Language change in Report on the Work of the Government by Premiers of the People's Republic of China

Renkui Hou
Guangzhou University, Guangzhou, China
hourk0917@163.com

Chu-Ren Huang
The Hong Kong Polytechnic University, Hong Kong
churen.huang@polyu.edu.hk

Kathleen Ahrens
The Hong Kong Polytechnic University, Hong Kong
Kathleen.ahrens@polyu.edu.cn

Abstract

The present paper explored the focusing topics change and language change in Report on the Work of the Government by Premiers of the People's Republic of China (hereinafter Report texts). The text clustering and correspondence analysis showed the focusing topics change in selected three periods Report texts. The Report texts were represented by the clause length distribution and clustered. The clustering result showed the differences of clause length usages in the Report texts. The relationship between clause length and word length was studied. The average word length decreases with clause length and were fitted using the function, $y = ax^b$ based on the Menzerath-Altmann Law. The relationship between the three periods Report texts represented by the fitted parameters, *a* and *b*, were explored.

1 Introduction

Language change has been main concern of linguists from centuries in many parts of the world and is the topic of research for classical linguistic studies. However, these early works tended to focus on sound and sound changes, tracing back to work by Pāṇini in 4th century BCE. Biber (2012) argued strongly that reference works that describe different linguistics levels, i.e., lexical, grammatical, and lexico-grammatical, should consider register difference.

Language is the mode of political discourse. For example, the president candidates will demonstrate their ideas and policies in the debates. Van Dijk (1997) defined the political discourse based on three dimensions: the actors, the political scope of the discourse and the context of communication. From the above definition, a discourse is considered as 'political' when it is produced by a political actor carrying out a political action (e.g. to govern, legislate, protest or vote) in an institutional context of communication. Randour et al. (2020) conducted a systematic literature review of 164 scientific articles from the Scopus database and confirmed that political discourse is generally limited to the discourses of (institutionalized) political elites and most specifically to oral monological speeches.

The Menzerath-Altmann law originates from the fact that the length of a construct influences the lengths of its immediate constituents in different language domains. It is summarized as "the greater the whole, the smaller its parts" by Paul Menzerath after he detected the dependency of syllable length on word length (Menzerath 1954). Altmann generalized this hypothesis to all the language levels, formulating it as "The longer a language construct, the shorter its components" (Altmann 1980). Hřebíček (1992, 1995, 1997) showed that the whole hierarchy of textual levels is based on this dependency, and called this the Menzerath-Altmann law.

Altmann (1980) gave the theoretical derivation and the corresponding differential equation of the MA law, as shown in Equation (1).

$$\frac{y'}{y} = -c + \frac{b}{x} \quad \text{Equation (1)}$$

The solution to this differential equation is shown in the Formula (1):

$$y = ax^b e^{-cx} \quad \text{Formula (1)}$$

where *y* is the mean size of the immediate constituents, *x* is the size of the construct, and parameters *a*, *b*, and *c* depend mainly on the levels of the units under investigation.

A large number of observations have shown that parameter *c* is close to zero for higher levels of language whereas lower levels lead to very small values of

parameter *b*; only for intermediate levels is the full formula needed (Köhler, 2012).

The two simplified formulas were obtained when higher and lower levels were studied respectively. Formula (2a) has become the most commonly used "standard form" for linguistic purposes (Grzybek, 2007).

$$y = ax^b \qquad \text{Formula (2a)}$$
$$y = ae^{-cx} \qquad \text{Formula (2b)}$$

This paper explores the language change in Chinese political discourse, Report on Work of the Government by Premiers of the People's Republic of China, based on the content development and Menzerath-Altmann law from the perspective of quantitative linguistics.

