Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-022-01488-8 OPEN

Regional varieties and diachronic changes in Chinese political discourse

Renkui Hou[1]✉, Chu-Ren Huang[2]✉ & Kathleen Ahrens[3]✉

The present paper explores the synchronic variations and diachronic changes in political discourses in Hong Kong (HK) and in Mainland of People's Republic of China (PRC). The relationship between lengths of linguistic constructs and their immediate constituents (including sentences and clauses, and clauses and words) are fitted using the function $y = ax^b$ based on the Menzerath–Altmann (MA) law to capture the characteristics of language as self-organizing complex systems. We found that the two fitted parameters a and b, as distinctive characteristics of complex systems, can distinguish two regional variants of political speeches from HK and PRC over different periods in time. We also found that the same parameters can capture language changes between different periods of political speeches from the PRC. More specifically, we found that regional variations and historical changes show different degrees of salience at different constituency levels. In addition, we found compounding effects between historical change and regional variations. That is, the two regional variants of political speeches are closer to each other at the earliest diachronic period as compared with the latter two periods, as represented by the fitted parameters of the relationship between sentence and clause lengths. Our results provide strong support for the hypothesis for the MA Law capturing the characteristics of language as a complex self-organizing system, as the two fitted parameters account for the interaction of diachronic language change and synchronic variation.

[1] School of Humanities, Guangzhou University, Guangzhou, China. [2] Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China. [3] Department of English, The Hong Kong Polytechnic University, Hong Kong, China. ✉email: hourk0917@163.com; churen.huang@polyu.edu.hk; kathleen.ahrens@polyu.edu.hk

Introduction

Language is a dynamic, complex system that constantly adapts to its environment (Wang, 2006; Beckner et al., 2009). One way to model such complex systems is through the dynamic interaction between the individual components and the collective system, as the behaviors of neither the individual components nor the collective system are predictable in isolation (Holland, 1996). In other words, models can capture a system in terms of the balance maintained between individuals and a collection of individuals. In addition, complex systems may be organized hierarchically, such as the constituent relations seen in language, and the dynamic relations between an element and its constituents are posited to follow the Menzerath–Altmann (MA) law. That is, the length of a linguistic construct interacts with the length of its constituents, and their correlation must remain constant within a linguistic system. Paul Menzerath summarized it as “The larger the whole the smaller the parts” after he detected the dependency of syllable length on word length (Menzerath, 1954, p. 101). Altmann generalized this hypothesis to all language levels, formulating it as “The longer a language construct, the shorter its components” (Altmann, 1980). Interestingly, the MA law is shown to be applicable to biologically complex systems when the construct-constituency relation is interpreted as the system–component relation (e.g. Shahzad et al., 2015).

Altmann (1980) gave the theoretical derivation and the corresponding differential equation of the MA law, as shown in the following equation:

$$\frac{y'}{y} = -c + \frac{b}{x}$$

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

$$y = ax^{b}\mathrm{e}^{-cx} \tag{1}$$

In this formula, y is the mean size of the immediate constituents, x is the size of the construct, and parameters a, b, and c primarily depend on the levels of the units under investigation.

Many observations have shown that parameter c is close to zero for higher levels of language, whereas lower levels tend to have very small values of parameter b. Therefore, only intermediate levels require the full formula (Köhler, 2012, p. 149). Formula (2) has become the most commonly used “standard form” for linguistic purposes (Grzybek and Stadlober, 2007). Köhler (1984) interpreted the parameters in Formula (2) and assumed that a represents a quantity that is dependent on the language and language levels, while b might represent a shortening tendency and describe the range of structural information that has to be stored for each language component. Köhler (1989) proposed that the mechanism of shortening is a consequence of memory limitations: the longer the construct, the more space must be reserved for structural information between the constituents, therefore, the size of the constituents must be reduced. Cramer (2005) showed that parameters a and b might be correlated.

$$y = ax^{b} \tag{2}$$

In this study, the MA law is interpreted as a description of Holland’s (1996) ‘hidden order.’ The fact that parameters a and b remain constant for a specific constituency relation in a complex system suggests that they represent an equilibrium state that is the characteristic of the system and a norm that ‘guides’ self-organization. In other words, regardless of whether the construct or the constituents change first, the adaptation results in maintaining the parameters constant. Previous studies have applied the MA law to diachronic changes (e.g. Jiang et al., 2020) or synchronic variations (e.g. Hou et al., 2019a, 2019b; Chen and Liu, 2022), but not to both changes and variations simultaneously.

The current study Is among the first to investigate the complex dynamics of changes and variations. This study examines two comparable domain-specific diachronic corpora in order to compare language change and variation in a controlled manner. The two political corpora in Chinese are composed of two regional variants from the Mainland of the People’s Republic of China (PRC) and Hong Kong SAR(HK), respectively. This study aims to show that the MA law holds not just for historical changes or regional variations but is also able to account for their interactions.

Synchronic and diachronic perspectives of language. Traditional linguistic research on political discourse has primarily focused on either a synchronic or a diachronic perspective. An example of a synchronic study is Savoy (2018), who examined the verbal style and rhetoric of the candidates of the 2016 US presidential primary elections and showed that Trump adopted a more direct communication style, selecting simple words and producing short sentences compared with other candidates. Yu (2013) revealed that feminine political figures tend to use emotional words more frequently and employ more personal pronouns than masculine political figures. Randour et al. (2020) also identified that the variation of political speech should be observed across continents, policy domains, sub-genres, forms of discourse, and various actors that speak the language of politics.

Other studies have focused on political issues, events, and actors, including ideology or identity construction over time (Lu and Ahrens, 2008; Wang, 2017; Wodak and Boukala, 2015) or concentrated on specific linguistic characteristics of the discourses, such as Wang and Liu’s (2018) analysis of Trumpian discourse, or Roitman’s (2014) study of the use of pronouns by French presidential candidates. Moreover, additional studies apply quantitative methodologies to diachronic change. Burgers and Ahrens (2020), on the other hand, undertook an examination of diachronic change in TRADE metaphors in over 200 years of US State of the Union Addresses and found certain types of metaphors are more likely to be used over longer time periods. Jiang et al. (2020) investigated quantitatively whether Queen’s English has experienced diachronic changes and drifted towards common people’s English; the indices they used include lexical richness and complexity. Savoy (2018) summarized that the temporal period of the documents constitutes an important factor in explaining the variations between presidents or prime ministers after reviewing the related studies on political texts. Kubát and Cech (2016) examined vocabulary richness, secondary thematic concentration, and text activity, as the indices, in US presidential addresses through more than two centuries and found individual variations by different presidents but no diachronic changes.

Recent studies focusing on similar languages pose an interesting challenge in terms of changes versus variations, as the diversity of similar languages is likely due both to language variations (synchronic change) and changes over time (diachronic change) (Zampieri and Nakov, 2021). In fact, Thomason (1997) argues that diachronic change always leads to language variations. Xu et al. (2022), on the other hand, suggest that language variations can lead to language changes over time, and language varieties can be the result of localized language changes, based on their study of light verbs in Mainland and Taiwan variations of Chinese.

In light of the results of the above two groups of studies, we observe that diachronic changes and synchronic variations interact and should not be studied in isolation. This suggests a self-organizing complex systems approach that can model both language changes and variations, as well as their interactions.

The corpus we will look at in this study provides a novel approach to this issue, as it covers a period of historical changes in Hong Kong from 1984 to 2014. Since Hong Kong was returned to the People's Republic of China as a Special Administrative Region (SAR) in 1997 from British colonial rule, it has undergone significant changes in the political and other systems. Interestingly, in both the colonial period and the SAR period, Hong Kong is run by government officials outside of the central political leadership in London and Beijing, respectively. This unique situation allows for a synchronic contrast of the language used by leaders in HK (colonial leaders or SAR leaders) with leaders in the PRC. Recent work (Ahrens and Zeng, 2022) has shown that leaders in China and HK have employed metaphors of EDUCATION in major government speeches and reports before and after HK's handover in different ways, which suggests that it would be valuable to examine language change and variation further with respect to the MA Law.

Language as a complex system. Language as a complex system is fertile ground for the study of the MA law because of its well-analyzed layers of unit-constituent relations, ranging from phonemes to syllables, morphemes, words, phrases, clauses, and sentences (Lyons, 1968). That is, the MA law can be used to measure the complexity of relationships between language units at various levels (Altmann, 1980; Torre et al., 2021), or to provide empirical evidence for the relations between two linguistic levels (Köhler, 1993).

Previous research has validated the MA law between a linguistic unit and its constituents at different linguistic levels. For example, Köhler (1982) conducted the first empirical test to confirm the MA law based on the interaction between sentences and clauses. Buk and Rovenchak (2008) analyzed the correlation between clause length (measured both in words and in syllables) and sentence length (measured in clauses). They found that Formula (3), one of the generalizations of MA law (Wimmer and Altmann, 2005), best predicts the results. Benešová (2016) tested the potential validity of the MA law on samples in different languages in order to test the MA law as a language universal law.

$$y = ax^{b}e^{-c/x} \tag{3}$$

In more recent literature, Xu and He (2020) studied the relationship between English sentences and clauses based on the MA law in academic spoken and written registers and showed that the fitted parameters can differentiate these two registers. Jiang and Ma (2020) showed that the relationship between sentence and clause in original and translated texts abided by the MA law and the fitted parameters could differentiate the translational language from the original. They also identified that the parameter values of the MA law in Formula (2) could be used in automatic text classification. Berdicevskis (2021) performed a study of MA law at two syntactic levels based on the data from the Universal Dependencies collection. He showed that the MA law works well at the sentence–clause–word level. Mačutek et al. (2021) proposed the linear dependency segment as a new linguistic unit by showing that the relation between sentence length in clauses and clause length in linear dependency segments abides by the MA law in two Czech dependency treebanks.

Additional studies focus on other constituency levels, most often at the clause-word level. Tuldava (1995) found a highly significant statistical dependence between average word length and clause length. Hou et al. (2019b) fitted the relationship between clause and word length based on Formula (2) and proposed an index to represent the formality degree of Chinese registers and to calculate the distance between different registers. Hou et al. (2020a) fitted the relationship between clause and word lengths based on Formula (1) and showed that the three fitted parameters could differentiate various Chinese registers. Hou et al. (2020b) demonstrated that the relationship between clause length and word length in PRC political speeches from different periods (1978–1982, 1997–2001, and 2016–2020) abides by the MA law and the fitted parameters undergo diachronic changes.

In the meantime, some studies examine variations of the hierarchical relationships between language units at different levels. Chen and Liu (2022) showed that the MA law fittings at the 'sentence > clause > word' levels outperformed the 'clause > word > character' levels and classified the texts from *Press* and *Science* registers using the fitted parameters based on the MA Law. Lastly, Fenk-Oczlon and Pilz (2021) found significant and complex correlations between the length of linguistic units and the number of words, phonemes, and population. Their findings indicated that the length of a language unit is a crucial mediating variable in the hierarchical and synergetic linguistic system. Overall, these studies showed the versatility of the application of the MA law at various constituency levels without directly addressing the optimal model of the complex system involving multiple hierarchical levels.

In language, the sentence and the word are generally considered to be the two most salient and intuitive levels. Therefore, it is not surprising that some of the earliest studies on inter-level correlations looked at the interaction between texts or sentences and words, even though, strictly speaking, they do not have a direct unit-constituency relation. These include the well-known Zipf's law (Zipf, 1935) for text-word relation and Arens (1965) for sentence–word relation. Grzybek and Stadlober (2007), in fact, reported that Altmann's (1980) study was inspired by and in reaction to the proposed relation between sentence length (word quantity) and word length by Arens (1965).

Another possible reason that early research has focused on the interaction of words with large text units is that these two levels are easily accessible in raw text, while other intermediate levels require annotation to identify. This explains why the first studies of Zipf's law in Chinese use characters instead of words, a convention established by Zipf's own study (1949). The character is the natural choice since Chinese, unlike most other writing systems, conventionally marks the boundary of characters instead of words. Similarly, earlier work on the MA law in Chinese also focused on characters (e.g. Prün, 1994). Chen et al. (1993), a study based on the word-segmented Sinica Corpus, was probably the first study to show that Zipf's law predicts the distribution of Chinese words better than Chinese characters. These earlier works showed that the choice of linguistic units to model language as a complex system is far from trivial.

