions equally contributes to formality construct in the Coh-Metrix formality model.

To address the above issue, the present study aimed to develop a Coh-Metrix formality model based on human judgments of formality, in order to investigate how these five Coh-Metrix dimensions contribute to formality construct, and whether this model is robust to measure formality.

Coh-Metrix

Coh-Metrix (http://www.cohmetrix.com; Graesser et al. 2004; McNamara et al. 2014) was developed to analyze texts on the multilevel theoretical framework, such as words, syntax, the explicit text base, the situation model, and the discourse genre and rhetorical structure (Graesser and McNamara 2011). Modules of Coh-Metrix use lexicons, part-of-speech classifiers, syntactic parsers, templates, corpora, latent semantic analysis and other components, which are widely used in computational linguistics. The current public web site provided approximately 100 measures for colleagues to use. A principal components analysis (PCA) was performed on 37,520 texts in order to simplify the analysis and identify central constructs of text complexity (Graesser, McNamara, and Kulikowich 2011).

The PCA extracted eight dimensions that accounted for 67% of the variance in variations among texts. The top five of these dimensions were incorporated in Coh-Metrix-TEA (Text Easability Assessor; http://tea.cohmextrix.com). The five dimensions of Coh-Metrix-TEA were analyzed by Nelson et al. (2011) in the comparative assessment of text complexity metrics. The five primary dimensions of Coh-Metric-TEA are listed and succinctly defined below.

Narrativity. Narrativity tells a story, with characters, events, places, and things that are familiar to readers. Narrativity is closely affiliated with everyday oral conversation.

Word Concreteness. Concrete words evoke mental images and are more meaningful to the reader than abstract words.

Syntactic Simplicity. Sentences with few words and simple, familiar syntactic structures are easier to process and understand. Complex sentences have structurally embedded syntax, which increases the difficulty of comprehension.

Referential Cohesion. High cohesive texts contain words and ideas that overlap across sentences and the entire text, forming threads that connect the explicit textbase.

Deep Cohesion. Causal, intentional, and other types of connectives help the reader form a more coherent and deeper understanding of the text at the level of the causal situation model.

These five dimensions cover five of the 6 levels in the multilevel theoretical framework: genre, situation model, textbase, syntax, and words. Each of the five dimensions is expressed in terms of ease of comprehension. Text difficulty is defined as the opposite of ease, so principal component scores are reversed in measures of text difficulty (see Graesser and McNamara 2011; McNamara et al. 2014 for details).

This paper used these five dimensions to develop two measures of formality, unweighted (the average of five dimension scores) and weighted (trained on the corpus with the human formality annotation). The goal is to detect and evaluate the measure of formality as perceived by humans. Three approaches were adopted to evaluate the models of formality: (1) the human construct of formality, (2) Flesch-Kincaid grade level, and (3) text genres.

Method

Text Samples

The TASA corpus (The Touchstone Applied Science Associates, Inc.) includes academic texts for students from kindergarten to the first year of college in the United States (Zeno et al. 1995). In order to get an even distribution of genres, we randomly selected 12 text excerpts from each of three genres at each grade level with stratified random sampling. Therefore, 468 texts (12 3 13) were randomly selected for the analyses in this study. One text from each genre and each grade was put in a subset (1 3 13); therefore, 12 subsets were obtained in total. 78 texts in two subsets were randomly selected for training human judges to scale formality on texts, and 390 texts in 10 subsets for human rating. For both training and rating sessions, human judges were presented only 39 texts in one subset each time to avoid the fatigue of scaling the text formality. Table 1 illustrated the means and standard deviation of the word count and the sentence count in the 10 rating corpora.

