Comparing Two Measures for Formality

Haiying Li, Zhiqiang Cai, Arthur C. Graesser

202 Psychology Building; University of Memphis; Memphis, TN, 38152-3230
hli5@memphis.edu|zcai@memphis.edu|graesser@memphis.edu

Abstract

Formality is an important dimension of language style. Texts of different genres tend to have different degrees of formality. F-score (formality-score) is a most popular measure for formality to differ genres. It uses a method of combining proportions of words of different types, with nouns, adjectives, articles and prepositions as positive elements, and adverbs, verbs and interjections as negative elements. This measure is limited to the word level; however, indicators of genre are not restricted to the word level. This paper proposes a new composite component score of formality, "CF-score", which consists of a combination of five Coh-Metrix component scores: narrativity, referential cohesion and deep cohesion, syntactic simplicity, word concreteness. F-score and CF-score are compared with TASA corpus in diverse text genres. Results of General Linear Model showed CF-score formality measure was more powerful and more rational with text genres than F-score.

Introduction

In sociolinguistic domain, researchers showed their interests in studies on style of language variation. Labov (1972) claimed the most urgent issue to be solved was the quantification of dimension of style. Bell (1984) asserted style as one element of language variation was often measured but seldom explained. Heylighen and Dewaele (1999, 2002) proposed formality was a universal dimension of stylistic variation and provided a measure for formality with a combination of proportions of words of different types to differ among genres. This measure is restricted to word level; however, the indicators of genre are not restricted to the word level, but involved in the syntax and discourse level.

This paper proposes a composite component score of formality, "CF-score", which consists of a combination of five Coh-Metrix component scores: narrativity, referential cohesion and deep cohesion, syntactic simplicity, word concreteness. F-score and CF-score are compared with TASA corpus in diverse text genres. Formality measured at multi-dimensional level is hypothesized to better differ genres than F-score.

This paper is organized in four sections. First, the paper briefly describes the definition and measure of formality. Measure of formality focuses on Heylighen and Dewaele's F-score. Then, Coh-Metrix, an automated text analysis tool, is introduced as it is used to generate the composite component score of formality. Second section describes the methodology of generating CF-score and the research design for comparing two measures, F-score and CF-score. Results and discussion are reported in section three. Tthe paper ends with conclusions and future work.

Formality

Formality of language is defined as the expression that does not vary with changes of context and it has the characteristics of "continuum" from least formal to most formal (Heylighen and Dewaele 1999, 2002). This definition denotes semantic ambiguity from the unstated implication could be resolved in the formal language style (Gorfein 1989). For example, spatial-temporal context makes adverbs of time and space meaningful if both listen and speaker are in the face-to-face conversation. For example, "here" and "today" are explicit in the sentence "Mary is here today" if both listener and speaker clearly know the context of the time and space. Without the situational context, references of "here" and "today" are implicit. On the contrary, the specific noun such as "classroom" and concrete date such as "November 20, 2012" are required if speaker and listen are not in the same situational context. Instead, we will use this sentence "Mary is in the classroom in November 20, 2012" to avoid semantic ambiguity and fuzziness in expression (Levelt 1989).

Formal language explicitly states the sufficient references, assumptions and background knowledge without the dependence of context. In contrast, the informal language embedded the above information in the context. Therefore, Context-dependence or context-independence is one of the most important indicators of formality (Heylighen and Dewaele 1999, 2002).

The extent of shared context or common ground between speaker and listener during the conversation determines the choice of type of discourse during communication (Clark and Marshall 1981). If speaker and listener share more background, it is easier for them to understand referred expressions with less effort (Krauss and Weinheimer 1966). Consequently, they tend to use less formal language dependent on their shared context (Heylighen and Dewaele 1999, 2002). If they do not have more shared common ground, they tend to choose more formal language independent on the context to avoid ambiguity and fuzziness (Biber 1988; Clark and Brennan 1991; Heylighen and Dewaele 1999, 2002).

The next section introduces the measure of formality in previous research.

Measure for Formality

A lot of studies on the quantification of formality have attempted to use linguistic features. These linguistic measures included proportion of explicitness versus implicitness (Hasan 1984; Leckie-Tarry 1995), lexical density to differ written texts from spoken (Halliday 1985), nouniness in politeness speeches (Brown and Levinson 1979), and the ratios between the amount of nouns and verbs, nouns and pronouns, and the parts of speech related to the nominal like nouns, adjectives and articles, and the verbal like auxiliary verbs and adverbs (Fielding and Fraser 1978). Even though the researcher proposed these theories, they did not evaluate the theories with empirical studies.