1.1 Literature review

Millar and Trask (2015) demonstrated that languages change (even their spelling rules) throughout their history. Previous studies about language change focused on sound and sound changes, word and word changes mostly. Lieberman et al. (2007) studied the regularization of English verbs over the last 1200 years and how the rate of regularization depends on the frequency of word usage. Lexicostatistics was used to calculate the evolutionary history of a set of related languages and varieties (Bakker et al. 2009, Barbancon et al. 2013). Baker (2011) focused on words that have changed their frequency and meaning in the study of change in British English over the twentieth century. Degaetano-Ortlieb and Teich (2018) have used relative entropy for detection and analysis of periods of diachronic linguistic change. Campos et al. (2020) set a corpus-driven methodology to quantify automatically diachronic language distance between chronological periods of several languages. The results showed that a diachronic language distance based on perplexity detects the linguistic evolution that had already been explained by the historians of the three languages.

There is a long tradition of linguistic research on political discourse. Van Dijk's (1997) definition of political discourse brings together studies focusing on discourse produced by political elites in an institutional context with the aim of carrying out a political action (Randour et al. 2020). There are some studies aims at studying political issues, events and actors, such as ideology or identity construction (Wang 2007, Wodak and Boukala 2015). Other studies concentrated on specific linguistics characteristics of the discourses, such as Wang and Liu's (2018) analysis of Trump discourse, or Roitman's (2014) study of the use of pronouns by French presidential candidates. Lu and Ahrens (2008) studied the metaphor usage in political discourse. Savoy (2018) examines the verbal style and rhetoric of the candidates of the 2016 US presidential primary elections.

Yu (2008) demonstrated that machine learning methods can be trained to classify congressional speeches according to political parties. Better performance levels can be achieved when the training examples are extracted from the same time period as the test set. This means that the congressional speeches have different stylistic features in different periods. Yu (2013) explored the correlation between language usage and gender, and reveals that (political) feminine figures tend to use emotional words more frequently and employ more personal pronouns than men.

Previous research has validated the MA law at different language levels. Tuldava (1995) examined the dependence of average word length on clause length, finding a statistically highly significant interdependence between average word length and clause length, indicating that there are other factors that influence average word length. Motalová et al. (2014) and Ščigulinská and Schusterová (2014) verified the validity of the MA law applied to contemporary written and spoken Chinese respectively. Benešová (2016) tested the potential validity of the MA law on samples in different languages and attempted to test the concept of this language universal. Hou et al. (2017) studied the relationship between sentence length and clause length in Chinese language and concluded that the relationship in formal written register can be fitted by the MA law.

There are some researches for studying the theory and formula per se. Köhler (1984) interpreted the parameters in Formula (2a) and assumed that *a* represents a quantity depending on the language and language levels, *b* might represent a shortening tendency and might describe the range of structural information that has to be stored for each language component. Cramer (2005) showed that parameters *a* and *b* might be linked by a systematic connection through a correlation analysis. Hou et al. (2019) fitted the relationship between clause length and word length in different Chinese registers and concluded that the relationship between the fitted parameters, *a* and *b*, in each register can be fitted by the linear regression. The result of linear regression is different in different Chinese register texts.

There are few studies on Chinese language change and on linguistic characteristics of political discourse. This paper will explore the language change in Chinese political discourse based on topic words and the MA law.

1.2 Data and methodology

The Report texts on Works of Government of China in three different periods were selected to establish the corpus. These three different periods are 1978-1982, 1997-2001 and 2016-2020 respectively. The first five years, 1978-1982, are the initial stage of the reforming and opening up in China. Hong Kong was returned to China

in 1997. The last five years, 2016-2020, is the 13th five-year plan was initiated and finished.

The texts of Report were segmented and Parts of Speech tagged using the Chinese Lexical Analysis System created by the Institute of Computing Technology of the Chinese Academy of Sciences (ICTCLAS).

The text vectors were established based on Vector Space Model and bag of words. Text clustering was used to validate whether the selected linguistic characteristics can differentiate the Report texts in different periods. The differences can be demonstrated and interpreted. They reflect the language change and the focusing topic transition in three periods Reports. Correspondence analysis was used to analyze the correlation between these characteristics and different periods report texts.

The Formula (2a) was used to fit the average word length distribution in the clause with certain lengths (i.e., the relationship between clause and words). Then, the texts can be represented by these fitted parameters, *a* and *b*, and were displayed in a 2-dimensional space. Thus, the relationship between Report texts from these periods can be explored.