Procedure

Coh-Metrix was performed on 390 texts from the TASA corpus to obtain the five Coh-Metrix dimension scores: narrativity, word concreteness, syntactic simplicity, referential cohesion and deep cohesion. An unweighted Coh-Metrix model was computed with the average of the five-dimension scores as proposed by Graesser et al. (2014)

[Unweighted formality score = (referential cohesion + deep cohesion – narrativity – syntactic simplicity – word concreteness)/5]. The weighted Coh-Metrix formality model was developed based on the human annotated formality corpus.

Table 1 Descriptive Statistics for the Rating Corpus (390).

GL	Word Count			Sentence Count		
	LA	SS	S	LA	SS	S
1	277(14)	269(14)	275(16)	34(10)	34(10)	36(6)
2	295(26)	276(27)	288(20)	31(5)	30(5)	32(5)
3	292(25)	283(22)	281(17)	27(6)	28(4)	29(5)
4	291(30)	266(12)	268(11)	22(6)	24(2)	26(4)
5	284(25)	278(20)	281(22)	19(5)	23(5)	26(5)
6	292(28)	297(27)	279(22)	16(4)	21(6)	20(5)
7	287(24)	294(33)	277(25)	18(3)	17(3)	22(5)
8	286(29)	291(24)	277(19)	15(4)	18(3)	21(4)
9	290(23)	292(28)	274(21)	15(5)	19(4)	20(6)
10	288(22)	303(19)	277(22)	15(4)	17(7)	20(5)
11	296(29)	304(10)	271(23)	13(2)	17(4)	19(3)
12	299(34)	289(28)	279(19)	13(5)	14(3)	18(4)
13	299(21)	321(16)	300(19)	12(4)	15(3)	16(2)

Note. N = 390. GL means Flesch-Kincaid grade level; LA means language arts; SS means social studies; and S means science. The first number is the mean and the standard deviation is within the parentheses.

Human Rating

Four native English speaking undergraduates participated in human judgments of formality: two females and two males. Their average age was 20.75 [19, 25]. First, they discussed the formality of three passages classified as non-academic, intermediate and academic by Snow and Uccelli (2009). All of them agreed academic passage had high formality; non-academic, low formality; and intermediate with the formality between the middle of high and low.

One week later, they were presented instructions of formality judgment. For example, formal texts tend to have more abstract words, complex syntax, and discourse cohesion than informal text. Formal text is typically in the expository rather than in the narrative genre. A written official document is more formal than a spoken conversation. Public speech tends to be more formal than a personal letter. Face-to-face conversations between two political leaders are more formal than conversations between two friends. Also, professional letters between professors are more formal than private letters between close friends. Human judges were allowed to use their own perception to scale formality, but required to keep consistency during the entire scaling. The judges rated the overall text formality using a 6-point scale with 1 (least formal) to 6 (most formal).

Inter-rater reliability was assessed by the intra-class correlation coefficient (ICC) with a two-way random model and absolute agreement type. In the training section, the inter judge reliability was extremely high (Cronbach's α = .80, ICC coefficient = .67; Landis and Koch 1977). After discussing the disagreements, judges independently coded the 390 texts with one subset at one time. The reliability in the rating sessions reached the threshold of inter judge reliability (Cronbach's α = .71, ICC coefficient = .68). In the analyses, we used the average of four judges' scaling scores.

Statistical Analyses

Multiple regression models with 10-fold cross-validation using Weka were performed to assess the extent to which the five components best capture the formality construct. Meanwhile, the weighted Coh-Metrix formality model was constructed based on the coefficients.

Intra-class correlations were used to compare the weighted formality model with the unweighted model and the averaged and individual human judgments of formality.

One-way ANOVAs were conducted to assess the extent to which formality model was sensitive to variation in text genres and grade levels.

Results and Discussion

We developed the weighted Coh-Metrix formality model to measure formality as humans perceived from five Coh-Metrix components, including narrativity, word concreteness, syntactic simplicity, referential cohesion, and deep cohesion. The model was evaluated by the comparisons with human construct of formality, and the prediction of grade level and genres.