Lastly, a recent series of papers is the first to systematically investigate the correlation models between different levels of linguistic constituency in Chinese, covering the character, the word, the clause, and the sentence levels. For example, Hou et al. (2017) applied the MA law to the sentence–clause correlation in different Chinese register texts and showed that the sentence–clause relationship abides by the MA law in written formal registers, but not in daily informal registers; Hou et al. (2019b) applied the law to the clause–word level. Hou et al. (2019a) systematically compare the fitted models of register classification based on sentence–word/character, sentence–clause, and clause–word/character correlations. They concluded that: (1) Immediate constituency relations (i.e. sentence–clause and clause–word/character) capture the differences between the related complex systems of different registers. (2) Characters as sociological words (Chao, 1968) are a good imitation of linguistic words and are provided as a plausible alternative in classification, with a slight loss in the results. (3) The clause–word relation provided slightly better results. These studies laid the ground for applying the MA law to the interaction of adaptation at various constituent levels.

Given the well-attested application of the MA law to the correlation of the immediate constituency levels for modeling language as complex systems, there are still two unanswered questions: First, can the MA law model the interaction of different kinds of adaptation? Second, do different levels of immediate constituency correlations model different aspects of the complex systems? In order to approach these two modeling issues, we study a corpus of a specific domain that has both diachronic changes and synchronic variations.

Research questions. Our first research question (RQ1) is to examine whether the MA law can model both regional variations and diachronic changes in PRC and HK political speeches. Our second research question (RQ2) is to examine whether fitted parameters of the MA law can differentiate variations and changes in PRC and HK political speeches. In other words, are the six systems, at three different times and two different locations, equally distinct? If not, what are the characteristics of the model that identify the same location or the same time?

Why are we interested in the possible distinctions between language changes and variations in a model of language as a self-organizing complex system? Language (both written and spoken) are the most frequently and thoroughly documented human behaviors. They are also, at the same time, direct or indirect documentation of other human behaviors. Close correspondences between linguistic variations and events that cause variation in human behavior have been established. The correlation between the types of linguistic variations and the causes of human behavior changes, however, may vary based on the preceding event.

For instance, the changes in the use of models marking different power relations between speakers and addressees have been shown to reflect the changes in the power dynamics of a society, e.g. Winter and Gärdenfors (1995), Leech (2012), Millar (2009), and Wang et al. (2022). Linguistic encoding of weather, for example, has been shown to vary typologically. Huang et al. (2021) and Dong et al. (2021) showed that the variations in the use of different linguistic devices to encode kinesis in China correspond to the actual distribution of weather patterns in China based on the different kinetic energy of the weather events. Su et al. (2021) showed that historical and geographic variations in professional gender segregation in China can be mapped to the use (or lack) of gender modification of professional terms. Wang et al. (2022) showed that the use of different speech act constructions pragmatically reflects the different social dynamics in two culturally different societies: Guangzhou and Hong Kong. Overall, this line of research has produced interesting results in terms of how language encoding corresponds to collective human behavior changes. However, this line of research has yet to establish a model of how language changes and varies as a complex system.

The above results are consistent with the macro-view that collective human behaviors in general, and language in particular, are complex self-adaptive systems, and can be modeled as such. Beckner et al. (2009) assume that language is a complex adaptive system where both internal and external factors contribute to self-organization. This complex adaptive system view provides a direct way to model language variations and changes (Wang, 2006). From the complex system approach, the productive work hitherto has focused on how to model and identify/classify different specific variations. As good as the MA law and other models are in identifying changes in complex systems, the literature so far has provided little study of the interaction between multiple kinds of adaptation. This paper, hence, takes the first step to fill this gap by carefully examining the fitted MA law at different constituent levels and for both diachronic change and synchronic variation.

Data and methodology

The HKBU Corpus of Political Speeches (Ahrens, 2015) is an important resource for those interested in the study of political rhetoric. In this study, we select from the policy addresses of HK Colonial Governors (from 1984 to 1996) and Hong Kong Special Administrative Region Chief Executives (from 1997 to 2014), and the report on the Work of the Government by Premiers of the People's Republic of China (from 1984 to 2013).

Hong Kong, which was a colony of Great Britain for over one hundred years, was returned to China in 1997. The HKBU Corpus of Political Speeches was created to represent this point in time. We extracted data from three representative periods, each of which are fives year long, and approximately 10 years apart diachronically. The first sample from roughly 10-year prior to 1997 (namely 1984–1988) constitutes the British period; the second sample (1997–2001) from the first years of Hong Kong as a Special Administrative Region (SAR) of China; and the last sample (2010–2014) from about 10 years after the handover. The corpus of political speeches on the Work of the Government by Premiers of the PRC was also created according to the corresponding periods (i.e., 1984–1988, 1997–2001, and 2009–2013). In this study, we focus on the HK political speeches written in Chinese in order to compare them to PRC political speeches, and be treated as two varieties of Chinese.

The relationship between linguistic constructs and their immediate constituents are explored based on the MA law in this study. It is necessary to define the sentence, clause and word in Chinese. The sentence is considered to be a basic expression unit in all languages and has its own problems, for example, the difference between a word and a sentence can be quite fuzzy in polysynthetic languages. There is no such problem in Chinese. In contrast to Indo-European languages, it is difficult to define the terms "sentence" and "clause" in Chinese as sentences/clauses in Chinese are often defined in terms of characteristics of speech rather than texts (Huang and Shi, 2016; Chao, 1968; Zhu, 1982). In the written Chinese language, characteristics of speech are marked by punctuation at the end of a sentence. Thus, the sentence is defined as the Chinese text separated by periods, exclamation marks, and question marks operationally (Hou et al., 2017, 2019a; Chen and Liu, 2022). The sentence length is measured in the number of its immediate constituents, clauses (Hou et al., 2019a; Chen and Liu, 2022). In Köhler (1982), the number of clauses is determined by counting the number of finite verbs in a sentence. In this study, however, the number of clauses is determined by the number of commas, semicolons or colons in a Chinese sentence (Hou et al., 2017; Chen and Liu, 2022).

The minimal value of sentence length is one if none of the above-mentioned punctuation is present in a sentence according to this definition. For example, in the following sentence, there are three commas, three semicolons and one colon, thus, this sentence is composed of eight clauses.

政府/NN 在/P 小心/AD 考慮/VV 過/AS 各/DT 方面/NN 提出/VV 的/DEC 觀點/NN 和/CC 意見/NN 之後/LC, /PU 決定/VV 實施/VV 四/CD 項/M 改革/NN; /PU 一/CD 增加/VV 民選/NN 區議員/NN 的/DEG 數目/NN, /PU 以/MSP 進一步/AD 增强/VV 區議會/NR 的/DEG 代表/NN 地位/NN; /PU 二/OD 荃灣/NR 及/CC 葵涌/NR / /PU 青衣/NR 各自/AD 成立/VV 本身/PN 的/DEG 區議會/NR; /PU 三/CD 擴大/VV 區議會/NR 在/P 管理/VV 地方/NN 設施/NN 方面/NN 的/DEG 功能/NN;

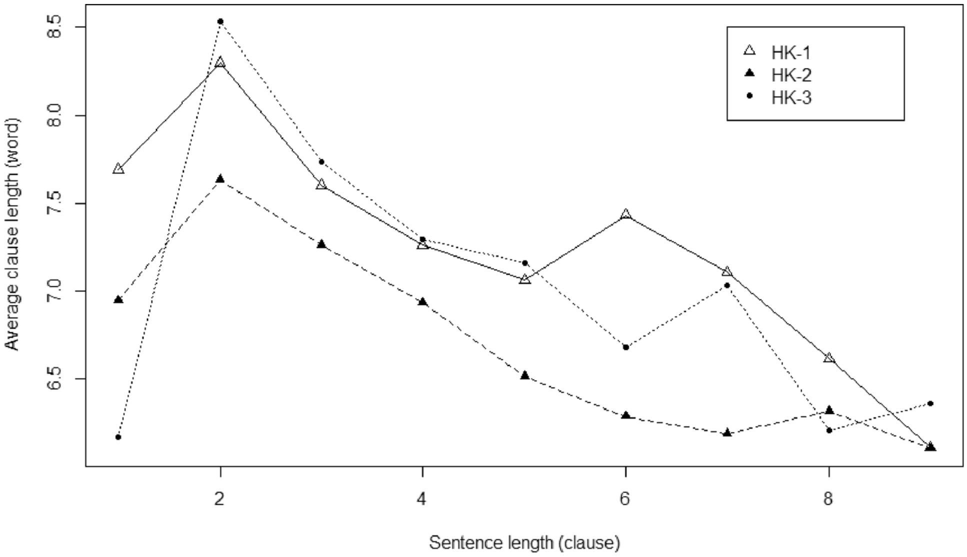

Fig. 1 The average clause length distributions in sentences in HK political speeches. HK-1, HK-2, and HK-3 represent 1984–1988, 1997–2001, and 2010–2014 Hong Kong political speeches, respectively.

/PU 四/CD 爲/P 市政局/NR 服務/NN 範圍/NN 以外/LC 的/DEG 地方/NN, /PU 成立/VV 一/CD 個/M 新/VA 的/DEC 區域議局/NR 。/PU

After carefully considering the views and opinions put forward by various parties, the government decided to implement four reforms: (I) Increasing the number of elected District Board members to further enhance the representative status of district boards; (II) Tsuen Wan and Kwai Chung/Tsing Yi each set up their own district councils; (III) Expanding the functions of District Councils in the management of local facilities; and (IV) To set up a New Regional Council for places outside the scope of the Urban Council.

The definition of wordhood and the operational criteria for identifying word boundaries have been hotly debated and have received ongoing research attention (e.g., Huang and Xue, 2012, 2015). The current study follows the word segmentation results of the HKBU corpus (Ahrens, 2015). Clause length is defined as the number of words in a clause (Hou et al. 2019b, 2020a; Chen and Liu, 2022). Word length is defined as the number of syllables in a word.[1]

The numbers of sentences, clauses, and word tokens in PRC and HK political speeches are 16,677, 44,743, and 33,3273, respectively, as shown in Table 1 in Appendix.

The relationships between linguistic units and their immediate constituents (i.e., sentence and clause, and clause and word respectively) were studied and fitted by Formula (2) in PRC and HK political speeches (sections "Relationship between sentence and clause" and "The relationship between clause and word "). Mačutek et al. (2021) showed that there should be an intermediate linguistic unit between clause and word in Czech dependency treebanks. However, no such intermediate level has been attested in Chinese and the existing literature showed that the clause-word constituency relation is well fitted by the MA Law (e.g. Hou et al., 2017; Chen and Liu, 2022). Thus, the word is considered to be the immediate constituent of the clause and clause and word are considered to be immediate neighbors in Chinese. The texts that are analyzed in this research can be represented by the fitted parameters, *a* and *b*, and are displayed in one two-dimensional space in order to explore the relationship between political speeches from different periods and regions.

In the process of fitting the relationship between linguistic units and their immediate constituents, Formula (2) shows that this function is nonlinear; however, it can be transformed into a linear function in order to avoid the impact of the initial parameter estimates on the fitted result.

$$y = ax^b \tag{2}$$

Taking the logarithm of both sides of Formula (2) gives

$$\ln(y) = \ln(a) + b\ln(x)$$

Then, defining

$$Y = \ln(y);\ X = \ln(x)$$

The linear function stated is obtained, as shown in Formula (4):

$$Y = bX + \ln(a) \tag{4}$$

The determination coefficient (R^2) was used to validate the fitted results of this linear regression as like residual sum-of-square for the validation of nonlinear regression result; it shows the goodness-of-fit of the model to the empirically collected data, indicating the proportion of variance in the data that can be explained by the model (Conway and White, 2012; Baayen, 2008). In quantitative linguistics, a fit is generally considered very good if R^2 is greater than or equal to 0.9 (Popescu et al., 2009, p.16; Chen and Liu, 2022). A fit with $R^2 > 0.8$ is good and a fit with $R^2 > 0.75$ is accepted (Chen and Liu, 2022).

