Linguistic Style and Social Historical Context: An Automated Linguistic Analysis of Mao Zedong's Speeches

Ying Duan, Nia M. Dowell, Arthur C. Graesser, and Haiying Li

Institute for Intelligent Systems & University of Memphis, TN 38152, USA
{yduan1|ndowell|a-graesser|hli5}@memphis.edu

Abstract

Times of crisis and prosperity may be the most defining moments of leadership and therefore one of the most important contexts in which to study leadership processes. In the present research we explored linguistic patterns of cognitive complexity, social representations and social coordination of Mao Zedong speeches during different socio-historical contexts, namely times of prosperity and crisis. The texts of Mao Zedong were analyzed using a computerized text analysis tool, Linguistic Inquiry Word Count (*LIWC*), to explore how his linguistic style was influenced by the social climate. The Pearson's correlations and structural equation modeling results show, during times of prosperity, Chairman Mao's linguistic style increased in cognitive complexity, social representations and social coordination.

Introduction

Political leadership remains an important topic of research; receiving attention from a variety of fields, including political science, psychology, discourse analysis, and sociology (Ammeter, Douglas, Gardner, Hochwarter, and Ferris 2002). Within this research, a number of methodological approaches have been employed to gain deeper insights into leadership processes (Hancock, Beaver, Chung, Frazee, Pennebaker, Graesser, and Cai,). However, scholars have widely acknowledge leadership as being inextricably bound to language, discourse, and communication (Bligh, Kohles, and Meindl 2004).

The language and discourse of famous political leaders has proven to be a fruitful avenue exploring psychological states, cognitive functioning, and more macro-level social dynamics and strategies of influence (Bligh, Kohles, and Meindl, 2004; Guerini and Stock, 2010) . This is in line with previous research showing linguistic and discourse properties are diagnostic of a number of psychological and social phenomena, such as personality, depression, deception and emotion (Agarwal and Rambow, 2010; D'Mello, Dowell, and Graesser, 2009; D'Mello and Graesser, 2012; Mairesse and Walker, 2010; Rude, Gortner, and Pennebaker, 2004; Saxbe, Yang, Borofsky, and Immordino-Yang, 2012). If political leadership is manifested in language and discourse in a fashion that reflects the socio-historical context, then it is worthwhile to conduct systematic investigations of the linguistic and discourse patterns of the language of leaders.

Leadership processes cannot be adequately studied out of a historical context (Faris and Parry 2011). In order to better understand the language of leaders' speeches in history, it is necessary to use a framework that includes contextual variables, namely those things which surround it in time and place and which give it its meaning. These historical contextual variables define the political, social, cultural, and economic setting for a particular idea or event, which in some cases prove problematic for leadership (Faris and Parry 2011). From this perspective, we hypothesize linguistic style is reflective of the socio-historical contexts.

Markers of linguistic style have been linked to a number of interesting psychological features. For instance, personal pronouns can provide information about group processes. Increases in first person plural ("we", "us", "our") and third person plural ("they", "them") pronouns can signal distinctions group identity or social representations (perdue). Linguistic patterns can also reveal the complexity of an individual's thinking (slacher, chung). For example, the use of discrepancies (e.g. "would", "should") and tentative (e.g. "maybe", "perhaps") is associated with more cognitively complex language. Seemingly unimportant language, such as assents (e.g. "yes", "OK") or fillers (e.g. "you know" and "I mean"), is revealing of group cohesion and social coordination.

The present research uses automated linguistic analyses to explore the speeches of a long-tenured Chinese autocratic leader, during times of conflict and stability. Linguistic Inquiry Word Count (LIWC) (Pennebaker, Booth, and Francis, 2007), was used to investigate the linguistic and discourse variation of 293 original Chinese language texts delivered by Chairman Mao. LIWC is a computerized text analysis tool that reports the percentage of words in a text that are in either grammatical (e.g. articles, pronouns, prepositions), psychological (e.g.

emotions, cognitive mechanisms, social), or content categories (e.g. home, occupation, religion). LIWC provides roughly 80 word categories, but also groups these word categories into broader dimensions. More generally, our novel contribution to the understanding of political leadership would show how leadership style can be gleaned from language and discourse use in varying socio-historical context, namely crisis and prosperity.
This is example text. It is 10 point Times New Roman. The first sentence after the heading begins without a paragraph indent.

Method

Corpus

The texts were collected from *Selected Works of Mao Zedong*. There were 293 original Chinese language speeches that Chairman Mao delivered between the years of 1925 and 1957. The corpus included Mao Zedong's official reports, orders, claims, assertions and interviews.