Best-fit Models

Results of multiple regression with 10-fold cross-validation in Weka showed all of the five components were robustly attributed to human formality judgments with different weights (r = .76). Based on the multiple linear regression coefficients, the weighted Coh-Metrix formality model was constructed: [Weighted Formality = 4.18 – .71 * Narrativity – .26 * Syntactic Simplicity – .14 * Word Concreteness + .14 * Referential Cohesion + .10 * Deep Cohesion].

This model supports the claim that the components at the multi-textual levels are attributed with different weights to measuring formality construct (Li et al. in press). This model also provides an empirical foundation for the theoretical assumption that these five Coh-Metrix dimensions are aligned with the human-perceived formality construct. Formality increases with low narrativity, simple syntax, less concrete words, and high referential cohesion and deep cohesion. Narrativity is a strongly robust measure for formality construct (see Graesser et al. 2014; Li et al. in press). The other components played the minor roles in the contribution to the formality model as compared to the narrativity component. One reason is that narrativity has such apparent text characteristics (e.g., easy to process and to understand, story-like) that are easy for humans to detect, to process and to understand. However, other features such as syntax, word use, and cohesion are involuntarily ignored.

Model Comparisons

This section evaluated the weighted Coh-Metrix formality in terms of the human construct of formality, Flesch-Kincaid grade level, and genres.

Human Formality Construct. Intra-class correlations were computed to compare models with human construct of formality for a test set. Table 2 showed that the weighted formality model had considerably strong reliabilities with the average of the human judgments than the unweighted formality model. This model was as good as, if not better than the individual human judges (Rater 1, 2, 3 and 4). However, the unweighted formality had considerably lower reliabilities with each individual human judge. Thus, the weighted formality model was a more reliable and valid measure.

Table 2 Inter-rater Reliability Matrix between Models and Human Raters in the Testing Dataset.

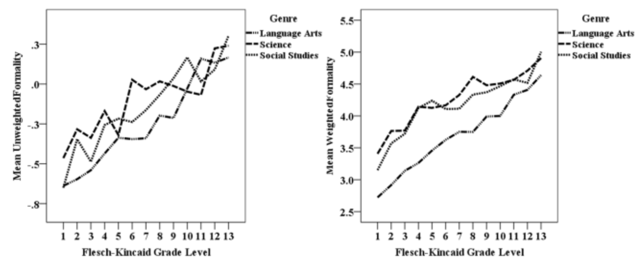
Raters	WF	UF	Rater1	Rater2	Rater3	Rater4
Rater1	.701	.429	–––	–––	–––	–––
Rater2	.648	.382	.680	–––	–––	–––
Rater3	.559	.348	.347	.487	–––	
Rater4	.703	.377	.748	.669	.345	–––
Average	.852	.578	.853	.861	.730	.853

Note. N = 390. WF means weighted formality score; UF means unweighted formality score; and Averaged means the average of human judgment of formality.

Grade Levels and Genres. One-way ANOVAS were performed to evaluate the models' sensibility to the variation in Flesch-Kincaid grade levels and text genres, respectively. There were no significant interactions between grade levels and genres for both weighted and unweighted formality models, but significance was found in main effects. Both formality models significantly predicted grades and genres, so the model that had the ability to explain more variances was the better model (Li et al. in press). The weighted formality model explained 61% of variance in the grade level and 31% in text genres, whereas the unweighted model explained only 38% of the variance in the grade level and 4% in genres. These results demonstrated that the weighted formality model had a considerably higher capability to predict both FK grade level and genres.

Figure 1 plots the unweighted and weighted formality scores as a function of the three genres (language arts, social studies, and science) and the 13 grade levels. The formality scores increased linearly as a function of the grade levels. The relative formality scores showed the expected ordering of science, social studies and language arts. The weighted model could substantially distinguish the narrative texts from the two informational texts. This pattern was more apparent in the weighted model than the unweighted model. The two informational genres (science and social studies) were moderately distinguished by the weighted formality model, but not by the unweighted model.