2 The change of the thematic words

The Report texts, as political discourses, speak about sharply defined topics and concentrate more or less on the core of the information. Usually the topics are linguistically represented by a particular number of nouns (or even proper nouns) and first order predicates, namely verbs and adjectives. The nouns represent the concrete and abstract concepts in the Report. The verbs represent the action from the subjects or on the objects which are all represented by nouns. The nouns can be modified by the adjective(s) and one noun phrase can be established. The nouns, adjectives and verbs are summarized as thematic words (Popescu et al. 2009).

The nouns, verbs and adjectives, occurring more than 50 times in all texts, were selected to represent the Report texts in three periods. Each report text is represented as one vector using the relative occurrence frequencies of these words.

The hierarchical clustering analysis was used to cluster the texts. Kullback-Leibler Divergence (Relative Entropy) were adopted to compute the distance between text vectors. The sum of squares of the deviations was used to calculate the distance between two clusters. The hierarchical clustering result, dendrogram, is shown in Figure 1.

From Figure 1, we can see that the Report texts from one period were clustered into one cluster. This means that there are systematical similarities of the content words usages in the same period texts and there are systematical dissimilarities between different periods Report texts. The internal differences of content words usages in one period texts, 1997-2001 and 2016-2020 respectively, are small compared with that in 1978-1982 texts. This may be caused by the drastic social changes in that period. The height of common ancestors of these three periods Report texts showed the relationship between the content words usages.

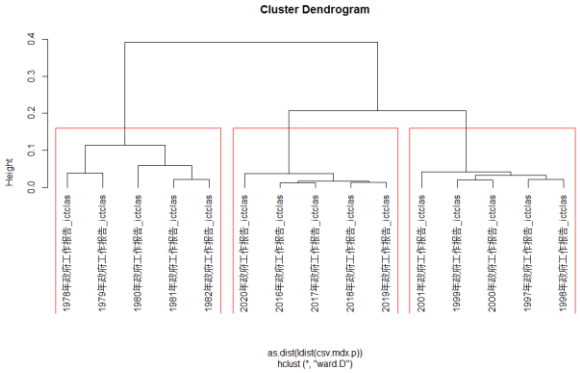

Figure 1: The result of hierarchal clustering of texts represented by the content words

Correspondence analysis is a summary technique which outputs a correspondence plot. A 2D correspondence plot is the most useful depiction of complex reality because it reduces the number of dimensions of variation to the manageable two dimensions represented by the x-axis and the y-axis. Its unique feature is the fact that it captures both the column (content words) and row (periods of report) categories of the cross-tabulation table in the same space.

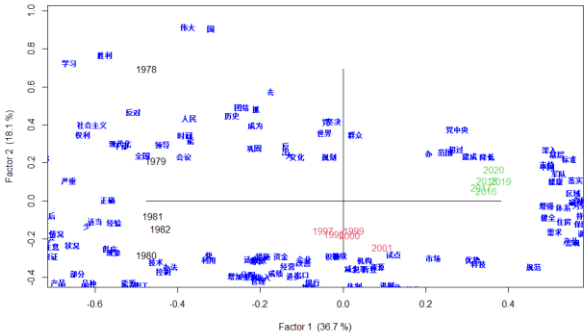

Figure 2: The correspondence analysis result of Report texts in three periods

The result of correspondence analysis is shown in Figure 2. The Figure 2 shows three periods Reports texts (1978-1982, 1997-2001, 2016-2020) were clustered according to their usages of different content words (nouns, verbs and adjectives). Both the reports texts and content words are displayed in the plot. From Figure 2, we can see that some words co-occurred with some Reports from different periods. For example, "健康/住房/增强 (health/housing/enhance)" overlapped with the Reports from 2016-2020. It means these issues are the concerning

focus of Government of these five years. “经验/学习/现代化 (experience/learn/modernization)” were the concerning focus of Government in the initial stage of the reform and opening. “企业/机构/试点/改善 (enterprise/organization/experimental unit/improvement” were focused by Government in 1997-2001.