Measures

Historical measures. Major economic, historical, and social data were collected over the span of Mao's leadership in order to study the relationship between those factors and Chairman Mao's linguistic style. The factors that represent the population, economic status, and the stability of China were *Population, Gross Domestic Product* per capita (GDP), and *Lack of Disturbances* (LD). The population and GDP data were based on figures of Maddison (2009). The values of these two variables were converted to standardized values according to the following formula: [(x-minimum)/ (maximum-minimum)].

Lack of Disturbances (LD), as a historical measurement, reflected major events/states that occurred during the time span of Mao's speeches. Specifically, LD included civil wars, other wars, and border disputes. For instance, "civil wars" were internal country conflicts, such as the 1st, 2nd, and 3rd civil war between Kuomintang and the Communist Party for the control of China. "External conflicts" refer to the wars between the Chinese army and outside countries, such as the anti-Japanese war and Korean War. "Border Disputes" refers to the border conflicts between China and its neighboring countries, such as China-India border dispute and the China-Soviet dispute. The events were coded according to scholarly publications (Cheek 2002; Li and Wang 2008). The codes for a composite variable called *Disturbances* were aggregated over these three forms of conflict and converted to the following scale: 0= no incident occurred, 1= one incident occurred, and 2= two incidents occurred. For the purposes of the current investigation, these Disturbances measures were reverse coded to create the Lack of Disturbances measure (LD). Therefore, higher numbers indicated more stability or lack of disturbances.

Word categories. Linguistic constructs of social coordination, cognitive complexity, and social representations were measured using word categories from an automated linguistics tool, Chinese LIWC. Specifically, this analysis focused on three types of words categories. The two indices that made up the social coordination construct were *Assent* (e.g. "yes", "OK") or *Fillers* (e.g. "you know" and "I mean"). There were two indices that represented the cognitive complexity construct: *Discrepancy* (e.g. "should", "would") and *Tentative* (e.g. "maybe", "perhaps", and "guess"). The social representations construct included six personal pronoun indices, namely 1st person singular (e.g. I, my, and me), 1st person plural (e.g. we, our, and us), 2nd person singular (e.g. you, you'll), 2nd person plural (Categories in Chinese LIWC only, e.g. you), 3rd person singular (e.g. she, he, her, his, and him), and 3rd person plural (e.g. they, their, and them).

Procedure

Data analysis. There were some Missing values of GDP per capita from 1925 to 1928 and 1939 to 1949. These values were represented by the average values of their adjacent years. For instance, the missing values from 1939 to 1949 were computed by averaging the values of 1938 and 1950. Our primary data analyses used SEM to test the fit of a series of models to our predictions.

Item parceling. To increase accuracy of parameter estimates, word types of the first person pronoun, the second person pronoun, and the third person pronoun were parceled respectively (Bandalos 2002). Hall, Snell, and Foust (1999) have provided evidence from a simulation study that parceling can lead to better parameter estimates if there are omitted variables that lead to shared variance among items in a parcel. Parceling can also increase statistical power compared with either path analysis or loading every item from a scale on one factor because fewer parameters are tested (Tempelaar, Gijselaers, van der Loeff, and Nijhuis 2007).

SEM approach. SEM analyses were carried out with the LISREL 8.70 software program (Mels 2004) using maximum likelihood estimation. We used the two-step measurement model and the full structural model approach that is frequently recommended in the literature (Kline 2005). ***Fit indices.*** The fit of each model was assessed using the recommendations from Hu and Bentler (1999) for samples of $N > 200$. The fit of nested models was tested using the chi-square test of difference (Kline 2005). The chi-square test for a good-fitting model should be nonsignificant. CFIs (confirmatory fit index) from a good-fitting model should be greater than .95. RMSEAs (root mean square error of approximation) should be less than .08, or the confidence interval should straddle .05. SRMRs (standardized root mean residual) from a good-fitting model should be less than .06 (Hu and Bentler 1999). All

statistical tests were assessed using an alpha level of $p <$ .05.

Results and Discussion

Descriptive statistics and correlations among all measures are shown in Table 1. We began by fitting a measurement-only model, which is equivalent to fitting a set of confirmatory factor analyses on each factor while simultaneously allowing all factors to correlate with each other. The 2nd person pronoun showed a negative regression weight and the value is much smaller than the other two kinds of pronouns under the latent variable social representation, therefore, this indicator was removed.