Biber (1988) used 67 linguistic features to differ spoken and written texts with the method of factor analysis. The extracted six factors include involved versus informational production, narrative versus non-narrative concerns, explicit versus situation-dependent reference, overt expression of persuasion, abstract versus non-abstract information, on-line informational elaboration. Even though this multi-dimensional analysis provides a much clearer variation across speech and writing, this method could not distinguish all genres significantly. In addition, six underlying factors representing linguistic features seem too much for some researchers. Therefore, one combination of dimension is required to represent the style of language of variation to differ genres.

Heylighen and Dewaele (1999, 2002) proposed formality as a universal dimension of stylistic variation. They claimed there were two types of formality: surface formality and deep formality according to the extent of required attention to form for the sake of convention or for the sake of accurate interpretation. The formal style of language has the features of detachment, accuracy, rigidity, cognitive load, and dense information. These features are represented by high proportions of nouns, adjectives, prepositions and articles in texts. In contrast, informal style has the features of flexibility, directness, implicitness, involvement and less information. These features are indicated by pronouns, verbs, adverbs and interjections. Thus, they provided a measure for formality with the formula below:

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

This measure of formality is called as "F-score". "F-score" refers to formality score within the range from 0 to 100 in percentage. The higher F-score, the more formal the text tends to be. F-score is recently used to evaluate the concept of formality at the sentence level with the human judgment (Lahiri, Mitra, and Lu 2011). Teddiman (2009), however, found F-score failed to distinct formality of online diary entries from online diary comments. Thus, a more effective measure for formality is put on the agenda to satisfy the fine-genre analysis.

The goal of the current paper is to develop a composite component score to measure formality extending from coarse word categories to fine-grained categories, and from lexical level to syntactic, semantic and discourse level with automated text analysis tool, Coh-Metrix.

Coh-Metrix

Coh-Metrix was developed to analyze texts on multiple linguistic features with indices at levels of words, syntax, the explicit text base, the situation model, and the discourse genre and rhetorical structure (Graesser, McNamara, Louwerse, and Cai 2004; McNamar, Graesser and Kulikowich 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.

Coh-Metrix includes hundreds of indices related to readability or text difficulty, cohesion and language, 53 indices among which were evaluated and used to perform the Principal Components Analysis (PCA). Table 1 showed the linguistic indices in each dimension with primary loading greater than .35. The orthogonal varimax rotation extracted eight components in terms of Eigen value greater than 1. The first five components were labeled as narrativity, syntactic simplicity, word concreteness, referential cohesion and deep cohesion according to the underlying linguistic features in each dimension. The rest three components including no more than three indices were not considered in this study.

Narrativity refers to texts that are story-like or informational with linguistic features such as nouns, adjectives, verbs, pronouns and adverbs etc. Part of these linguistic features was used by F-score, which treated nouns and adjectives as positive elements, and verbs, adverbs and pronouns as negative elements. However, if more linguistic features such as syntactic structure ease and cohesion were

included, the word type features are mutually represented in one underlying feature "Narrativity."