Figure 1 Mean of Unweighed and Weighted Formality of Each Genre at Different Flesch-Kincaid Grade Levels.



FK grade level is sensitive to sentence length, word length, and word frequency. The grade level is robustly decreased as a function of narrativity and syntactic simplicity, and moderately decreased with word concreteness. Word frequency heavily loads on the narrativity dimension and sentence length on the syntactic dimension (Graesser et al. 2014). Thus, it is not surprising the weighted model predicts grade level better than unweighted model.

Our comparisons of models with the variation in grade level and genre confirmed that both formality metrics measured at the multi-textual levels could distinguish grade level and genre. However, the weighted model constructed by human construct of formality explained the most variance, as compared with the unweighted formality model constructed by theoretical frameworks. Model comparisons with the human formality construct support the claim that this model best predicts the construct of formality as perceived by humans, as we would expect.

Conclusions

The concept of formality originated intuitively and evolved theoretically. Researchers developed measures of formality from the word level to the multiple textual levels, and conducted empirical studies to evaluate the models from the aspect of the ability to predict genres and grade level. This study incorporated human judgments of formality to develop a better measure of formality as perceived by humans. The Coh-Metrix formality does increase when texts have low narrativity, syntactic simplicity and word concreteness, but high referential cohesion and deep cohesion. The findings imply that narrativity is the most robust component in the formality construct, as compared to other components, such as word, syntax and cohesion.

The future studies should analyze the texts with the high agreement and disagreement that human judges scaled. Thus, we may explore why some texts are easy to reach agreement in formality scales and some are difficult; and what linguistic features occur in these texts. Moreover, the weighted Coh-Metrix model may be evaluated on other unannotated corpora in order to test the generalizability of the weighted model to other languages.

References

Andrén, M.; Sanne, J. M.; and Linell, P. 2010. Striking the Balance between Formality and Informality in Safety-critical Communication: Train Traffic Control Calls. *Journal of Pragmatics* 42: 220–241.

Atkinson, J. M. 1982. Understanding formality: The Categorization and Production of 'Formal' Interaction. *The British Journal of Sociology 33*: 86–117.

Bell A. 1984. Language Style as Audience Design. *Language in Society* 13(2): 145–204.

Biber, D. 1988. *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Brooke, J.; and Hirst, G. 2014. Supervised Ranking of Co-occurrence Profiles for Acquisition of Continuous Lexical Attributes. In *Proceedings of COLING*, 2172–2183. Dublin City University and Association for Computational Linguistics.

Carroll, J. B. 1966. Vectors of Prose Style. In T. A. Sebeok ed. *Style in Language*, 283–292. New York: Wiley.

Fang, A. C.; and Cao, J. 2009. Adjective Density as a Text Formality Characteristic for Automatic Text Classification: A study based on the British National Corpus. In *Proceedings of the PACLIC*, 130–139.

Graesser, A. C.; and McNamara, D. S. 2011. Computational analyses of Multilevel Discourse Comprehension. *Topics in Cognitive Science,* 3, 371–398.

Graesser, A. C.; McNamara, D. S.; Cai, Z.; Conley, M.; Li, H.; and Pennebaker, J. 2014. Coh-Metrix Measures Text Characteristics at Multiple Levels of Language and Discourse. *Elementary School Journal* 115: 210–229.

Graesser, A. C.; McNamara, D. S.; Louwerse, M. M.; and Cai, Z. 2004. Coh-metrix: Analysis of Text on Cohesion and Language. *Behavior Research Methods, Instruments, and Computers* 362: 193–202.

Graesser, A. C.; McNamara, D.S.; and Kulikowich, J. 2011. Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *Educational Researcher* 40: 223–234.

Hall, E. T. 1976. Beyond Culture. New York: Anchor Press.