Correlations were added between the errors of First person pronoun and Filler, the errors of Social Coordination and Cognitive Complexity, and the errors of Social Coordination and Social Representation. Since LIWC is used to count the percentage of words in a specific category, there exists some ambiguous definition and gap between the categories of words. Thus, we let the above index errors correlate. This also indicated some potential relationships between the words: the unexplained parts of the Social Representation and Social Coordination constructs have some overlap. The conclusion also can apply to the Social Coordination and Cognitive Complexity Constructs. With all the other measurements, the measurement model showed an excellent fit to the data, $\chi^2(20) = 25.57$, p = .18, *CFI*=. 99, *SRMR*= .027, *RMSEA*= .030, 90% CI [<.001, .062] (See Table 2). This suggests that the factor is hypothesized to drive, and therefore fitting a structural equation model is warranted.

The analyses proceeded and we fit the structural model. The model showed an excellent fit $\chi^2(21) = 25.90$, p=.21 *CFI*=1.00, *SRMR*= .028, *RMSEA*= .028, 90% CI

Variables	1	2	3	4	5
1. Population (P)					
2. Gross Domestic Product per capita (GDP)	.674**				
3. Lack of Disturbances (LD)	.802**	.741**			
4. First Person (FP)	.056	.029	.014		
5. Second Person (SP)	.044	-.049	-.039	.081	
6. Third Person (TP)	.219**	.210**	.248**	.174**	-.054
7. Discrepancy (D)	.299**	.196**	.258**	.094	.001
8. Tentative (T)	.161**	.228**	.250**	.015	.003
9. Assent (A)	.331**	.346**	.405**	.181**	-.003
10. Fillers (F)	.254**	.243**	.244**	.352**	-.003
M	.4884	.5551	1.2491	2.3015	1.7177
SD	.2583	.1849	.5389	1.1272	.9422

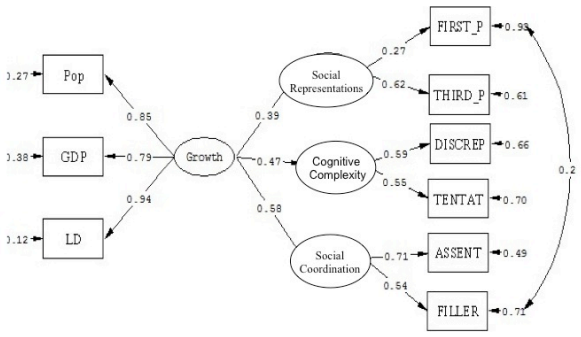
Table 1: Descriptive and Correlations Statistics
Note. N=293. * $p < .05$; ** $p < .01$.

[<.001, .060] and supported our hypotheses. As indicated in Table 2, the structural model was not significantly different from the measurement model, $\Delta\chi^2(1) = .33$, p = .5657, suggesting that the model tested our prediction fit very well.

Fit index	Measurement model	Structural model
$\chi^2(df)$	25.57 (20)	25.90 (21)
CFI	.99	1.00
SRMR	.027	.028
RMSEA	.030	.028

Table 2: Indicators of Fit for the Measurement Model and Structural Model
Note: *CFI= confirmatory fit index. SRMR= standardized root mean residual. RMSEA= root mean square error of approximation. CI= confidence interval.*

Figure 1 illustrates the structural model: all the three paths in the model were statistically significantly different from zero. The result indicated that the Social Coordination construct was most influenced most by Growth (standardized path loading= .58), the second was Cognitive Complexity (standardized path loading= .47), followed by Social Represenations (standardized path loading= .39). Thus, the variances of the latent constructs can be explained by Growth are 33.6%, 22.1%, and 15.2%.



One limitation pertains to the measure reliabilities less than ideal (Cronbach's α for Growth, Social Coordination,

	6	7	8	9	10
1. Population (P)					
2. Gross Domestic Product per capita (GDP)					
3. Lack of Disturbances (LD)					
4. First Person (FP)					
5. Second Person (SP)					
6. Third Person (TP)					
7. Discrepancy (D)	.015				
8. Tentative (T)	.049	.323**			
9. Assent (A)	.313**	.233**	.254**		
10. Fillers (F)	.268**	.214**	.166**	.390**	
M	3.0102	.6970	.5803	.1261	.2940
SD	1.3897	.5630	.4988	.3322	.3365

Figure 1: Structural Equation Model showing the relationship between prosperity and linguistic style

Cognitive Complexity, and Social Representations were .815, .427, .483, and .202, respectively. According to Kline (2005), the values of reliabilities in this study are from unacceptable to very good. If we had used a larger number in these measurements, we might have been able to create more reliable secondary factors, which would increase

measurement precision. This research is a valuable addition to the linguistic and discourse variation research by supporting previous findings that indicate people demonstrate consistent changes in linguistic style as a function of socio-historical climate. The present findings suggest that linguistic style plays an important role in representing individual change in political leaders during significant social periods.