Dimension	Label	Description
Narrativity	DENPRPi	Pronoun
	NOUNi	Noun
	SYNNP	Number of modifiers per noun phrase
	READASW	Word length, number of syllables
	FRCLacwm	CELEX content word frequency
	VERBi	Verb
	PRO3i	Third person pronoun
	FRCLaewm	CELEX Log frequency for all words
	WRDAacwm	Age of acquisition for content words
	ADJi	Adjective
	WRDFacwm	Familiarity for content words
	PRO1i	First person pronoun
	FRCLmcsm	CELEX Log minimum for content words
	ADVi	Adverb
	DENNEGi	Negation density
	MEDwtm	Minimal Edit Distance, part of speech
	AGLSPSVi	Agentless passive voice density
Referential Cohesion	CRFPC1um	Content word overlap, adjacent sentences
	CRFPCaum	Content word overlap, all sentences
	CRFBA1um	Argument overlap, adjacent sentences
	CRFBAaum	Argument overlap, adjacent sentences
	MEDawm	Minimal Edit Distance, all words
	LEXDIVVD	Lexical diversity, VOCD, all words
	LEXDIVTD	Lexical diversity, MTLD, all words
	CRFBN1um	Noun overlap, adjacent sentences, binary
	TYPTOKc	Type-token ratio, content word lemmas
	LSAGN	LSA given/new, sentences
	CRFBSaum	Noun overlap, all sentences, binary
	LSAassa	LSA overlap, adjacent sentences
	LSApssa	LSA overlap, all sentences in paragraph
Syntactic Simplicity	READASL	Sentence length, number of words
	CAUSV	Causal verb
	STRUTt	Sentence syntax similarity, across paragraphs
	INTEi	Intentional verbs incidence
	LexDensity	Lexical diversity
	SYNLE	Left embeddedness, words before main verb
Word Concreteness	WRDIacwm	Imagability for content words
	WRDCacwm	Concreteness for content words
	WRDMacwm	Colorado norms, content words
Deep Cohesion	CONCAUSi	Sentence syntax similarity, paragraphs
	CONLOGi	Logical connectives
	CONi	All connectives
	CONTEMPi	Temporal connectives
	INTEC	Ratio, intentional particles to intentional verbs
	CAUSC	Ratio, casual particles to causal verbs

Table 2: Linguistic indices in each dimension with loading greater than .35.

Referential cohesion was denoted by explicit arguments and ideas that overlapped across all sentences or within adjacent sentences. Linguistic feature in this dimension included noun overlap, argument overlap, content word overlap, lexical density, and latent sematic analysis, etc. Similarly, deep cohesion was indicated by the connectives that link sentences or clauses such as causal, intentional, spatial or temporal connectives from the perspective view of the deeper semantic level.

Syntactic simplicity, the level of difficulty of sentences, was designated by sentence length, causal verbs, sentence syntax similarity, etc. The higher scores indicated the sentence had fewer words and simpler syntactic structure. Word concreteness was signified by content words related to concreteness and imagability.

Therefore, Coh-Metrix makes it possible for us to obtain five comprehensive dimensional scores which will be used to compute CF-score at the multi-textual level. Meanwhile, Coh-Metrix also provides the indices of all the word types which will be used to compute F-score.

The goal of this paper is to test which measure of formality could better predict text genres, F-score or CF-score.

Method

Corpus

TASA corpus (The Touchstone Applied Science Associates, Inc., http://nora.hd.uib.no/corpora/2004-1/0322.html) is used in this study. TASA corpus includes academic textbooks for students from kindergarten to the first year of college in the United States (Zeno et al 1995). The TASA corpus has nine text genres consisting of 119,627 paragraphs taken from 37,651 samples with 10,829,757 words (See Table 2). The majority of genre is language arts (42.62%), followed by social studies (27.82%) and science (14.26%). Other genres such as business, health, miscellaneous, home economics are minority of genres, not exceeding 4%. There is an unknown genre (5.88%) in this corpus. We, therefore, used TASA in this paper because it is a large corpus with miscellaneous fined text genres and has proven to be able to represent other texts and differences between text genres.

Genre	N	Word Count		WPS	
		Mean	SD	Mean	SD
Business	1079	302.17	19.80	18.58	4.64
Health	1359	279.18	20.46	14.99	5.64
Home Economy	283	269.29	20.33	20.47	4.58
Industrial Arts	142	278.49	23.63	16.29	3.17
Language Arts	16044	285.94	24.59	17.49	10.10
Miscellaneous	675	293.95	21.64	19.92	8.45
Science	5356	283.96	22.99	15.61	6.17
Social Studies	10501	292.91	24.75	20.01	8.96
Unknown	2212	282.93	24.41	18.62	5.17
Total	37651	287.64	24.56	17.99	8.87

Table 2: Descriptive statistics of words per sentence and number of cases in each genre.

Note: WPS = Word per sentence.

Procedure

Coh-Metrix is used to analyze TASA corpus and obtain the individual indices of linguistic features and the five dimension scores.

To compute F-score, the following word types are needed: nouns, adjectives, prepositions, articles as positive elements, pronouns, verbs, adverbs, and interjections as negative elements. These indices are computed in permillage. Thus, percentage is first computed for these eight word categories. Then F-score is computed according to formality formula proposed by Heylighen and Dewaele (1999):

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

F-score is a value within the range from 0% to 100%.

A combination of formality score is computed with the combination of the five-dimension scores with the formula below:

CF-score = (referential cohesion + deep cohesion – narrativity – syntactic simplicity – word concreteness)/5.