3 The change of relationship between clause length and word length

The sentence, as the maximal grammatical unit and minimal statement unit, is considered to be a basic linguistic unit in all languages. Chinese sentences are often defined in terms of characteristics of speech (Huang and Shi 2016; Lu 1993). Chao (1968) and Zhu (1982) defined a sentence as an utterance with pauses and intonation changes at its boundaries.

A common approach for identifying sentences in syntactically annotated corpora (e.g., Chen et al., 1996; Chen et al., 2003; Huang and Chen, 2017 for Sinica TreeBank) is to mark all segments between punctuation marks that indicate utterance pauses as sentences. Such punctuation marks include commas, semicolons, colon, periods, exclamation marks, and question marks. Wang and Qin (2014) and Chen (1994) also adopted this operational definition and called such units *sentence segments*. Wang and Qin (2014) considered the lengths of *sentence segments* to be relevant to language use in Chinese. Sentences (as defined by Chen et al., 2003; Huang and Chen, 2017) and sentence *segments* (as defined by Chen, 1994; Wang and Qin, 2014) are roughly equivalent to clauses.

3.1 The distribution of clause length

Clause length is defined as the number of words included. Words is considered to be the segments delineated by blank spaces in the texts segmented by a Chinese lexical analysis system. The occurrence frequencies of clauses with certain lengths were calculated in three periods texts and the relative frequency distributions of the clauses in three periods texts are shown in Figure 3.

Figure 3 shows that the relative frequency distributions of clause length in three periods Report texts are similar. The relative occurrence frequencies of clauses increase firstly and then decrease with the increasing of lengths. The occurrence frequencies of one- and two-word clauses are highest in 1978-1982 and are lowest in 2016-2020. The clauses lengths concentrate on the 3-10 words. The percentages of clauses with 1-15 words lengths are more than 95% in three periods texts.

Each text is represented by the relative frequency of clause lengths. Correspondence analysis was used to analyze the texts. The correspondence analysis result, correspondence plot, is shown in Figure 4.

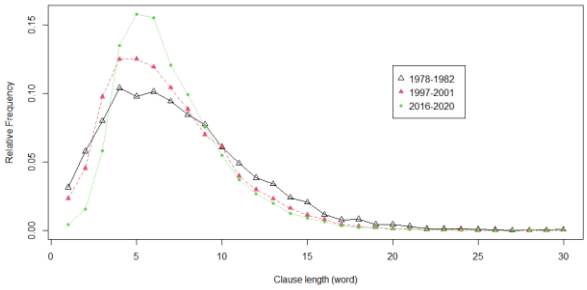

Figure 3: The frequency distribution of clause length in terms of words

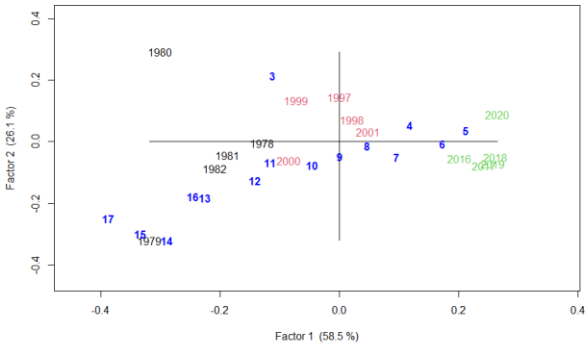

Figure 4: The result of correspondence analysis of texts

From Figure 4, there are differences of clause length usages between Report texts from three different periods. Combined with Figure 3, the short clauses are frequently used in 2016-2020 Report texts and the long clauses are frequently used in 1978-1982.

3.2 Fitted results of average word length in clauses

The average word length in clauses with certain lengths was calculated as the number of Chinese characters in the given clauses divided by the number of words in those clauses. We calculated the average word length distribution in each period texts and fit them using the Formula (2a). The fitted result is shown in Table 1 and Figure 5.