Relationship between sentence and clause

In this section, we will determine whether the relationship between sentence and clause in PRC and HK political speeches can be fitted by Formula (2) and abides by the MA law. If so, we will further explore our first research question of whether there are synchronic variations of fitted parameters in PRC and HK political speeches, and whether the fitted parameter shows the diachronic change in different periods.

The correlations between average clause lengths and sentence lengths in HK political speeches during three periods were established, as shown in Fig. 1.

From Fig. 1, the average clause length in 1-clause sentences from HK political speeches is significantly smaller than that in 2-clause sentences in each period. In other words, the MA Law does not apply when 1-clause sentences are included. The relative occurrence frequency of sentences with 1–5 clauses in 1984–1988 HK political speeches is 97.2%. The relative occurrence

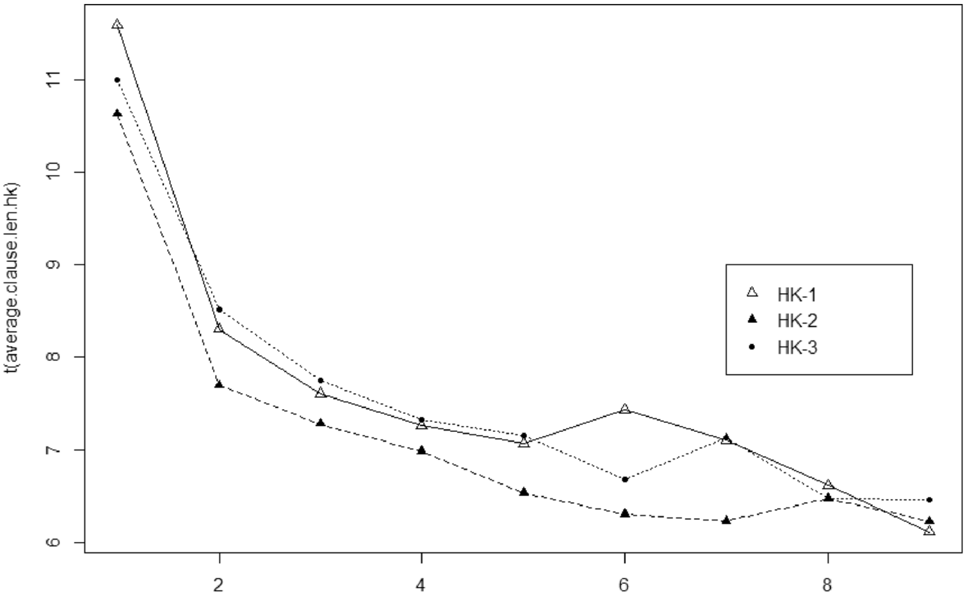

Fig. 2 The average clause length distributions in HK political speeches (headlines not included). HK-1, HK-2, and HK-3 represent 1984–1988, 1997–2001, and 2010–2014 Hong Kong political speeches, respectively.

frequencies of sentences with 1–6 clauses in HK political speeches from 1997–2001 and 2010–2014 are 97.06% and 98.11%, respectively. This means that most of the sentences in HK political speeches are composed of 1–6 clauses.

Figure 1 also shows that the average clause length decreases when only sentences with more than 1 clause are considered in HK political speeches. There is an obvious decreasing tendency of average clause length in 2–5, 2–7, and 2–6 clause sentences in 1984–1988, 1997–2001, and 2010–2014 HK political speeches, respectively. Both the relative occurrence frequencies of 2–5 clause sentences in 1984–1988 HK political speeches and 2–7 clause sentences in 1997–2001 HK political speeches are 75.3%, respectively. The relative occurrence frequency of 2–6 clause sentences in 2010–2014 HK political speeches is 68.9%. That is, for about 70% of the sentences (except for 1-clause sentences) from HK political speeches, the average clause length decreases with sentence length.

In order to study the MA law at all pairs of constituency levels, full sentences are required. We noticed that in HK political speeches the 1-clause sentences in our initial analysis are high in frequency, but are exceptionally short in length. Careful checking revealed that most of these are headlines and that all of the headlines are phrases and not complete sentences, as in (1)–(3) below.

(1) 伍〇/CD 、/PU 基本/JJ 設施/NN 的/DEG 發展/NN

50. Development of Infrastructure

(2) (/PU 乙/OD)/PU 土地/NN 供應/NN

(B) Supply of Land

(3) (/PU 戊/OD) /PU 公用/JJ 事業/NN

(E) Public Utilities

There are two options to model the constituency relations of these headlines. The first is to treat them as one-clause sentences. The second is to treat them simply as stand-alone clauses that do not form full sentences. The deciding factor between these two options is most likely a stylistic one: i.e. whether full sentences are allowed in the headlines. In addition, we observed that the headlines have a disproportional share of one-clause sentences in HK. For example, there are 70 headlines among 191 one-clause sentences in 1984, and 49 headlines among 144 one-clause sentences in 1985. In addition, since all of the headlines in the Hong Kong data consist of single phrases, without exception, it is reasonable to assume that the stylistic guide in the HK government does not allow full sentences to appear as headlines. Therefore, we propose that the headlines in HK political speeches should not be considered sentences.

In Fig. 1, the default hypothesis that all the headlines are sentences was adopted when calculating the average clause length. The average clause length distributions in sentences of HK political speeches based on the hypothesis that headlines are clauses but not sentences are shown in Fig. 2.

From Fig. 2, it can be seen that the average clause length decreases with the increase in sentence length. Different from Fig. 1, the average clause lengths of one-clause sentences are compatible with the prediction of the MA law when the headlines are considered as clauses only.

Overall, the average clause length shows a decreasing tendency when the headlines from HK political speeches are not treated as sentences, although there are some small discrepancies during specific periods of time. For example, the average clause length in 5-clause sentences is slightly less than in 6-clause sentences in 1984–1988. From the sentence length distributions, the sum of relative occurrence frequencies of sentences including 1–8 clauses in the HK political speeches from three periods were 99.81%, 98.58%, and 99.23% respectively. Based on the suggestion that the MA law is cognitively motivated by short-term memory constraints (Köhler, 2012), our result could be predicted by Miller's (1956) seven plus/minus two rule for the capacity of short-term memory.

Formula (2) was used to fit the relationship between lengths of sentences and clauses (average clause length distributions) in HK political speeches from three periods. The fitted results are shown in Table 1 and Fig. 3. The determination coefficients (R^2) in Table 1 showed that the fitted results of average clause length distributions in HK political speeches from 1984–1988 and 1997–2001 are good and are very good in 2010–2014.

The black and red dots represent the observed and fitted values of the average clause length, respectively, in Fig. 3 (and similarly in the figures that follow). Figure 3 shows that the fitted values of average clause length in HK political speeches are similar to the observed values, which also shows that the fitted results are good.

The mean clause length distributions in sentences in 15 HK political speeches from different periods were fitted using

Formula (2) based on MA law. The fitted results were good and are shown in Table 2 in Appendix. The HK political speeches are represented by the fitted parameters *a* and *b* and are displayed in a 2-Dimensional plane, as shown in Fig. 4. It is difficult to see an obvious regular diachronic change of these two fitted parameters in HK political speeches from the first period to the third period.

Similarly, the average clause length distributions were established in different periods of PRC political speeches, as shown in Fig. 5. Unlike HK political speeches, there are few headlines, most of which are sentences. For example, there are only 10 headlines among 91 sentences in 1984 PRC political speeches and 8 of them are full sentences. This shows the PRC style convention allows full sentences to appear as headlines. Thus, all headlines are treated as sentences when calculating the average clause length in sentences. The relative occurrence frequencies of sentences with 1–8 clauses in PRC political speeches from these three periods are 98.16%, 99.11%, and 98.57%, respectively. The average clause length decreases with sentence length in 1–8 clauses sentences from Fig. 5. For sentences with the same length, the average clause length in the first period (1984–1988) of PRC political speeches is larger than that in the other two periods. The average clause length has a slower downward trend in PRC political speeches from 2009 to 2013.

Formula (2) was used to fit the average clause length distributions in sentences in PRC political speeches. The fitted results are shown in Fig. 6 and Table 2. The black and red dots represent the observed and fitted values of the average clause length in PRC political speeches, respectively, in Fig. 6, which shows that the fitted values of the average clause length are similar to the observed values. This means that the relationships between sentences and clauses in PRC political speeches from three periods can be fitted by Formula (2). The determination coefficients (R^2) shown in Table 2, which are more than 92% in the three periods respectively, also show that the fitted results are all very good.

Table 1 The fitted results of the average clause length distribution in sentences in HK political speeches.

	a	*b*	R^2
1984–1988	10.581	−0.237	87.69%
1997–2001	9.707	−0.225	89.59%
2010–2014	10.361	−0.227	93.76%

The fitted results show that the relationship between sentence length and clause length in PRC political speeches abides by the MA law. This conclusion is consistent with Hou et al. (2017), which showed that the relationship between sentence and clause lengths in Chinese formal written registers abides by the MA law. The value of fitted parameter *b* in 2009–2013 PRC political speeches is more than in the other two periods of PRC political speeches, which means that the decreasing tendency of average clause length is small and the range of structural information for each clause of sentences that need to be stored for processing is also small and the link between clauses in a sentence is becoming weak.

Then, Formula (2) was used to fit the average clause length distributions in sentences from 15 PRC political speeches during three time periods. The fitted results are shown in Table 3 in Appendix.

The fitted result in the 1985 PRC political speech is not satisfactory from R^2, 53%. The fewer number of sentences (272 in total) may be one of the factors that led to this unsatisfactory fitted result. Except for 1985, the fitted results of the relationships between the lengths of sentences and clauses in the third period PRC political speeches (2010, 2012, and 2013) are not as good as in the other two periods, with the determination coefficients 62.20%, 64.23%, and 67.68% respectively. Hou et al. (2017) demonstrated that the relationship between sentence and clause does not abide by the MA law in daily informal registers (*Sitcom conversation* and *TV Talk Show*), which are colloquial. This may be the result of PRC political speeches becoming more colloquial in nature. In addition, the not-so-good-fitted results may also be caused by the fact that the word seems not to be directly “below” the linguistic unit of the clause as in Czech in Mačutek et al. (2021). Hou et al. (2021) showed that the relationship between sentence and clause lengths in Chinese political discourses from 2016–2020 was not fitted by Formula (2) and validated the first hypothesis; PRC political speeches are becoming more colloquial. The colloquial register narrows the distance between communicators and sounds friendlier to people. This is consistent with the findings of Jiang et al. (2020), which showed the Queen’s English has experienced diachronic changes and drifted towards common people’s English using quantitative indices.

Except for 1985, each PRC political speech was represented by the fitted parameters. The relationship between different periods of PRC political speeches are shown in Fig. 7. From Fig. 7, it seems that the values of *a* in fitted results from the first period PRC political speeches are larger than in the other two periods. The *t*-test was used to validate whether the difference between the

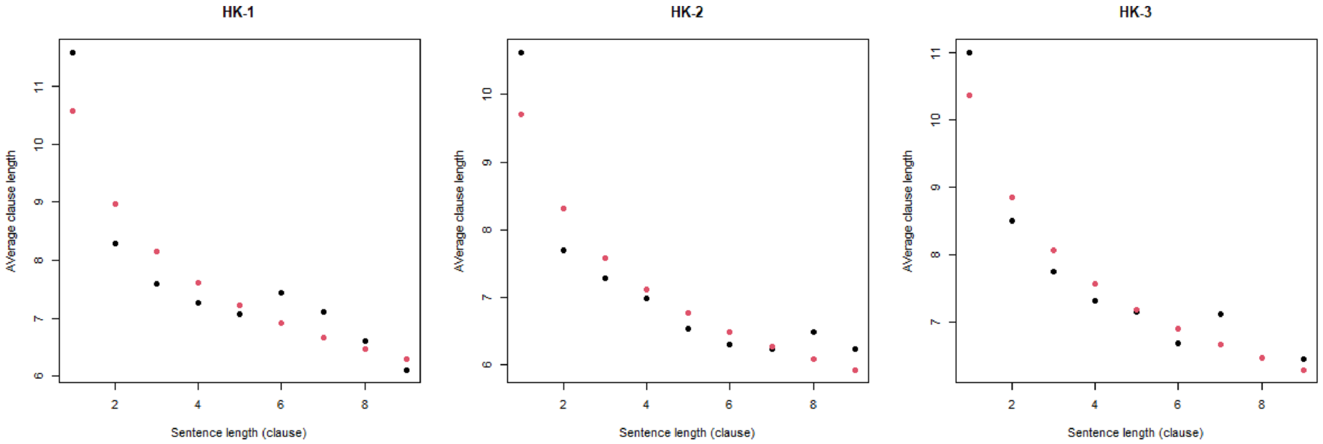

Fig. 3 The fitted results of the average clause length distribution in sentences in HK political speeches. HK-1, HK-2, and HK-3 represent 1984–1988, 1997–2001, and 2010–2014 Hong Kong political speeches, respectively.