Dimension scores are standardized scores. CF-score, therefore, is a standardized score with "0" as the mean, above "0" representing more formal, below "0" representing less formal.

To compare F-score and CF-score, percentage of F-score is converted into Z score in order to make F-score have the same value unit as CF-score with the formula below:

$$z = (x - \mu)/s \text{ (Marx and Larsen, 2006)}$$

among which, z is the standardized score we want, x is the observation value, μ is the population mean, and s is the standard deviation of the population.

Based on this formula, F-score is computed:

$$\text{F-score}_Z = (\text{F-score} - 49.2112)/6.01559$$

In this formula, 49.2112 is the mean of F-score in all the cases in TASA corpus; and 6.01559 is the standard deviation. F-score represents the value of each observation. Thus, we converted percentage F-score to Z F-score, which is labeled as F-score_Z (We still use F-score instead of F-score_Z to represent measure for formality proposed by Heylighen and Dewaele for convenience).

Research Design

Independent variable is text genres with nine types. Dependent variable is F-score and CF-score respectively. General Linear Model (GLM, Univariate; McCullagh and Nelder, 1989) is used to examine whether the differences in formality measured by F-score or CF-score exist among nine genres. If formality measured by both F-score and CF-score significantly differ between text genres, which measure has more power?

Results and Discussions

Univariate Analysis

GLM was performed with genres as independent variable, and F-score and CF-score as dependent variable respectively to examine which measure would better distinguish nine texts genres.

Results of GLM Univariate analysis with CF-score as the dependent variable and genres as independent variable showed formality of texts differed statistically among genres across nine types, $F\ (1, 8) = 1455.16$, $p < .001$ (See Table 3). The effect size is .237. Tukey's post-hoc comparisons (Kirk 1982) of nine genres indicate that language arts (M = -.23, 95% CI [-.24, -.23]) had significantly least formal style among nine genres ($p < .001$). Health (M = .10, 95% CI [.08, .12]) didn't show difference in formality from Science (M = .10, 95% CI [.09, .11]), Social Studies (M = .15, 95% CI [.14, .15]) and Miscellaneous (M = .16, 95% CI [.12, .19]), $p > .05$. However, these four genres showed significantly less formal style than industrial arts (M = .24, 95% CI [.19, .29]), unknown (M = .33, 95% CI [.32, .35]), home economy (M = .45, 95% CI [.41, .49]) and business (M = .51, 95% CI [.46, .53]), $p < .05$. Industrial arts were significantly less formal than unknown, home economy and business, $p < .05$. Unknown was significantly less formal than home economy and business ($p > .05$). Home economy and business did not show difference in formality, but they were most formal texts among nine genres ($p < .05$).

Genre	Mean (CF-score) & CI
Language Arts	-.23 [-.24, -.23]
Health	.10 [.08, .12]
Science	.10 [.09, .11]
Social Studies	.15 [.14, .15]
Miscellaneous	.16 [.12, .19]
Industrial Arts	.24 [.19, .29]
Unknown	.33 [.32, .35]
Home Economy	.45 [.41, .49]
Business	.51 [.49, .53]

Table 3: Mean of formality with measure of CF-score

Results of GLM Univariate analysis with F-score as the dependent variable and genres as independent variable showed formality of texts differed statistically among genres across nine types, $F\ (1, 8) = 780.218$, $p < .001$ (See Table 4). The effect size is .143. Tukey's post-hoc comparisons (Kirk 1982) of nine genres indicate that language arts (M = -.40, 95% CI [-.42, -.39]) and Health (M = -.34, 95% CI [-.39, -.29] had no significantly difference in formality, but both had significantly least formal style among nine genres ($p < .001$) with the mean below "0'. Miscellaneous (M = .06, 95% CI [-.01, .13]) showed statistically less formal style than social studies (M = .26, 95% CI [.24, .28]), unknown (M = .40, 95% CI [.36, .44]), industrial arts (M = .42, 95% CI [.27, .57]), business (M = .50, 95% CI [.45, .55]), science (M = .48, 95% CI [.45, .50]), home economy (M = .63, 95% CI [.52, .74]), $p < .001$. Social studies were statistically less formal than business, science and home economy ($p < .05$). Social studies, unknown and industrial arts did not show significantly difference in formality ($p > .05$). Business did not show difference in formality from science, but were significantly less formal than home economy ($p < .05$). Science and home economy did not show difference in formality, but showed statistically most formal style among nine text genres ($p < .05$).