Table 1: The fitted result of the average word length in clauses with certain length

	a	*b*	R^2	*Adjusted* R^2
1978-1982	2.452	-0.144	86.49%	85.45%
1997-2001	2.457	-0.132	88.02%	87.1%
2016-2020	2.654	-0.180	88%	87.08%

From Figure 5, we can see that the average word length decreases with the clause length. The average word

length distribution can be fitted by the Formula (2a) in each period text. The determination coefficient, R^2, in Table 1 showed that the fitted result is good. Based on the fitted result, we conclude that the relationship between clause and word abides by the MA law.

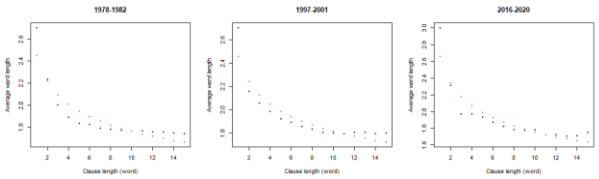

Figure 5: The fitted result of average word lengths in clauses

The average word lengths in 15 texts from three periods were fitted by the MA law. The fitted results were shown in Appendix 1. The values of R^2 demonstrate that the relationships between clause and word lengths in all Report texts abide by the MA law.

The Report texts from three periods are represented by the two fitted parameters, *a* and *b*, of the average word length in clause with certain length. They are displayed in a two-dimensional space, as shown in Figure 6.

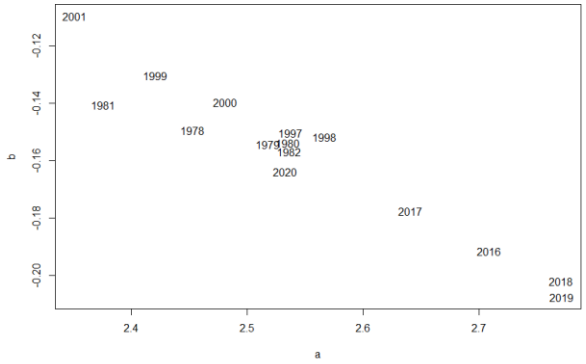

Figure 6: The relative position of Report texts from each period

From Figure 6, the two parameters, *a* and *b*, correlated negatively, which means *b* decreases with *a*. The *b* values are smallest in Report texts in 2016-2020, which means the extent of the decreasing of the average word length in clauses is maximum. The big *a* values in 2016-2020 Report texts mean that average word length in short clauses are larger than the other two periods texts. The ranges of fitted parameters are similar, but the relationship between these two parameters is not similar in texts from 1978-1982 and 1997-2001.

4 Conclusion

This paper studied the changes of language and focusing topics of Chinese political discourse represented by Reports on Work of the Government by Premiers of the People's Republic of China. We selected the Report texts in three periods, 1978-1982, 1997-2001 and 2016-2020, to establish the corpus of political discourse. Text clustering result showed that the thematic words in Report texts of these three periods are changing. Correspondence analysis showed that the correlation between Report texts and thematic words in correspondence plot. From correspondence plot, it can be seen that the changes of thematic words usages in these three periods Report texts.

Then the Report texts were represented by the clause length distribution and analyzed using correspondence analysis. The result of correspondence analysis showed that the more short clauses were used and the less long clauses were used with time from 1978-1982, 1997-2001 to 2016-2020. The average word length in clauses with certain lengths were calculated, which decreases with the clause length. Formula 2a was used to fit the average word length. The fitted result showed that the relationship between clause and word lengths abide by the MA law.

The 15 Report texts were represented by the fitted parameters of average word length. The parameters were used to represent the Report texts. The two-dimensional space was used to show the relationship between the texts. The result showed that the parameters *b* in 2016-2020 Report texts are smaller than that in Report texts from 1978-1982 and 1997-2001. It needs to be explored for the relationship between the fitted parameters and the development of Chinese language.

Acknowledgements. We would like to thank the anonymous reviewers for their insightful and helpful comments.

Funding support. Research on this paper was funded by National Social Science Fund in China (Grant No. 16BYY110), the Hong Kong Polytechnic University Grant 4-ZZFE, National Natural Science Fund in China (Grant No. 61866035).

