


# Multiscale structural complexity analysis of the Chinese classics *A Dream of Red Mansions* and *All Men Are Brothers*

Jing Feng(冯靖), Ping Wang(王萍), and Changgui Gu(顾长贵)
**Citation:**Chin. Phys. B, 2026, 35 (**1**): 010506. DOI: 10.1088/1674-1056/ae1fe9
http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

**What follows is a list of articles you may be interested in**

---

## Interval multiscale sample entropy: A novel tool for interval-valued time series complexity analysis

Ping Tang(唐萍), Bao-Gen Li(李宝根), and Yang Wang(王阳)
Chin. Phys. B, 2025, 34 (**12**): 120508. DOI: 10.1088/1674-1056/adea56

## A novel baseline perspective visibility graph for time series analysis

Huang-Jing Ni(倪黄晶), Zi-Jie Song(宋紫婕), Jiao-Long Qin(秦姣龙), Ye Wu(吴烨), Shi-Le Qi(戚世乐), and Ming Song(宋明)
Chin. Phys. B, 2025, 34 (**8**): 080504. DOI: 10.1088/1674-1056/adce9a

## Model-free prediction of chaotic dynamics with parameter-aware reservoir computing

Jianmin Guo(郭建敏), Yao Du(杜瑶), Haibo Luo(罗海波), Xuan Wang(王晅), Yizhen Yu(于一真), and Xingang Wang(王新刚)
Chin. Phys. B, 2025, 34 (**4**): 040505. DOI: 10.1088/1674-1056/adb733

## Associated network family of the unified piecewise linear chaotic family and their relevance

Haoying Niu(牛浩瀛) and Jie Liu(刘杰)
Chin. Phys. B, 2025, 34 (**4**): 040503. DOI: 10.1088/1674-1056/adb26a

## Impact of the chaotic semiconductor laser output power on time-delay-signature of chaos and random bit generation

Chenpeng Xue(薛琛鹏), Xu Wang(王绪), Likai Zheng(郑利凯), Haoyu Zhang(张昊宇), Yanhua Hong, and Zuxing Zhang(张祖兴)
Chin. Phys. B, 2025, 34 (**3**): 030504. DOI: 10.1088/1674-1056/ada7dd

---

# Multiscale structural complexity analysis of the Chinese classics *A Dream of Red Mansions* and *All Men Are Brothers*

Jing Feng(冯靖)[1], Ping Wang(王萍)[2], and Changgui Gu(顾长贵)[3,†]

[1] *Higher Vocational and Technical College, Shanghai University of Engineering Science, Shanghai* 200437, *China*
[2] *School of Science, Jiangsu University of Science and Technology, Zhenjiang* 212100, *China*
[3] *Business School, University of Shanghai for Science and Technology, Shanghai* 200092, *China*

(Received 11 October 2025; revised manuscript received 4 November 2025; accepted manuscript online 17 November 2025)

Text, as a fundamental carrier of human language and culture, exhibits high structural and semantic complexity. Its systematic analysis is essential for understanding linguistic patterns and cultural transmission. *A Dream of Red Mansions* and *All Men Are Brothers*, two masterpieces of Chinese classical literature, have long been central to debates regarding the authorship of their later chapters. Previous studies, often based on word-frequency statistics, function word distributions, entropy measures, and complex network analyses, have provided valuable insights into stylistic differences; however, they remain limited in capturing cross-scale structural features. To address this gap, we apply a multi-scale structural complexity approach based on character-frequency time series to analyze the structural evolution of both novels under various segmentation strategies. Our results reveal significant differences in peak complexity positions, overall complexity levels, and intra-textual variations between the two works, which are closely linked to changes in authorship and stylistic patterns. This study not only provides new quantitative evidence for resolving authorship disputes in classical literature but also demonstrates, from the perspective of structural complexity, the profound depth and unique charm of Chinese literary expression, highlighting the richness of Chinese language and culture. Moreover, it emphasizes the potential of structural complexity analysis as a versatile tool for textual analysis and style attribution.

**Keywords:** a dream of red mansions, all men are brothers, word-frequency, multi-scale structural complexity

**PACS:** 05.45.Tp, 05.45.Xt

**DOI:** 10.1088/1674-1056/ae1fe9 **CSTR:** 32038.14.CPB.ae1fe9

## 1. Introduction

Human language, as a highly complex social system, is the product of a long-term interplay between cultural development and biological evolution.[1] With the continuous advancement of statistical theory and the rapid growth of computational techniques, scholars have carried out systematic and in-depth studies of linguistic systems, producing important achievements in several directions.[2] For example, word-frequency distributions have been shown to follow Zipf's law in a wide variety of languages,[3–7] not only in modern languages such as English and French, but also across diverse linguistic contexts. In addition, scale-free properties have been consistently observed at the levels of words and letters,[8–10] indicating a strong self-organizing nature of language. Complex network theory has also been employed to characterize the topological structures of linguistic systems, uncovering relational networks among words and enabling comparative studies and classification of languages.[11–14] These approaches have significantly advanced our understanding of the universal principles of language. However, when applied to Chinese classical novels, which are lengthy, structurally layered, and linguistically unique, such methods often fall short in capturing cross-scale organizational features, thereby limiting a comprehensive representation of their internal complexity.