Genre	Mean (F-score) & CI
Language Arts	-.55 [-.42, -.39]
Health	-.41 [-.39, -.29]
Miscellaneous	.16 [-.01, .13]
Science	.39 [.45, .50]
Business	.50 [.45, .55]
Social Studies	.50 [.24, .28]
Unknown	.52 [.36, .44]
Industrial Arts	.56 [.27, .57]
Home Economy	.63. [.52, .74]

Table 4: Mean of formality with measure of F-score$_Z$

Effect size of Univariate analysis with CF-score ($\eta 2$ = .237) was greater than F-score ($\eta 2$ = .143). This indicated that CF-score could explain 23.7% variance in formality by text genres. In contrast, F-score could only explain only 14.3% variance in formality by text genres. Therefore, a combinational dimension measure of CF-score for formality at the multi-textual level had more power than the measure of F-score for formality at the word level in terms of genre prediction (See Figure 1).

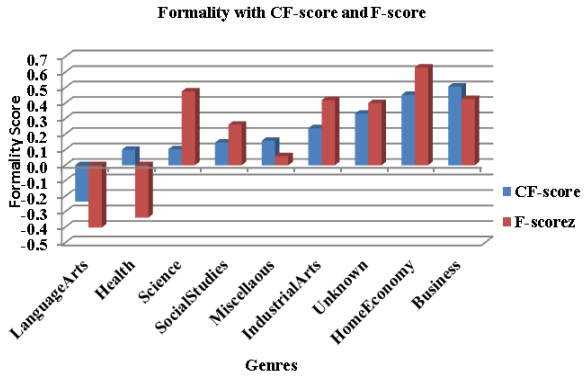

Figure 1. Means of formality with CF-score and F-scorez

Both of these two measures for formality could significantly distinct nine text genres. Tukey post hoc comparison, however, showed the differences in these two measures. Tukey post hoc showed homogeneous subset with harmonic mean with unequal cases in each genre had much explicit group classification of genres with CF-score than F-score (Kirk 1982).

In addition, according to subjects of curriculum, language arts contained more narrative texts. Based on Biber's multi-dimensional analysis (1988), language arts tended to be less formal than other text genres due to the feature of narrativity. In comparison of CF-score and F-score, language arts did show least formal style. However, by comparison with other genres, results from CF-score made more sense, because only formality of language arts was below zero. F-score indicated formality of both language arts and health was below zero. Furthermore, language arts are significantly different from health in CF-score, but no significant difference exists in F-score. Therefore, CF-score has much more power to differ language arts from health than F-score.

Moreover, the formality with CF-score tended to be more rational than F-score was reflected in two similar text genres home economy and business. Formality of these two genres did not show significant difference in CF-score, but did show significant difference in F-score. As we know, home economy is one specific branch of business. Therefore, these two text genres should show similar degree of formality. This hypothesis was approved by CF-score, but was failed by F-score.

Moreover, health, to some extent, is one specific type of science. Therefore, these two text types should show similar degree of formality. This theoretical hypothesis was supported by CF-score, showing these two genres no significant difference in CF-score. However, F-score showed these two genres had substantial statistic difference in formality.

The above hypotheses supported by CF-score could also be explained by the curriculum of these subjects in Common Core Standards. Both subjects such as language arts and science are in the curriculum from Kindergarten. Even if they are different in language style (narrative versus informational), the difference between two genres should not have that dramatic discrepancy in formality considering the readability of texts across 12 grades. Industrial arts, home economy and business are collected in curriculum at higher grade levels. This assumes formality of these genres to be more formal than those included at emergent level. These curriculum-based hypotheses were supported by formality measured by CF-score from least formal to most formal in order: language arts, health, science, social studies, miscellaneous, industrial arts, unknown, home economy and business (See Figure 1).

In summary, formality measured by CF-score with a combination of comprehensive and multi-level dimension scores could better distinguish fine genres or subgenres than F-score with a combination of single word-level scores.