References

Altmann, G. (1980). Prolegomena to Menzerath´s law. Glottometrika 2, 1-10.

Bakker, D., Muller, A., Velupillai, V., Wichmann, S., Brown, C.H., Brown, P., Egorov, D., Mailhammer, R., Grant, A. and Holman, E.W. (2009). Adding typology to lexicostatistics: A combined approach to language classification. Linguistic Typology 13(1), 169–181.

Baker, P. (2011). Times may change, but we will always have money: diachronic variation in recent British English. Journal of English Linguistics, 39(1), 65–88.

Barbançon, F., Evans, S., Nakhleh, L., Ringe, D. and Warnow, T. (2013). An experimental study comparing linguistic phylogenetic reconstruction methods. Diachronica 30, 143–170.

Benešová, M. (2016). Text segmentation for Menzerath-Altmann law testing. Palacký University, Faculty of Arts.

Biber, D. (2012). Register as a predictor of linguistic variation. Corpus Linguistics and Linguistic Theory, 8(1), 9-37.

Campos, J., Otero, P., & Loinaz, I. (2020). Measuring diachronic language distance using perplexity: Application to English, Portuguese, and Spanish. Natural Language Engineering, 26(4), 433-454. doi:10.1017/S1351324919000378.

Chao, Y. R. (1968). A Grammar of Spoken Chinese. Berkeley and Los Angeles: University of California Press.

Chen, H. H. (1994). The contextual analysis of Chinese sentences with punctuation marks. Literary and linguistic computing, 9(4): 281-289.

Chen, Keh-jiann, Chu-Ren Huang, Li-ping Chang, and Hui-Li Hsu. (1996). Sinica Corpus: Design Methodology for Balanced Corpora. In. B.-S. Park and J.B. Kim. Eds. Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation. Seoul:Kyung Hee University. pp. 167-176.

Chen, Keh-Jiann, Chi-Ching Luo, Ming-Chung Chang, Feng-Yi Chen, Chao-Jan Chen, Chu-Ren Huang, and Zhao-Ming Gao. (2003). Sinica Treebank: Design Criteria, Representational Issues and Implementation. In Anne Abeillé (Ed.), Treebanks: Building and Using Parsed Corpora (pp. 231-248). Dordrecht; Boston: Kluwer Academic Publishers.

Cramer, I. (2005). The parameters of the Altmann-Menzerath law. Journal of Quantitative Linguistics, 12, 41–52.

Degaetano-Ortlieb, S. and Teich, E. (2018). Using relative entropy for detection and analysis of periods of diachronic linguistic change. In Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 22–33.

Grzybek, P. (2007). Do we have problems with Arens' law? A new look at the sentence-word relation. In P. Grzybek and E. Stadlober. Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of His 75th Birthday, 62, 205.

Hou, R, Chu-Ren Huang, Hue San Do & Hongchao Liu (2017): A Study on Correlation between Chinese Sentence and Constituting Clauses Based on the Menzerath-Altmann Law, Journal of Quantitative Linguistics. 24(4): 350-366.

Hou, R., C.-R. Huang, M. Zhou & M. Jiang. (2019). Distance between Chinese Registers Based on the Menzerath-Altmann Law and Regression Analysis. Glottometrics. 45: 24-56.

Huang, Chu-Ren and Shi, D. (2016). A Reference Grammar of Chinese. Cambridge: Cambridge University Press.

Huang, C.-R. & K.-J. Chen. (2017). Sinica Treebank. In N. Ide and J. Pustejovsky (eds), Handbook of Linguistic Annotation. Berlin & Heidelberg: Springer.

Hřebíček, L. (1992). Text in communication: Supra-sentence structure. Bochum, Brockmeyer.

Hřebíček, L. (1995). Text levels: Language constructs, constituents and Menzerath-Altmann law. Trier: WVT.

Hřebíček, L. (1997). Lectures on text theory. Prague: Academy of Sciences of the Czech Republic, Oriental Institute.