Chinese classical literature, as an important carrier of China's profound traditional culture, embodies not only immense artistic value but also deep-rooted national spirit and cultural heritage. Its linguistic richness and hierarchical structure offer an ideal context for exploring the complexity of the Chinese language from a multi-scale perspective. Among these works, *A Dream of Red Mansions* (ADRM) and *All Men Are Brothers* (AMAB) stand as monumental masterpieces, renowned for their expansive scope, artistic sophistication, and profound cultural influence.

Nevertheless, the authorship and division of labor in the creation of these works have long been subjects of scholarly debate, forming a central topic in literary studies. ADRM was originally published anonymously, leaving the question of authorship unresolved until well into the early 20th century.[15] The prevailing view attributes the first 80 chapters to Xueqin Cao and the remaining 40 chapters (Chapters 81–120) to E Gao, yet this attribution remains contested.[16] The case of AMAB is even more intricate: most scholars agree that the first 70 chapters were written by Nai'an Shi, while the final 30 chapters may have been added by Guanzhong Luo.[17] Investigating these questions not only addresses a fundamental issue

†Corresponding author. E-mail: gu_changgui@163.com

© 2026 Chinese Physical Society and IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

in literary history but also offers valuable insights into the evolution of textual style, the identification of authorship, and the transmission of literary styles across historical periods.

In studies of textual style, lexical choices, syntactic structures, and function words have been widely recognized as key indicators of authorial habits.[18–20] Word-frequency analysis remains the most common technique, where frequent and distinctive words are compared across text segments to highlight stylistic differences.[21] Entropy-based measures have been applied to examine textual rhythm and information distribution, offering insights into structural variability.[22] Function words, due to their high frequency and independence from thematic content, have proven to be reliable markers of writing style.[23] These methods are valued for their simplicity, interpretability, and statistical robustness. Nonetheless, they are mostly limited to the lexical or syntactic level and often fail to capture deeper organizational patterns, such as repetitive motifs or thematic transitions. This limitation becomes particularly evident in the case of monumental works such as ADRM and AMAB.[21]

In recent years, methods inspired by statistical physics and complex network theory have opened new perspectives for textual analysis. These approaches emphasize macro-level structures of texts. For instance, researchers have constructed character co-occurrence networks and chapter connectivity graphs to study narrative frameworks, community structures, and topological properties.[24–29] With the development of deep learning, graph neural networks have also been introduced to model and visualize literary structures, offering new ways to interpret narrative organization.[30–32] While these approaches demonstrate great potential in analyzing story structures and large-scale organization, they primarily focus on relational networks among characters, rather than quantifying linguistic complexity or tracking subtle, longitudinal shifts in authorial style. Thus, developing methods that can capture both fine-grained details and macro-level patterns across scales remains a pressing challenge.

To overcome the limitations of existing methods in multi-scale linguistic structure analysis, this study employs a multi-scale structural complexity (MSC) framework based on character-frequency time series.[33–35] This method is developed with reference to the approach proposed by Bagrov *et al.*, which was originally designed to describe the structural complexity of image data across multiple spatial scales. In our adaptation, the text is first converted into a time series that records the frequency of each Chinese character. A coarse-graining scale $k$ is then applied to measure the structural complexity at different resolutions, where each scale reflects a specific level of organization within the text. This framework enables the detection of both fine-grained repetitive patterns and macro-level organizational features, thereby revealing the hierarchical nature of linguistic structures.

Unlike traditional complexity measures such as entropy, fractal dimension, or synchrony, which primarily quantify randomness or self-similarity, multi-scale structural complexity emphasizes the interactions between local and global features and captures how these elements jointly shape the overall organization of the system. Thus, it provides a more comprehensive description of linguistic organization across scales. This capability is particularly valuable for lengthy and hierarchically structured works, such as Chinese classical novels, where cross-scale structural patterns are essential to understanding narrative composition and stylistic variation.

In this study, the method is applied to ADRM and AMAB under multiple segmentation schemes: full text, division into two parts, three parts, and segments of 10 chapters. This multi-level framework allows for a thorough characterization of structural differences and helps pinpoint key locations that may be related to changes in authorship or writing style. The remainder of this paper is organized as follows. Section 2 details data sources, text preprocessing steps, and the method of multi-scale structural complexity. Section 3 presents the empirical results. Finally, Section 4 concludes by summarizing contributions, outlining methodological limitations, and proposing directions for future research.