Conclusions and Future Work

This paper compared two measures for formality. One is F-score proposed by Heylighen and Dewaele (1999). F-score used the combination of proportion of word categories, including positive elements nouns, adjectives, articles and prepositions and negative elements verbs, adverbs, pronouns, and interjections. Another is CF-score developed by authors in this paper with the automated text analysis tool, Coh-Metrix. 53 linguistic indices generated by Coh-Metrix were used to perform Principal Component Analysis. Then five mulit-textural dimensions were used to compute the composite formality score. These five dimensions included

narrativity, referential cohesion, deep cohesion, syntactic simplicity, and word concreteness. Thus, the combination of formality score was obtained according to these five dimensions.

General Linear Model was performed separately on F-score and CF-score and results showed that CF-score could better distinguish texts among nine genres than F-score with more power. The different degree of formality in nine genres with CF-score showed more reasonable interpretation according to the classification of subgenres and the curriculum of these subjects guided by the Common Core Standards. The reason is that CF-score was developed considering multi-levels, including word categories, complexity of syntactic structure, latent semantic overlapping, and all types of cohesion. This comprehensive synthetic measure for formality is proven better than single word level measure to distinct formality among the fine-genres.

The limitation of this study is that we didn't use the same corpus in the study of Heylighen and Dewaele (1999) due to free access to their original corpus. The further work is suggested to conduct with the same corpus in their study for better evaluation of these two measures. Besides, in the future work, the human rating will be used to better evaluate these two measures.

References

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.

Brown, P. and Levinson, S. 1979. Universals in Language Usage: Politeness Phenomena. In E.N. Goody (ed.), *Questions and Politeness. Strategies in Social Interaction*, 56-289. Cambridge: Cambridge University Press.

Clark, H. H., and Brennan, S. E. 1991. Grounding in communication. In L. B. Resnick, J. M. Levine, and S. D. Teasley, *Perspectives on Socially Shared Cognition*, 127-149. Washington, DC: American Psychological Association.

Clark, H. H., and Marshall, C. R. 1981. Definite reference and mutual knowledge. In A. K. Joshi, B. Webber, and I. A. Sag (Eds.), *Elements of discourse understanding*, 10-63. Cambridge, England: Cambridge University Press.

Fielding, G. and Fraser, C. 1978. Language and Interpersonal Relations. In I. Markova (ed.), *The Social Context of Language*, 217-232. Chichester: Wiley.

Gorfein, D.S. ed. 1989. *Resolving Semantic Ambiguity*. New York: Springer Verlag.

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.

Halliday, M.A.K. 1985. *Spoken and written language.* Oxford: Oxford University Press.

Hasan, R. (1984) Ways of saying: ways of meaning. In R. P. Fawcett, M.A.K. Halliday, S.M. Lamb, A. Makkai (eds.), *The semiotics of Culture and Language.* Vol. 1 *Language as Social Semiotic*, 105-162. London & Dover: Pinter.

Heylighen, F. and Dewaele, J. M. 1999. Formality of Language: Definition, Measurement and Behavioral Determinants. Technical Report, B-1050, Free University of Brussels, Brussels, Belgium.

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

Krauss, R. M. and Weinheimer, S. 1966. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology* 4: 343-346.

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

Lahiri, S., Mitra, P., & Lu, X. (2011). Informality judgment at sentence level and experiments with formality score. *Computational Linguistics and Intelligent Text Processing* 446-457.

Leckie-Tarry, H. 1995. *Language and context. A functional linguistic theory of register.* London-New York: Pinter.

Levelt ,W. J. M. 1989. *Speaking. From intention to articulation*, Cambridge, MA: MIT. Press.

Kirk, R.E. 1982. *Experimental design: Procedures for the behavioral sciences.* Monterey, CA: Brooks/Cole Publishing Company.

Marx, M. L., and Larsen, R. J. 2006. *Introduction to mathematical statistics and its applications*. New Jersey: Pearson/Prentice Hall.

McCullagh, P. and Nelder, J. 1989. *Generalized Linear Models, Second Edition.* Boca Raton: Chapman and Hall/CRC.

McNamara, D. S. and Graesser, A. C. 2011. Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution,* 188-205. Hershey, PA: IGI Global.

Teddiman, L. 2009. Contextuality and Beyond: Investigating an online Diary Corpus. In *Proceedings of the Third International ICWSM conference*, 331-333. San Jose, Calif: Third International AAAI Conference on Weblogs and Social Media.

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.

Acknowledgments

This work was funded by the National Science Foundation BCS 0904909) for the project: Modeling discourse and social dynamics in authoritarian regimes. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these funding agencies, cooperating institutions, or other individuals.