Conclusion

Results showed that Growth had significant effects on the words use of Spoken words, Cognitive words, and Pronoun words. The influences from Growth to the three categories from the largest to the smallest are Spoken words, Cognitive words, and Pronoun words. These results support our prediction. The finding indicated that when times are difficult, as in the war and civil discontent, leaders use less first pronouns and third pronouns, the style of leader's speeches were less spoken but more coherent, and leaders used fewer cognitive words in their speeches since during that time their confidence was not high, and they were not totally involved and accepted by the public. To the contrary, when times are good, as in the case of a good economy and population growth, then the leaders were more confident and arrogant. Therefore, there will be more spoken words and cognitive words in their speeches, and leaders may use more first pronouns and third pronouns in their speeches.

Acknowledgements

This research was supported by the National Science Foundation (BCS 0904909, DRK-12-0918409), and U.S. Department of Homeland Security (Z934002/UTAA08-063). 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.

References

Agarwal, A., and Rambow, O. 2010. Automatic detection and classification of social events. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (pp. 1024–1034). Stroudsburg, PA, USA: Association for Computational Linguistics.

Ammeter, A. P., Douglas, C., Gardner, W. L., Hochwarter, W. A., and Ferris, G. R. 2002. Toward a political theory of leadership. The Leadership Quarterly, 13(6): 751–796. doi:10.1016/S1048-9843(02)00157-1

Bandalos, D. L. 2002. The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. Structural Equation Modeling, 9: 78–102.

Bligh, M. C., Kohles, J. C., ands Meindl, J. R. 2004. Charting the language of leadership: a methodological investigation of President Bush and the crisis of 9/11. The Journal of applied psychology, 89(3): 562–574. doi:10.1037/0021-9010.89.3.562

Cheek, T. 2002. Mao Zedong and China's Revolutions: A Brief History with Documents. Boston: Palgrave Macmillan.

D'Mello, S., Dowell, N., and Graesser, A. 2009. Cohesion Relationships in Tutorial Dialogue as Predictors of Affective States. In Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling (pp. 9–16). Amsterdam, The Netherlands, The Netherlands: IOS Press.

D'Mello, S., & Graesser, A. C. 2012. Language and Discourse Are Powerful Signals of Student Emotions during Tutoring. IEEE Transactions on Learning Technologies, 5(4): 304–317.

Faris, N., and Parry, K. 2011. Islamic organizational leadership within a Western society: The problematic role of external context. The Leadership Quarterly, 22(1): 132–151.

Guerini, M., & Stock, O. 2010. Predicting Persuasiveness in Political Discourses. In Proceedings of the International Conference on Language Resources and Evaluation. Valletta, Malta: European Language Resources Association.

Hall, R. J., Snell, A. F., and Foust, M. S. 1999. Item parceling strategies in SEM: Investigating the subtle effects of unmodeled secondary constructs. Organizational Research Methods, 2: 233–256. doi:10.1177/ 109442819923002

Hancock, J. T., Beaver, D. I., Chung, C. K., Frazee, J., Pennebaker, J. W., Graesser, A., and Cai, Z. 2010. Social language processing: A framework for analyzing the communication of terrorists and authoritarian regimes. Behavioral Sciences of Terrorism and Political Aggression, 2(2):x 108–132.

Hu, L., and Bentler, P. 1999. Cut off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6: 1–55.

Kline, R. 2005. Principles and practice of structural equation modeling (2nd ed.). New York, NY: Guilford Press.

Li, J. and Wang, S. 2008. The outline of China's modern history. Higher Education Press.

Maddison, A. 2009. Statistics on World Population, GDP and Per Capita GDP.

Mairesse, F., and Walker, M. A. 2010. Towards personality-based user adaptation: psychologically informed stylistic language generation. User Modeling and User-Adapted Interaction, 20(3): 227–278. doi:10.1007/s11257-010-9076-2

Pennebaker, J. W., and Graybeal, A. 2001. Patterns of Natural Language Use: Discourse, Personality, and Social Integration. Current Directions in Psychological Science 3:90-93.

Pennebaker, J. W., Booth, R. J., and Francis, M. E. 2007. Linguistic Inquiry and Word Count: LIWC. Austin, TX: LIWC.net.

Saxbe, D. E., Yang, X.-F., Borofsky, L. A., and Immordino-Yang, M. H. 2012. The embodiment of emotion: language use during the feeling of social emotions predicts cortical somatosensory activity. Social Cognitive and Affective Neuroscience, nss075. doi:10.1093/scan/nss075

Tempelaar, D., Gijselaers, W., van der Loeff, S. S., and Nijhuis, J. 2007. A structural equation model analyzing the relationship of student personality factors and achievement motivations, in a range of academic subjects. Contemporary Educational Psychology, 32: 105–131.