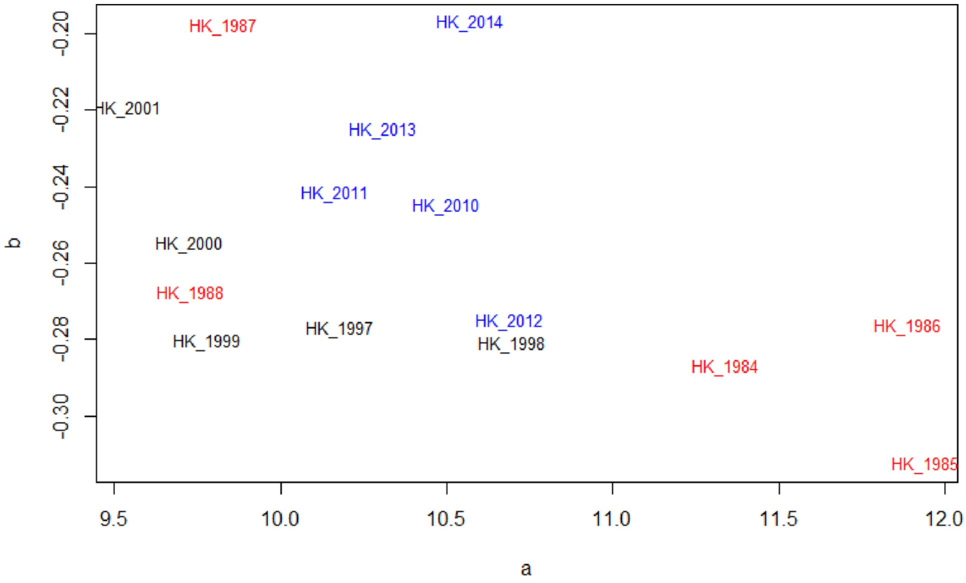

Fig. 4 The relation between different periods of HK political speeches represented by the fitted parameters of the relationship between sentence and clause lengths. HK political speeches from different periods were marked with different colors.

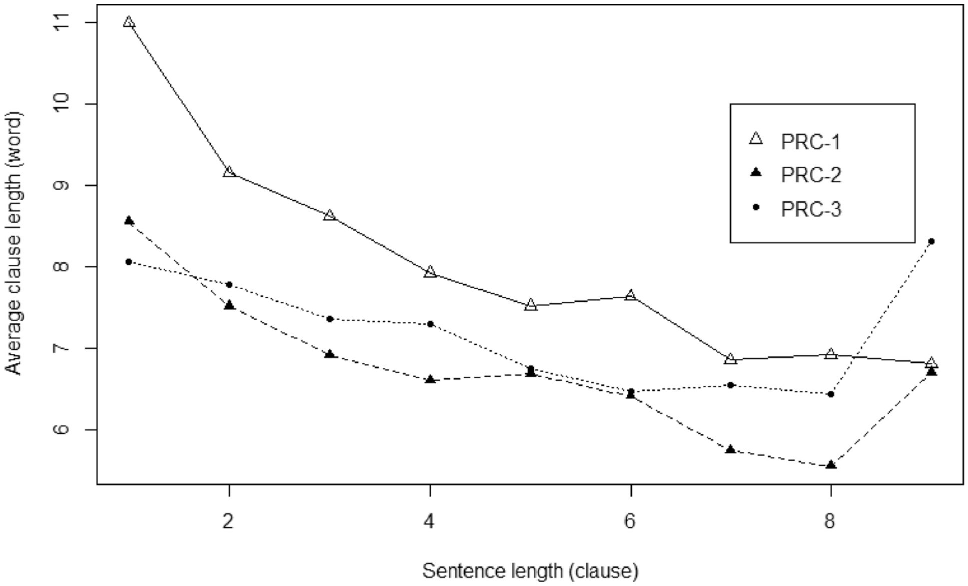

Fig. 5 The average clause length distributions in sentences in PRC political speeches. PRC-1, PRC-2, and PRC-3 represent 1984–1988, 1997–2001, and 2009–2013 PRC political speeches, respectively.

two group values of *a* is significant. The group means values of *a* are 11.092 and 8.559, respectively, in the first period and the latter two periods, and the *p*-value is much <0.05. This result shows that the difference between them is significant. Combined with the result of the t-test and Fig. 7, we can see that there is a large and significant difference between *a* values from the first period and the other latter two periods. This means that the average clause length in the sentences with the same length in the first period of PRC political speeches is longer than that in other two periods.

The large values of *b* in the third period showed that the average clause length has a small decreasing tendency with sentence length. The *b* values increased with time. The stored structure information with sentence length increases is small in the third period. The link between each clause in the sentences is weakened until the relationship between the sentence and clause does not abide by the MA law. The result of the *t*-test shows that there is a significant difference between the two group values of *b* in the third period and the first two periods of PRC political speeches. In addition, there are relatively obvious boundaries between different periods of PRC political speeches from Fig. 7. Therefore, it can be said that the fitted values, *a* and *b*, can differentiate the three different periods of PRC political speeches. In the meantime, the parameters have a diachronic change with time from the first period to the third period, respectively.

The PRC and HK political speeches represented by the two fitted parameters are displayed in Fig. 8. From this, it can be seen whether these two parameters can differentiate the Chinese varieties of PRC and HK political speeches.

Figure 8 shows that the first period PRC political speeches were closer to the HK political speeches, compared to the other two periods' PRC political speeches. The *t*-test showed that the average values of *a* are 10.460 and 8.423 in the HK political speeches and in the latter two periods of PRC political speeches, respectively; the p-value is much less than 0.05. This means the

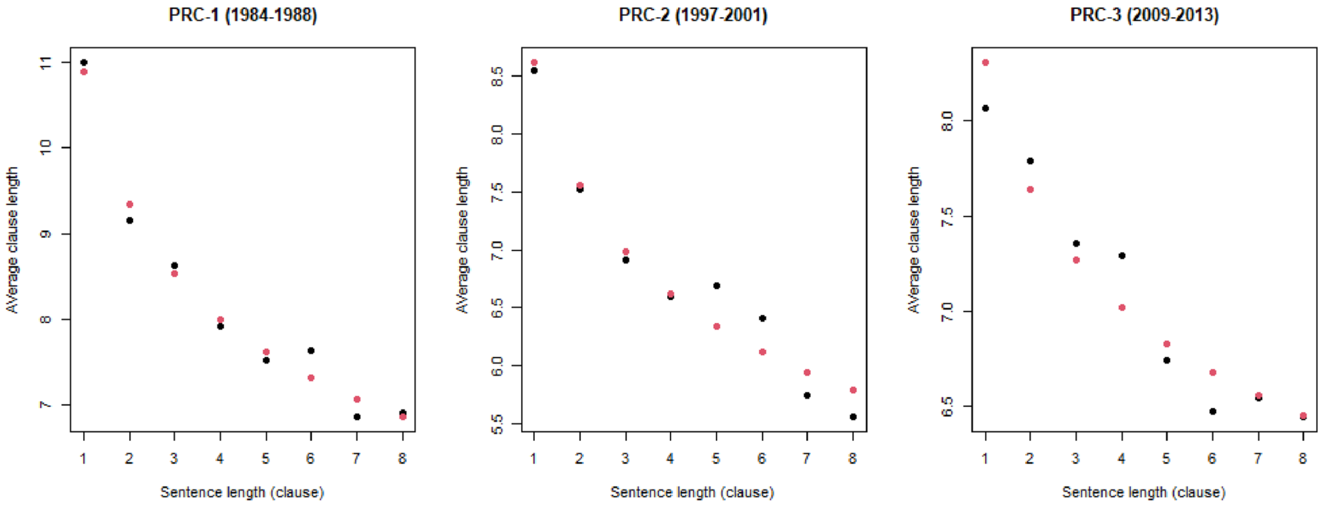

Fig. 6 The fitted results of the relationships between sentence and clause lengths in PRC political speeches. The black and red dots represent the observed and fitted values of the average clause length in PRC political speeches, respectively.

Table 2 The fitted results of the average clause length distributions in sentences in PRC political speeches.

	a	*b*	*R*^{2}
1984–1988	10.896	−0.223	97.83%
1997–2001	8.623	−0.191	93.82%
2009–2013	8.307	−0.122	92.74%

difference in *a* values between them is significant. Another t-test also showed that there is a significant difference between *b* values in the latter two periods of PRC political speeches and HK political speeches, and the *b* values are large in the latter two periods of PRC political speeches. Combined with Fig. 8, it can be said that the parameters *a* and *b* can differentiate the latter two periods' PRC political speeches and HK political speeches.

In PRC political speeches, parameter *a* decreases and parameter *b* increases with time. Particularly, in the third period of PRC political speeches, the large *b* means that the decreasing trend of average clause length is small and the increased structure information of a sentence stored in brain memory is small when sentence length increases. With this, the strength of the link between clause and clause in a sentence decreases. The extent of diachronic change of the fitted parameters, *a* and *b*, are largely from the first period to the latter two periods in PRC political speeches. The language change in PRC political speeches during these 30 years is large.

The distance between the three periods of HK political speeches is small, which means the language change from the first to the third period in HK political speeches is small. The smaller values of the fitted parameter *b* means that the decreasing tendency of average clause length is large and the structural information stored in the brain memory is large. The extent of the link between clause and clause in a sentence in HK political speeches is large.

The distance between different regional political speeches in the same period is becoming larger and larger from time to time. From Fig. 8, we also can deduce that the regional differences between PRC and HK political speeches are due to the diachronic change speed in PRC political speeches, which is larger than in HK political speeches. The same reason explains why the fitted parameters can differentiate PRC political speeches from different periods.

Thus, to answer RQ1, we conclude that the relationship between sentence and clause in PRC pollical speeches abides by the MA law and can be fitted by Formula (2), except for the three PRC political speeches in the third period. The fitted parameters can differentiate the different periods of PRC political speeches and have a diachronic change.

The average clause length in HK political speeches decreases with sentence length when headlines are considered as standing-alone clauses, but not 1-clause sentences. The average clause length distributions can be fitted by Formula (2) in this circumstance.

There are similar fitted parameter values in the three periods of HK political speeches compared to the three periods of PRC political speeches. The distance between HK political speeches and the second and third periods of PRC political speeches is large. We argue that the fitted parameters can differentiate these two regional political speeches, even though they started out fairly close to each other. The different speeds of diachronic change in Chinese political speeches between Mainland China and HK SAR led to an increasing divergence between the two varieties.

The relationship between clause and word

The average word lengths in clauses were 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 distributions in HK and PRC political speeches from different periods, as shown in Figs. 9 and 10, respectively. The headlines in HK political speeches were considered to be clauses when calculating the average word length distributions in clauses.

From Fig. 9, the average word length decreases with the clause length in HK political speeches from each period. The decreasing trend of the average word length in short clauses is large; contrastively, it is small in long clauses. Except for the 1-word clauses, the average word length in clauses with the same length from 1984 to 1988 HK political speeches is small, whereas in 2010–2014 HK political speeches it is large, and in 1997–2001 it is somewhere in between. This shows that the average word length increases diachronically in these three periods in the clauses with the same length.