Heylighen, F. and Dewaele, J. M. 2002. Variation in the Contextuality of Language: An Empirical Measure. *Foundations of Science* 7: 293–340.

Hovy, E. H. 1990. Pragmatics and Natural Language Generation. *Artificial Intelligence*, 43, 153-197.

Hymes, D. H. 1974. *Foundations in Sociolinguistics: An Ethnographic Approach* (Vol. 65). Philadelphia: University of Pennsylvania Press.

Klare, G. R. 1974. Assessing Readability. *Reading Research Quarterly* 10, 62–102.

Koslin, B. I.; Zeno, S.; and Koslin, S. 1987. *The DRP: An Effective Measure in Reading*. New York: College Entrance Examination Board.

Labov, W. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Lahiri, S.; Mitra, P.; and Lu, X. 2011. Informality Judgment at Sentence Level and Experiments with Formality Score. *Computational Linguistics and Intelligent Text Processing* 446–457.

Landis, R. J. & Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 159–174.

Li, H.; Graesser, A. C.; and Cai, Z. 2013. Comparing Two Measures of Formality. In C. Boonthum-Denecke & G. M. Youngblood eds., In *Proceedings of the 26th International Florida Artificial Intelligence Research Society Conference*, 220–225. Palo Alto, California: AAAI Press.

Li, H.; Graesser, A. C.; Conley, M.; Cai, Z.; Pavlik, P.; and Pennebaker, J. W. in press. A New Measure of Text formality: An Analysis of Discourse of Mao Zedong. *Discourse Processes*.

Linell, P.; Alemyr, L.; Jönsson, L. 1993. Admission of Guilt as a Communicative Project in Judicial Settings. Journal of Pragmatics 19, 153–176.

McNamara, D. S.; Graesser, A. C.; McCarthy, P. M.; and Cai, Z. 2004. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

Nelson, J.; Perfetti, C.; Liben, D.; and Liben, M. 2011. *Measures of text difficulty: Testing Their Predictive Value for Grade Levels and Student Performance*. New York: Student Achievement Partners.

Pennebaker, J. W.; Booth, R. J.; and Francis, M. E. 2007. *Linguistic Inquiry and Word Count: LIWC* [Computer software]. Austin, TX: LIWC.net.

Peterson, K.; Hohensee, M.; and Xia, F. 2011. Email Formality in the Workplace: A Case Study on the Enron Corpus. In *Proceedings of the Workshop on Languages in Social Media*, 86-95. Association for Computational Linguistics.

Sardinha, T. B.; and Pinto, M. V. eds. 2014. *Multi-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber*. Amsterdam: John Benjamins Publishing Company.

Snow, C. E.; and Uccelli, P. 2009. The Challenge of Academic Language. In Olson, D. R.; and N. Torrance. eds. *The Cambridge Handbook of Literacy* 112–133. Cambridge: Cambridge University Press.

Stenner, A. J. 1996. *Measuring Reading Comprehension with the Lexile Framework*. Durham, NC: Metametrics, Inc.

Teddiman, L. 2009. Contextuality and beyond: Investigating an Online Diary Corpus. In E. Adar, M. Hurst, T. Finin, N. Glance, N. Nicolov; and B. Tseng. eds. In *Proceedings of the Third International Conference on Weblogs and Social Media*, 331–333. Menlo Park, California: AAAI Press.

Zeno, S. M.; Ivens, S. H.; Millard, R. T.; and Duvvuri, R. 1995. *The Educator's Word Frequency Guide*. Brewster, NY: Touchstone Applied Science Associates.

Acknowledgements

This work was funded by the National Science Foundation (BCS 0904909), the Army Research Institute (W91WAW-07-C-0029), and the Minerva Initiative, Air Force Office of Scientific Research, Department of Defense (FA9550-14-0308). Any opinions are only those of the authors and do not necessarily reflect the views of these funding agencies, cooperating institutions, or other individuals.