Köhler, R. (1984). Zur Interpretation des Menzerathschen Gesetzes. In W. Lehfeldt & U. Straus (Eds.), Glottometrika 6, 177-183. Bochum: Brockmeyer.

Köhler, R. (2012). Quantitative syntax analysis (Vol. 65). Berlin: Walter de Gruyter.

Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A. Nowak. 2007. Quantifying the evolutionary dynamics of language. Nature 449(7163): 713-716.

Lu, J. (1993). The features of Chinese sentences. Chinese Language Learning. No. 1:1-6.

Lu, LW-L & K. Ahrens. (2008). Ideological influence on BUILDING metaphors in Taiwanese presidential speeches. Discourse & Society 19 (3): 383-408.

Menzerath, P. (1954). Die Architektonik des deutschen Wortschatzes (Vol. 3). F. Dümmler.

Millar, R.M. and Trask, L. (2015). Trask’ s Historical Linguistics. Abingdon-on-Thames: Routledge.

Motalová, T., Spáčilová, L., Benešová, B., Kučera, O. (2014). An application of Menzerath-Altmann law to contemporary written Chinese. Křížkovského, Olomouc: Univerzita Palackého v Olomouci.

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram,B.D., Köhler,R., Krupa,V.,Maˇcutek, J., Pustet, R., Uhlíˇrová, L., Vidya, M.N. (2009). Word frequency studies. Berlin-New York: Mouton de Gruyter.

Randour, F., Perrez, J., & Reuchamps, M. (2020). Twenty years of research on political discourse: A systematic review and directions for future research: Discourse & Society, 31(4), 428–443.

Roitman, M. (2014). Presidential candidates’ ethos of credibility: The case of the Presidential Pronoun I in the 2012 Hollande-Sarkozy Debate. Discourse & Society 25(6): 741-765.

Ščigulinská, J. & Schusterová, D. (2014). An Application of the Menzerath-Altmann Law to Contemporary Spoken Chinese. Palacký University in Olomouc. First Published 2014.

Tuldava, J. (1995). Informational measures of causality. Journal of Quantitative Linguistics, 2(1), 11-14.
Van Dijk TA. (1997). What is political discourse analysis? Belgian Journal of Linguistics 11: 11–52.
Wang, J. (2017). Representing Chinese Nationalism/Patriotism through President Xi Jinping's "Chinese Dream" discourse. Journal of Language and Politics. 16(6): 830-848.
Wang, Y and H. Liu. (2018). Is Trump always rambling like a fourth-grade student? An analysis of stylistic features of Donald Trump's political discourse during 2016 election. Discourse & Society 29(3):299-323.
Wang, K., & H. Qin. (2014). What is peculiar to translational Mandarin Chinese? A corpus-based study of Chinese constructions' load capacity. Corpus Linguistics and Linguistic Theory, 10(1), 57-77.
Wodak, R. & S. Boukala. (2015). European identities and the revival of nationalism in the European Union. Journal of Language and Politics 14(1): 87-109.
Yu, B. (2008). Classifying party affiliation from political speech. Journal of Information Technology & Politics. 5(1): 33–48.
Yu, B. (2013). Language and gender in congressional speech. Literary and Linguistic Computing. 29(1): 118–132.
Zhu, D. (1982). Lectures on Grammar. Beijing, China: Commercial Press.

Appendix 1:

Table：The fitted result of average word lengths in 15 Report texts from three periods

	a	*b*	R^2
1978	2.452905	-0.14949	0.792514
1979	2.517411	-0.15424	0.817301
1980	2.534983	-0.15377	0.901367
1981	2.375607	-0.14064	0.822611
1982	2.536153	-0.15683	0.953379
1997	2.53746	-0.15027	0.80805
1998	2.566591	-0.15172	0.862896
1999	2.420279	-0.13037	0.865351
2000	2.480595	-0.13977	0.91811
2001	2.351106	-0.10954	0.925222
2016	2.708322	-0.19148	0.871852
2017	2.640937	-0.17755	0.860396
2018	2.770532	-0.20207	0.809142
2019	2.771229	-0.2077	0.938813
2020	2.532724	-0.16393	0.891272