Figure 10 shows that the average word length also decreases with the clause length in PRC political speeches from each period. There is an outlier value of average word length in the 3-word clauses in PRC political speech from 2009 to 2013. The average word length in 3-word clauses from 2009 to 2013 is relatively

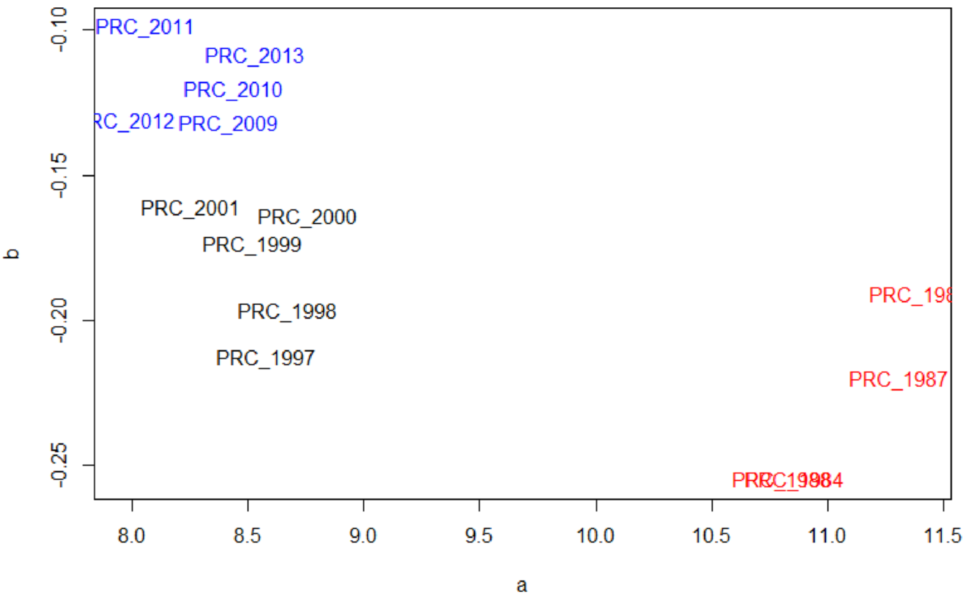

Fig. 7 The relationship between the different periods of PRC political speeches represented by the fitted parameters. PRC political speeches from different periods were marked with different colors.

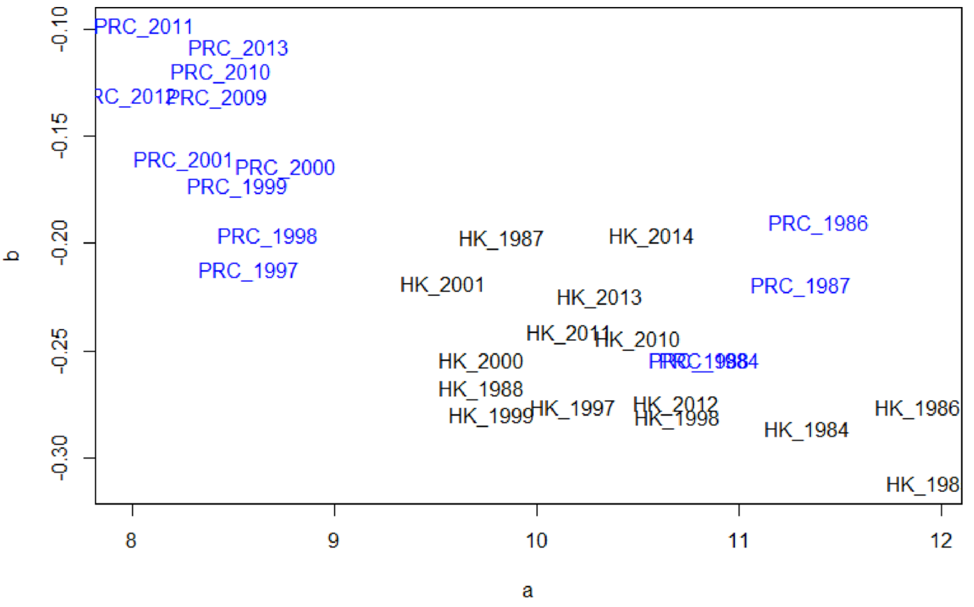

Fig. 8 The relationship between PRC and HK political speeches is represented by the fitted parameters of the relationship between sentence length and clause length. The political speeches from PRC and HK are marked with different colors.

small. We examined the clauses with 3 words in this period. The results show that there are more one-character words in these clauses, for example: “各位代表 (Ladies and gentlemen)” representing form of address, “一年来 (over the past year), 五年来 (over the past five years), 这五年 (in the past five years)” representing time. This leads to the small average word length in these clauses.

Similarly, the average word length in clauses with the same length increases diachronically with time. Compared to Fig. 9, the average word length in short clauses from PRC political speeches is longer than the average word length from HK political speeches. Figures 9 and 10 validated that the word lengths increase diachronically in both regional political discourses. The average word length in the clauses with the same length from PRC political speeches is larger than from HK political speeches.

Formula (2) was used to fit the average word length distributions in different periods of PRC and HK political speeches, as shown in Table 3.

From Table 3, it can be seen that the average word length distributions can be fitted by Formula (2) and the fitted results are good through the determination coefficients (R^2). The relationships between clause and word abide by the MA law in PRC and HK political speeches. The *b* values of the fitted result in PRC political speeches are smaller than in HK political speeches. This means that the decreasing tendencies of average word length in different periods of PRC political speeches are larger than in HK political speeches, and the structural information of the clause that needs to be stored for processing is large.

The average word length distributions in clauses in all 30 political speeches from PRC and HK were fitted using Formula (2).

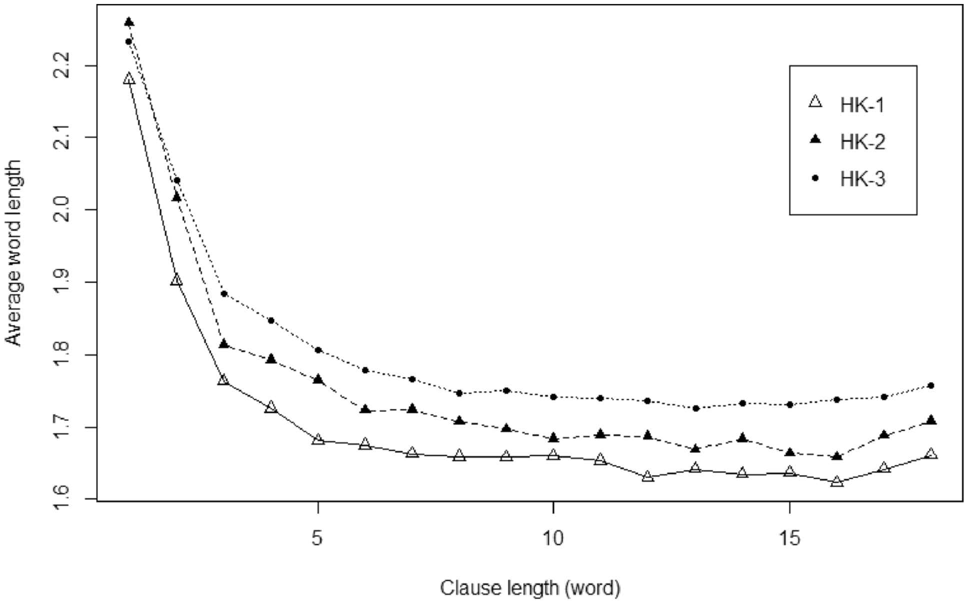

Fig. 9 The average word length distributions in clauses in HK political speeches. HK-1, HK-2, and HK-3 represent 1984–1988, 1997–2001, and 2010–2014 HK political speeches, respectively.

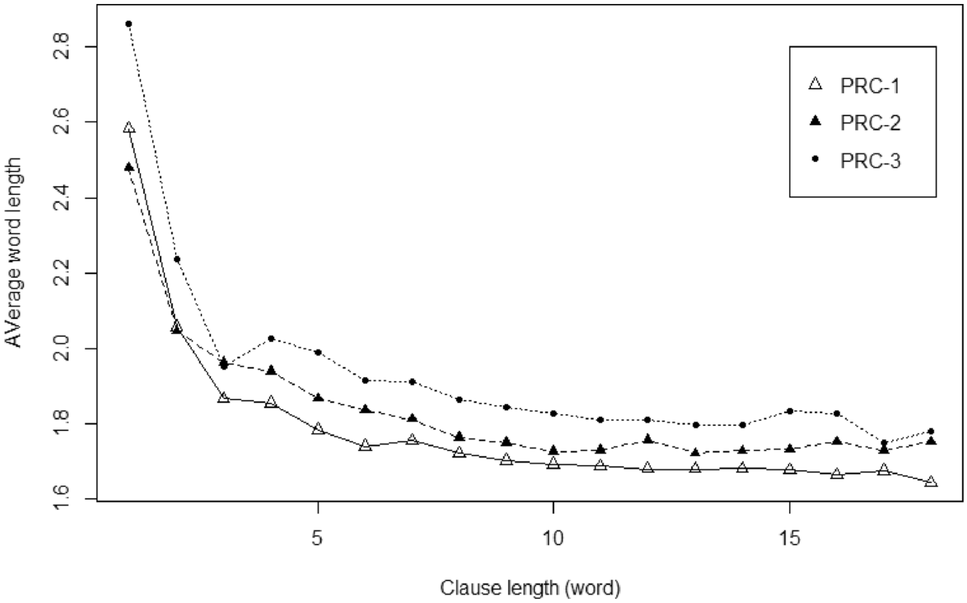

Fig. 10 The average word length distributions in clauses in PRC political speeches. PRC-1, PRC-2, and PRC-3 represent 1984–1988, 1997–2001, and 2009–2013 PRC political speeches, respectively.

The fitted results are shown in Table 4 in Appendix. The determination coefficients (R^2) of fitted results of the relationship between the lengths of clause and word in 2013 PRC political speeches and 1986 HK political speeches are 63.75% and 64.62%, respectively, thus the fitted results are not very good in these two political speeches. The values of R^2 of fitted results of the average word length distributions in other PRC and HK political speeches showed that the fitted results are good. The relationship between clause and word in each PRC and HK political speech abides by the MA law.

The PRC and HK political speeches were represented by the fitted parameters, a and b, of the relationship between clause and word lengths. They are visualized in a 2-Dimensional space, as shown in Figs. 11 and 12, respectively.

Figure 11 shows that the distance between the first-period and second-period PRC political speeches is small and the change of fitted parameters of average word length distribution in clauses is small. The distance between third-period PRC political speeches and the other two periods is relatively large. This means that the change of fitted parameters is relatively large. There is an obvious boundary between second and third-period PRC political speeches. The distances between 1997 and 2001 PRC political speeches are small compared to the other two periods of PRC political speeches.

Figure 12 shows that there is a rough boundary between different periods of HK political speeches, especially between the third and the other two periods. This means that the diachronic change in HK political speeches is large from the second to the third period. The fitted parameters, *a* and *b*, are more concentrated in 1997–2001 HK political speeches as with the same period of PRC political speeches. From Figs. 11 and 12, there are similar change tendencies of the two fitted parameters in PRC and HK political speeches. On the whole, the two parameters, *a* and *b*, have an increasing tendency with time in each regional political speech from PRC and HK, respectively.

Both regions' political speeches, represented by the two fitted parameters of average word length distributions, are visualized in one figure, as shown in Fig. 13.

Figure 13 shows that there is a rough boundary between PRC and HK political speeches even though some HK political

Table 3 The fitted results of the average word length distributions in clauses from different periods of PRC and HK political speeches.

		a	*b*	R^2
HK	1984-1988	2.022	−0.088	83.52%
	1997-2001	2.122	−0.098	87.96%
	2010-2014	2.128	−0.085	88.81%
PRC	1984-1988	2.303	−0.132	85%
	1997-2001	2.291	−0.114	89.38%
	2009-2013	2.521	−0.136	82.18%

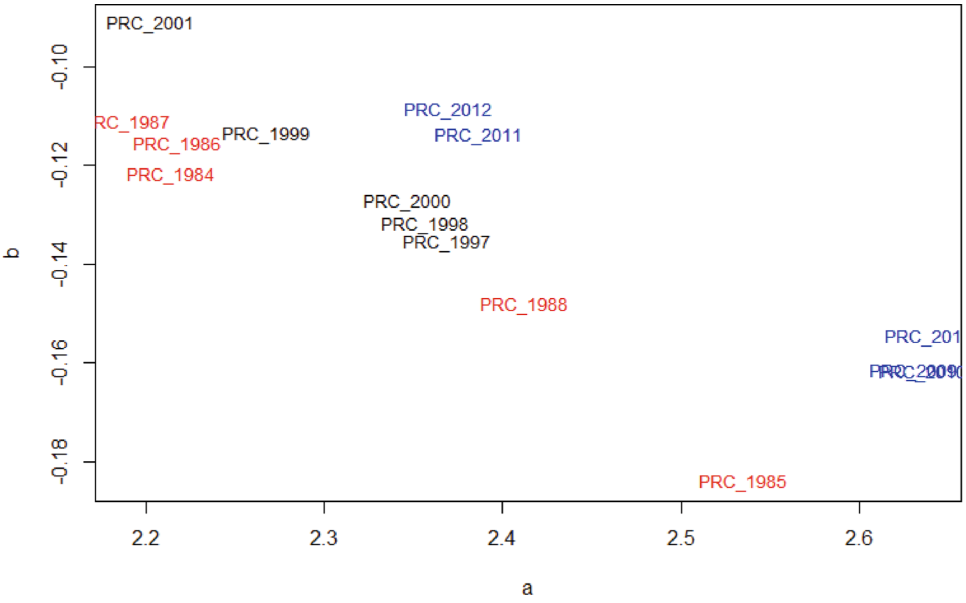

Fig. 11 The relationship between different periods of PRC political speeches represented by the fitted parameters of the relationship between clause length and word length. PRC political speeches from different periods were marked with different colors.

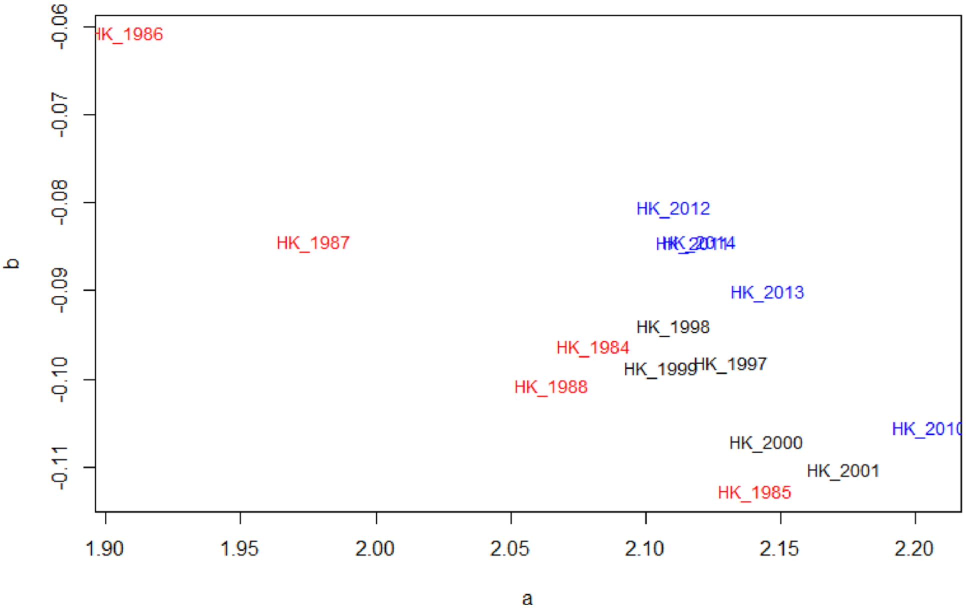

Fig. 12 The relationship between the different periods of HK political speeches represented by the fitted parameters of the relationship between clause length and word length. HK political speeches from different periods were marked with different colors.

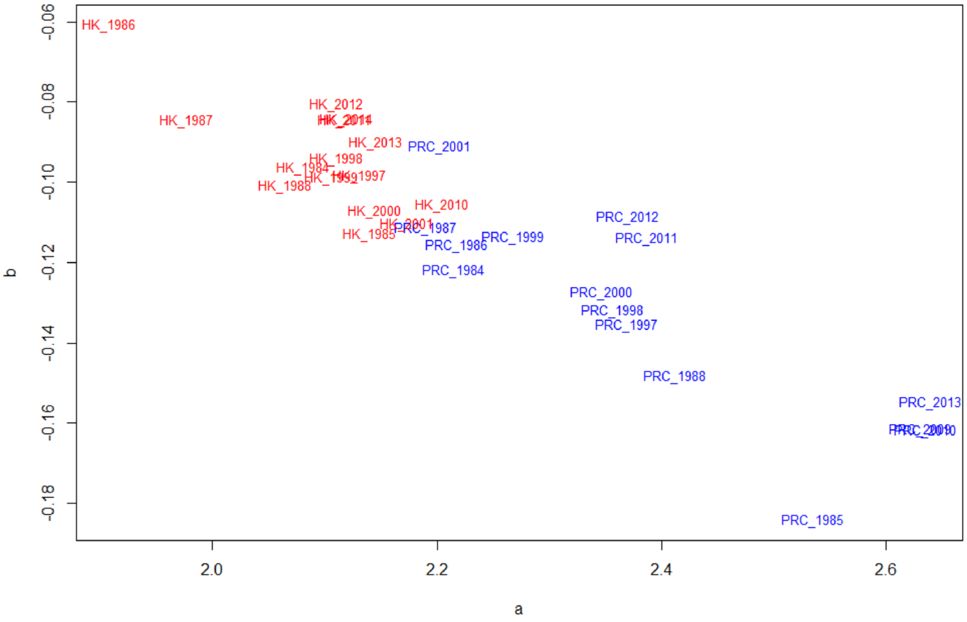

Fig. 13 The positions of PRC and HK political speeches represented by fitted parameters, *a* and *b*, of the relationship between clause length and word length. The political speeches from PRC and HK are marked with different colors, blue and red, respectively.

speeches are closer to PRC political speeches. The results of the *t*-test showed that there are significant differences in the group means of fitted parameters, *a* and *b*, between PRC and HK political speeches. The agglomerative hierarchical clustering analysis is used to cluster the political speeches from PRC and HK, in which Euclidean distance was used to calculate the similarity between different political speeches. The result of the clustering analysis is shown in Fig. 14 and Table 4.

From Fig. 14 and Table 4, the two HK political speeches and the five PRC political speeches are clustered into one same cluster, and the distance between them is small. Except for these two HK political speeches, there is an obvious boundary between HK and PRC political speeches. Basically, it can be considered that the fitted parameters can distinguish the political speeches between the two regions.

The values of *a* in PRC political speeches are larger than those in HK political speeches. This means that the average word length in the same length clauses in PRC political speeches is larger than in HK political speeches. It is possible that the disyllabization leads to the large average word length in clauses. The increase in word length may be used to avoid the communicational ambiguity that is led by the simplification of syllables based on the principle of least effort. The linguistic diachronic change leads to a change in speakers' behaviors.

In addition, most values of *b* in PRC political speeches are smaller than in HK political speeches, which means the extent of the decreasing of the average word length in clauses is large and thus, the clause information that must be retained for processing is also large in PRC political speeches. The words in clauses in PRC political speeches are closely related. In a word, we confirmed that the fitted parameters, *a* and *b*, can differentiate the regional variations of political speeches in the PRC and HK.

Moreover, there are large distances between PRC and HK political speeches from the same period, as can be seen in Figs. 15–17. The t-test also validates that there is a significant difference in *a* values between PRC and HK political speeches from the same period. There are small distances between political speeches in different periods from the same region, especially in HK. The *b* values decrease with *a* values from Figs. 13 and 15–17, and they are correlated negatively; however, whether their correlation in the two regional political speeches from the same period can be fitted using linear regression needs further exploration. Additionally, the diachronic change model of fitted parameters between clause and word lengths in political speeches may be further examined. From Fig. 13, we can see that the distance between different periods of PRC political speeches is larger than that between different periods of HK political speeches. This also means that the speed of change of PRC political speeches is a little larger than that of HK political speeches.

Table 4 The clustering result of the two regions' Chinese political speeches represented by the fitted parameters of the relationship between clause length and word length.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
HK	2	0	0	13
PRC	5	4	6	0

Thus, to answer RQ2, we conclude that the relationship between the lengths of clause and word abides by the MA law in PRC and HK political speeches and can be fitted by Formula (2). The fitted parameters, *a* and *b*, can differentiate PRC and HK political speeches. The different period's political speeches in the PRC and HK can also be differentiated roughly by these two parameters, as seen in Figs. 11 and 12, respectively. This means that there are diachronic changes in the fitted parameters of the two regions' political speeches. There are similar diachronic change tendencies of the two fitted parameters in PRC and HK political speeches, which is unlikely to be coincidental. This also means there are similar diachronic changes in Chinese regional varieties from the PRC and HK. Whether this shared tendency of changes by variants of the same language would apply more generally to other languages should be further explored.

Conclusion

This paper explored the interaction of diachronic changes and regional variation of political speeches in Chinese based on the MA law. The comparable corpus used in this study contains both the Premier's yearly reports in the People's Republic of China and

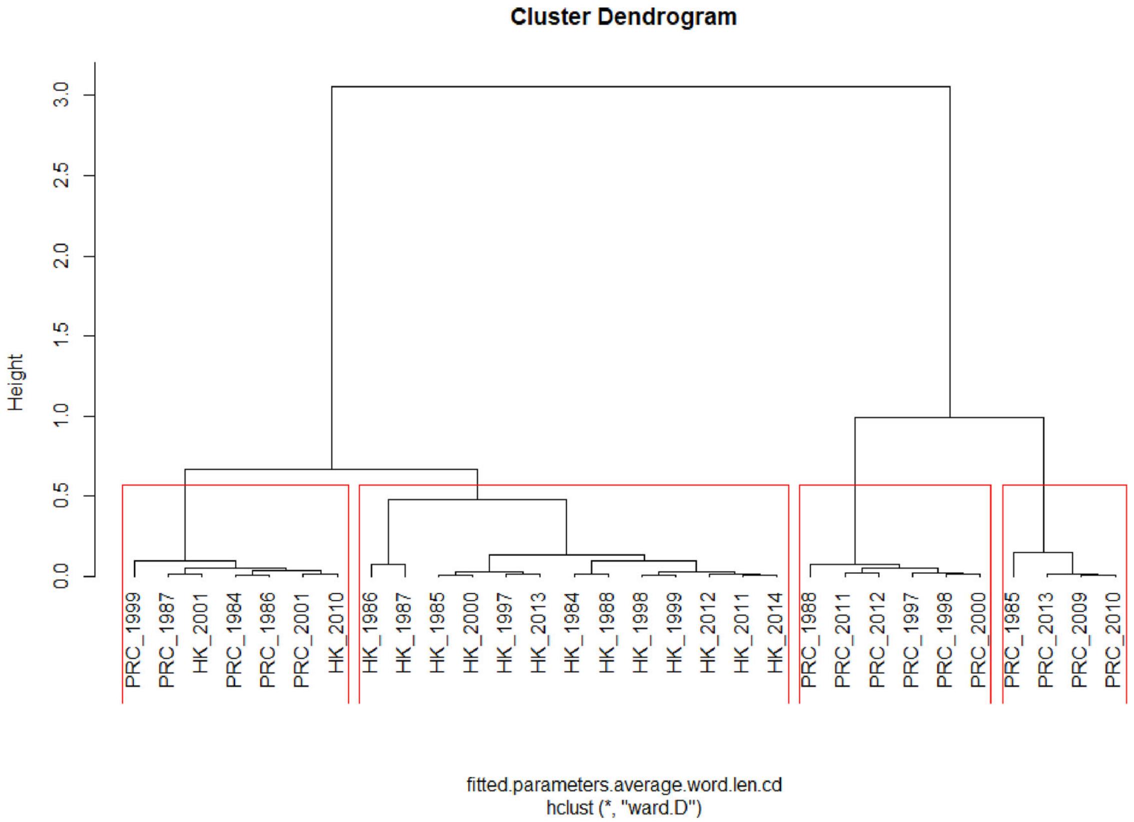

Fig. 14 The text clustering result of PRC and HK political speeches represented by the fitted parameters of the relationship between clause length and word length. All four clusters are marked with red color.

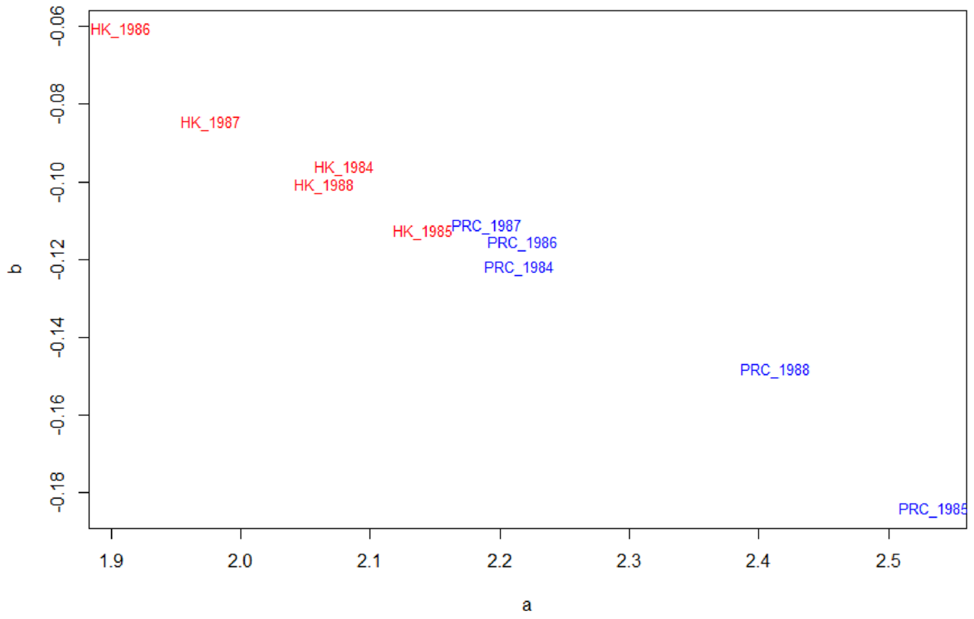

Fig. 15 The positions of the first period PRC and HK political speeches represented by the fitted parameters of relationship between the clause and word lengths, *a* and *b*. The PRC and HK political speeches from 1984-1988 are marked with blue and red, respectively.

the Chief Executive's yearly policy addresses in Hong Kong, SAR. Three different periods were selected for longitudinal studies of diachronic changes in both regions and the regional differences in Chinese political speeches were also explored synchronically. As such, this is the first study to model both diachronic changes and synchronic variations based on the same dataset, and the first attempt to show that the MA law provides a uniform model to capture the complex adaptation to the causes of historical changes and regional variations.

The first research question is if the MA law can be fitted to account for both regional variations and historical changes. We found that it can model regional variations, but was only

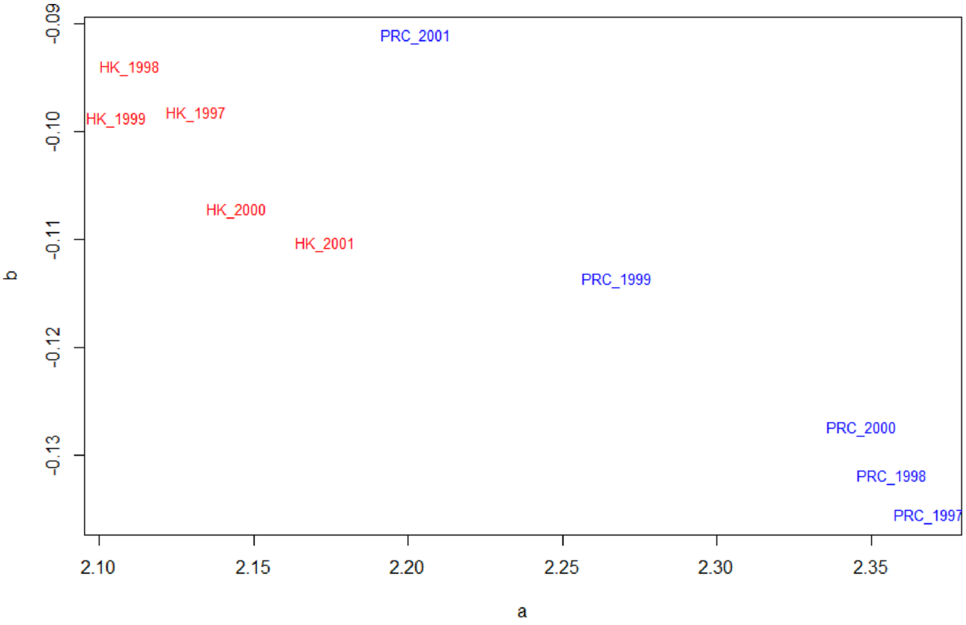

Fig. 16 The positions of the second period PRC and HK political speeches represented by the fitted parameters of relationship between the clause and the word lengths, *a* and *b*. The PRC and HK political speeches from 1997-2001 are marked with different colors, blue and red, respectively.

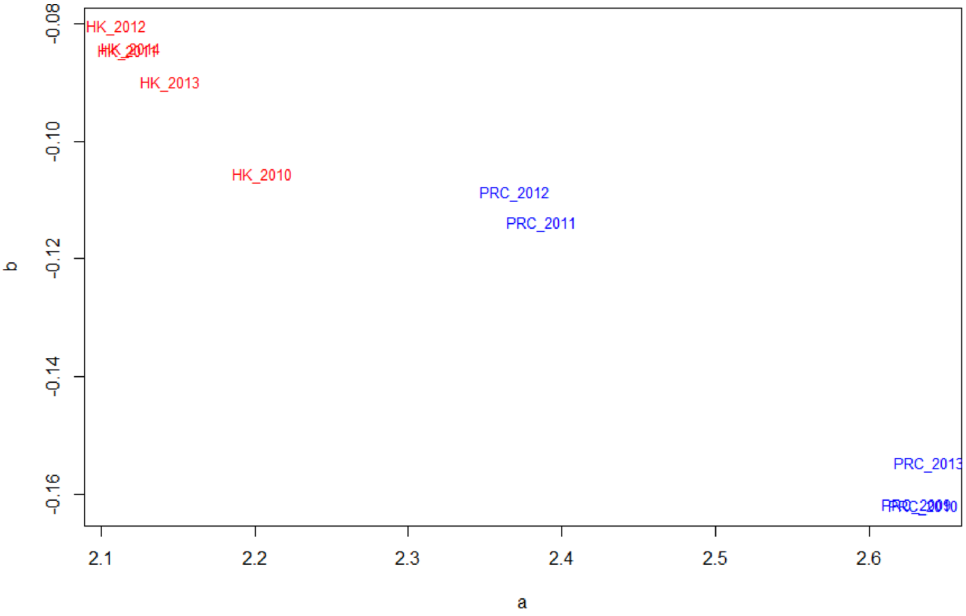

Fig. 17 The positions of the third period PRC and HK political speeches represented by the fitted parameters of relationship between the clause and word lengths, *a* and *b*. The PRC political speeches from 2009-2013 and HK political speeches from 2010-2014 are marked with different colors, blue and red, respectively.

successful in modeling diachronic change in the PRC corpus when the relationship between sentence length and clause length is considered. In the process, several interesting discoveries were made. First, we fit the relationship between sentence and clause successfully with the MA law after headlines were correctly interpreted as standing-alone clauses in HK political speeches. Second, there are no obvious differences between different periods of HK political speeches when they are represented by the fitted parameters, while the parameters of the fitted model separate PRC political speeches from different time periods. Third, we showed that the two variants are the closest in the earliest period. The distances between the latter two periods PRC political speeches and HK political speeches are significantly larger. The above two results suggest that the diachronic changes in PRC political speeches contributed most to the increasing differences between the two variants over time. Fourth, the fitted results of the relationship between sentence and clause in PRC political speeches in 2009–2013 are the least fitted among all results. The relatively high values of *b* in this period suggest that the link between clauses in sentences is growing weaker. We suspect that this is the consequence of the tendency to adopt a colloquial style of speaking during this period of time.

Instead of speculating that some language changes cannot be captured by the parameters of the MA law, a possible interpretation is that the political speeches by the Chief Executive in HK did not undergo significant changes. The fact that the latter part of our data covers the 18 years following the handover of Hong Kong in 1997 supports this interpretation. Note that the mantra of the HK Chief Executive during this period is 'unchanged,' as promised by China for the first 50 years of the Hong Kong SAR. It is very likely that the Chief Executive maintained the same language to signal the unchanged status of Hong Kong and to reassure the public.

In addition, following Hou et al. (2019b) and Hou et al. (2020b), the relationship between clause and word is also adopted and shown to successfully model both regional varieties of Chinese political speeches, offering further support to the positive answer to RQ1. The results show that the average word length distributions can be fitted by Formula (2) according to the MA law. The parameters, *a* and *b*, of fitted results can differentiate PRC and HK political speeches. There are also large differences in fitted parameters in PRC and HK political speeches from the same period. Significant discoveries include the following: First, unlike the fitted result based on sentence-clause correlation, the *b* values in PRC political speeches are smaller than in HK political speeches. This suggests that the PRC political speeches have a stronger tendency to decrease average word length than in HK. Second, the *a* values of fitted results in the PRC are larger than that in HK. This means that in clauses with the same length, the average word length in PRC political speeches is larger than in HK.

Lastly, the results from this study are based on sentence–clause correlation and the clause–word correlation can be combined to answer RQ2, which examined whether fitted parameters of the MA law could differentiate regional variations and diachronic changes in PRC and HK political speeches. In fact, the interesting finding reported in the previous paragraph of longer words having a faster rate of decrease compared to words of average length can be accounted for when results from both studies are put together. These two clause–word model tendencies of PRC and Hong Kong political speeches, in spite of the lack of diachronic change in the sentence–clause model of Hong Kong, are consistent with the account that the two variants share the tendency to move to shorter words. However, since the average word length in HK is already significantly shorter, the tendency is not strong enough to cause a percolation effect of changes at the sentence–clause level for HK. At the same time, the PRC can afford a sustained stronger tendency for decreasing word lengths since, at least during the period of our study, the average word length in PRC is still longer than in HK. In addition, we also found that there are similar diachronic changing tendencies of fitted parameters in the PRC and HK. It seems that there is a negative linear correlation between two fitted parameters in each period of PRC and HK political speeches.

There are two theoretical and methodological implications of the current study. First, the self-organizing complex system model, as represented by the MA law, allows us to study diachronic change and regional variations simultaneously. That is, we can now conduct empirical studies to explore the interactions of changes and variations. Language changes and variations are the two pillars of the evolution of language that have been assumed to interact but rarely if ever, studied in the same context. Our study not only proposes an empirical methodology but also demonstrates how changes and variations have compounding effects. For instance, our studies showed that the divergence of HK and PRC variations can be attributed mostly to the rapid diachronic changes in the PRC. Second, our direct comparison of modeling of sentence-clause correlation and the clause–word correlations suggests that the lower level correlation, e.g., clause–word level, seems to be more sensitive and can provide additional information. This is consistent with the prediction of lexicalist theories. This would also have strong implications for quantitative linguistics, as the majority so far have been dominated by the sentence–clause correlation.

Our current study demonstrates that the correlation between a linguistic unit and its immediate constituent, as modeled by MA law, serves as a good parameter to differentiate variations and changes in Chinese political speeches in the PRC and HK. In addition, we observed that parameters *a* and *b* characterize a complex and adaptive system according to MA law. The correlation between these two fitted parameters between clause and word in the PRC and HK corpora similarly changed over time, indicating that the application of MA law to study the aggregation and interactions of language units as a part of a complex system self-organizing in the face of different types of changes is a promising research topic for future studies.

Data availability

All data generated during this study are included in this published article as shown in the Appendix.
The dataset (corpus) analyzed during the current study is available in HKBU political corpus (https://digital.lib.hkbu.edu.hk/corpus/index.php).

Received: 27 May 2022; Accepted: 13 December 2022;
Published online: 24 December 2022

Note

[1] Different methods to measure word length in Chinese were presented in the literature. Deng and Feng (2013) measured word length by syllables, following Zipf's (1949) work, to explore the relationship between word length and the relative occurrence frequency. Chen and Liu (2016) argued that syllable is the most appropriate measurement unit of word length in spoken Chinese and the character component is the most appropriate measurement unit of word length in written Chinese. Hou et al. (2019b, 2020a) validated the application of the MA Law between clause and word lengths, with word length measured by the number of syllables. Chen and Liu (2022) investigate the MA law at the constituent levels of clause > word > character and concluded that sub-character units such as character components and strokes are parts of the writing system and are not proper constituents of a linguistic unit. Note that the number of syllables in a word equals the number of characters in the written formal Chinese register. Thus, it can also be said that Chinese word length is defined as the number of Chinese characters in a word; this is the current consensus. This choice is supported by Köhler's (2012) study that showed that word length is measured in terms of syllables in most quantitative studies.

References

Altmann G (1980) Prolegomena to Menzerath´s law. Glottometrika 2:1–10
Ahrens K (2015) Corpus of political speeches. Hong Kong Baptist University Library. http://digital.lib.hkbu.edu.hk/corpus/
Ahrens K, Zeng WH (2022) Referential and evaluative strategies of conceptual metaphor use in government discourse. J Pragmat 188:83–96. https://doi.org/10.1016/j.pragma.2021.11.001
Arens H (1965) Verborgene Ordnung. Pädagogischer Verlag Schwann
Baayen RH (2008) Analyzing linguistic data: a practical introduction to statistics using R. Cambridge University Press, Cambridge
Beckner C, Blythe R, Bybee J, Christiansen MH, Croft W, Ellis NC, Holland J, Ke J, Larsen-Freeman D, Schoenemann T (2009) Language is a complex adaptive system: position paper. Language Learn 59:1–26
Benešová M (2016) Text segmentation for Menzerath–Altmann law testing. Palacký University, Faculty of Arts
Berdicevskis A (2021) Successes and failures of Menzerath's law at the syntactic level. In: Proceedings of the second workshop on quantitative syntax (Quasy, SyntaxFest 2021), Sofia, Bulgaria. Association for Computational Linguistics, pp 17–32
Buk S, Rovenchak A (2008) Menzerath–Altmann law for syntactic structures in Ukrainian. Glottotheory 1:10–17
Burgers C, Ahrens K (2020) Change in metaphorical framing: metaphors of TRADE in 225 years of state of the union addresses (1790–2014) Appl Linguist 41(2):260–279
Chao YR (1968) A grammar of spoken Chinese. University of California Press, Berkeley and Los Angeles

Chen CY, Tseng SF, Huang CR, Chen KJ (1993) Some distributional properties of Mandarin Chinese—a study based on the Academia Sinica corpus. In: Huang CR, Chang CH, Chen KJ, Liu CH (eds) Proceedings of Pacific Asia conference on formal and computational linguistics I. Academia Sinica, Taipei, pp 81–95

Chen H, Liu H (2016) How to measure word length in spoken and written Chinese. J Quant Linguist 23(1):5–29. https://doi.org/10.1080/09296174.2015.1071147

Chen H, Liu H (2022) Approaching language levels and registers in written Chinese with the Menzerath–Altmann Law. Digital Scholarship in the Humanities. https://doi.org/10.1093/llc/fqab110

Conway D, White J M (2012) Machine learning for Hackers. O'Reilly Media, Inc

Cramer I (2005) The parameters of the Altmann–Menzerath law. J Quant Linguist 12:41–52

Deng Y, Feng Z (2013) A quantitative linguistic study on the relationship between word length and word frequency. J Foreign Language 36(3):29–39

Dong S, Yang Y, Ren H, Huang C-R (2021) Directionality of atmospheric water in Chinese: a lexical semantic study based on linguistic ontology. SAGE Open. https://doi.org/10.1177/2158244020988293.

Fenk-Oczlon G, Pilz J (2021) Linguistic complexity: relationships between phoneme inventory size, syllable complexity, word and clause length, and population size. Front Commun 6:626032

Grzybek P, Stadlober E (2007) Do we have problems with Arens' law? A new look at the sentence-word relation. In: Grzybek P, Stadlober E (eds) Exact methods in the study of language and text: dedicated to Gabriel Altmann on the occasion of His 75th Birthday. De Gruyter, Berlin, pp 205–217

Holland JH (1996) Hidden order: how adaptation builds complexity. Addison Wesley Longman Publishing Co., Inc.

Hou R, Huang C-R, Ahrens K, Lee Y-MS (2020a) Linguistics characteristics of Chinese register based on the Menzerath–Altmann law and text clustering. Digit Scholarsh Humanit 35(1):54–66

Hou R, Huang C-R, Ahrens K (2020b) Language change in report on the work of the government by Premiers of the People's Republic of China. In: Nguyen ML, Luong MC, Song S (eds) Proceedings of the 34th Pacific Asia conference on language, information and computation, 24–26 October. University of Science, Vietnam National University, Hanoi, Vietnam

Hou R, Huang C-R, Do HS, Liu H (2017) A study on correlation between Chinese sentence and constituting clauses based on the Menzerath–Altmann Law. J Quant Linguist 24(4):350–366

Hou R, Huang C-R, Liu H (2019a) A study on Chinese register characteristics based on regression analysis and text clustering. Corpus Linguist Linguist Theory 15(1):1–37. https://doi.org/10.1515/cllt-2016-0062

Hou R, Huang C-R, Zhou M, Jiang M (2019b) Distance between Chinese registers based on the Menzerath–Altmann Law and regression analysis. Glottometrics 45:24–56

Hou R, Huang C-R, Ahrens K (2021) Language change in Chinese political discourses based on the relationship between sentence and clause. In: Hu K, Kim J-B, Zong C, Chersoni E (eds) Proceedings of the 35th Pacific Asia conference on language, information and computation, 5–7 November. Shangai International Studies University, Shanghai. China

Huang C-R, Dong S, Yang Y et al. (2021) From language to meteorology: kinesis in weather events and weather verbs across Sinitic languages. Humanit Soc Sci Commun 8:4. https://doi.org/10.1057/s41599-020-00682-w

Huang C-R, Shi D (2016) A reference grammar of Chinese. Cambridge University Press, Cambridge

Huang C-R, Xue N (2015) Modeling word concepts without convention: linguistic and computational issues in Chinese word identification. In: Wang William S-Y, Sun Chao-Fen (eds) The Oxford handbook of Chinese linguistics. Oxford University Press, New York, pp 348–361

Huang C-R, Xue N (2012) Words without boundaries: computational approaches to Chinese word segmentation. Language Linguist Compass 6(8):494–505

Jiang X, Jiang Y, Hoi CKW (2020) Is Queen's English drifting towards common people's English—quantifying diachronic changes of Queens's Christmas messages (1952–2018) with reference to BNC. J Quant Linguist. https://doi.org/10.1080/09296174.2020.1737483

Jiang Y, Ma R (2020) Does Menzerath–Altmann Law hold true for translated language: evidence from translated English Literary texts. J Quant Linguist. https://doi.org/10.1080/09296174.2020.1766335

Köhler R (1982) Das Menzerathsche Gesetz auf Satzebene. In: Lehfeldt W, Straus U (eds) Glottometrika 4. Brockmeyer, Bochum, pp 103–113

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

Köhler R (1989) Das Menzerathschen Gesetz als Resultat des Sprachverarbeitungsmechanismus. In: Altmann, Schwibbe (eds) Das Menzerathsche Gesetz in informations – verarbeitenden Systemen. Olms, Hildesheim. pp 108–112

Köhler R (1993) Synergetic linguistics. In: Köhler R, Rieger B B (eds) Contributions to quantitative linguistics. Springer, Dordrecht, pp 41–51

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

Kubát M, Cech R (2016) Quantitative analysis of US Presidential inaugural addresses. Glottometrics 34:14–27

Leech G (2012) Modality on the move: the English modal auxiliaries 1961–1992. In: Facchinetti R, Palmer F, Krug M (eds) Modality in contemporary English. De Gruyter Mouton, Berlin, Boston, pp. 223–240

Lu L W-L, Ahrens K(2008) Ideological influences on BUILDING metaphors in Taiwanese presidential speeches. Discourse Soc 19(3):383–408

Lyons J (1968) Introduction to theoretical linguistics, vol 510. Cambridge University Press, Cambridge

Mačutek J, Čech R, Courtin M (2021) The Menzerath–Altmann law in syntactic structure revisited. In: Čech R & Chen X (eds) Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021), Sofia, Bulgaria. Association for Computational Linguistics, pp 65–73

Millar N (2009) Modal verbs in TIME: frequency changes 1923–2006. Int J Corpus Linguist 14(2):191–220. https://doi.org/10.1075/ijcl.14.2.03mil

Miller GA (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev 63(2):81–97

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

Popescu I-I, Mačutek J, Altmann G (2009) Aspects of word frequencies. RAM-Verlag, Lüdenscheid

Prün C (1994) Validity of Menzerath–Altmann's law: graphic representation of language, information processing systems and synergetic linguistics. J Quant Linguist 1(2):148–155

Randour F, Perrez J, Reuchamps M (2020) Twenty years of research on political discourse: A systematic review and directions for future research. Discourse Soc 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 Soc 25(6):741–765

Savoy J (2018) Analysis of the style and the rhetoric of the 2016 US presidential primaries. Digit Scholarsh Humanit 33(1):143–160

Shahzad K, Mittenthal JE, Caetano-Anollés G (2015) The organization of domains in proteins obeys Menzerath–Altmann's law of language. BMC Syst Biol 9(1):1–13

Su Q, Liu P, Wei W et al. (2021) Occupational gender segregation and gendered language in a language without gender: trends, variations, implications for social development in China. Humanit Soc Sci Commun 8:133. https://doi.org/10.1057/s41599-021-00799-6

Thomason S (1997) Language variation and change. In: Nunberg G, Wasow T (eds) The fields of linguistics. The Linguistic Society of America, Washington, DC, https://www.linguisticsociety.org/resource/languagevariation-and-change

Torre IG, DeRbowski Ł, Herna'ndez-Ferna'ndez A (2021) Can Menzerath's law be a criterion of complexity in communication? PLoS ONE 16(8):e0256133

Tuldava J(1995) Informational measures of causality. J Quant Linguist 2(1):11–14

Wang J (2017) Representing Chinese nationalism/patriotism through President Xi Jinping's "Chinese Dream" discourse. J Language Politics 16(6):830–848

Wang S, Liu R, Huang C-R (2022) Social changes through the lens of language: q big data study of Chinese modal verbs. PLoS ONE 17(1):e0260210

Wang WS-Y (2006) Language is a complex adaptive system. J Tsinghua Univ (Philos Soc Sci) 21(6):5–13

Wang Y, Liu H (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 Soc 29(3):299–323

Wimmer G, Altmann G (2005) Unified derivation of some linguistic laws. In: Köhler R, Altmann G, Piotrowski RG (eds) Quantitative linguistics. de Gruyter, Berlin, New York, pp 791–807

Winter S, Gärdenfors P (1995) Linguistic modality as expressions of social power. Nordic J Linguist 18(2):137–165. https://doi.org/10.1017/S0332586500000147

Wodak R, Boukala S (2015) European identities and the revival of nationalism in the European Union. J Language Politics 14(1):87–109

Xu H, Jiang M, Lin J, Huang C-R (2022) Light verb variations and varieties of Mandarin Chinese: comparable corpus driven approaches to grammatical variations. Corpus Linguist Linguist Theory 18(1):145–173. https://doi.org/10.1515/cllt-2019-0049

Xu L, He L (2020) Is the Menzerath–Altmann Law specific to certain language in certain registers? J Quant Linguist 27(3):187–203. https://doi.org/10.1080/09296174.2018.1532158.

Yu B (2013) Language and gender in congressional speech. Lit Linguist Comput 29(1):118–132

Zhu D (1982) Lectures on grammar. Commercial Press, Beijing, China

Zampieri M, Nakov P (eds) (2021) Similar languages, varieties, and dialects: a computational perspective. Cambridge University Press

Zipf GK (1935) The psycho-biology of language: an introduction to dynamic philology. Houghton, Mifflin, Oxford, England

Zipf GK (1949) Human behavior and the principle of least effort: qn introduction to human ecology. Addison-Wesley, Reading, MA

Acknowledgements

This study was partly funded by the Project of Strategic Importance at the Hong Kong Polytechnic University (No. ZE2J), Guangdong's Social Science Project in China (No. GD20CZY03), and National Social Science Fund in China (No. 16BYY110).

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-022-01488-8.

Correspondence and requests for materials should be addressed to Renkui Hou, Chu-Ren Huang or Kathleen Ahrens.