## 2. Data and methods

### 2.1. Data

ADRM and AMAB are two of the Four Great Classical Novels of China. The textual data analyzed in this study were obtained from digitized editions available on Wikisource (https://zh.m.wikisource.org/wiki/Wikisource/). The ADRM corpus contains 120 chapters, and the AMAB corpus corresponds to a 100-chapter edition. The raw texts were collated and proofread before conducting a character-frequency analysis for each novel. Key metadata of the two corpora are summarized in Table 1, where $N_c$ and $N_w$ denote the number of chapters and the total number of Chinese characters, respectively.

**Table 1.** Key metadata of the textual data.

| Corpus | Chapter total: $N_c$ | Word total: $N_w$ |
|---|---|---|
| ADRM | 120 | 726158 |
| AMAB | 100 | 669232 |

The dataset used for analysis is a time series representing the frequency of each Chinese character in the text. Punctuation marks were excluded, so that the series counts only Chinese characters. As the only continuously used writing system in the world for several millennia, Chinese characters embody the unity of form, sound, and meaning; they

also reflect the continuity and vitality of Chinese civilization. Through the structural analysis of character-frequency sequences, this study reveals the mathematical characteristics of the texts while also highlighting the unique advantage of Chinese characters in encoding complex narrative and cultural information. This intrinsic property of the Chinese writing system provides a solid foundation for applying quantitative methods to explore the structural complexity of classical literature. Accordingly, the constructed time series consist of 726158 and 669232 characters for ADRM and AMAB, respectively.

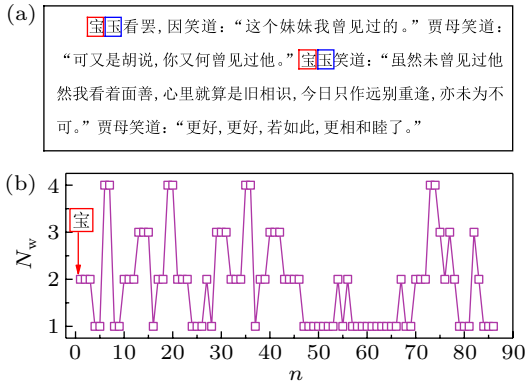



**Fig. 1.** The procedure for converting text into a time series. Panel (a) presents a sample excerpt from ADRM. Panel (b) illustrates the corresponding occurrence sequence of each character from panel (a), where $n$ denotes the sequential order of each character's appearance across segments, and $N_w$ represents its total occurrences in the entire text.

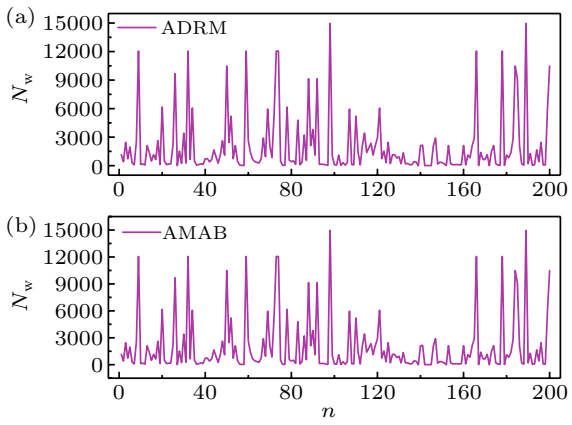



**Fig. 2.** The first 200 data points of the occurrence sequences constructed from the total occurrences of each Chinese character in the complete texts of ADRM and AMAB. (a) The sequence corresponding to ADRM. (b) The sequence corresponding to AMAB. Here, $n$ denotes the sequential order of each character's appearance across segments, and $N_w$ represents its total occurrences in the text.

The construction process can be illustrated using a short excerpt from ADRM (Fig. 1(a)). The first character (highlighted in the red box) appears twice in the entire novel, so that the first point in the series is $t_1 = 2$. The second character (highlighted in the blue box) also appears twice, giving $t_2 = 2$, and so on. For an excerpt of 86 characters, this procedure yields a sequence of length 86 (Fig. 1(b)). Figure 2 presents the first 200 points of the sequence for ADRM and AMAB generated by this method.

### 2.2. Method

In this study, we employed the method of multi-scale structural complexity analysis based on scalar time series data (S-MSC) to investigate the complexity of the stroke counts and semantic definitions in the Four Great Classical Novels of China.[34] This method represents an extension of the image-based structural complexity approach originally proposed by Bagrov *et al.*,[33] enabling the extraction of intrinsic structural complexity from purely one-dimensional time series data. This method does not require high-dimensional or image-based inputs. The entire approach consists of two main steps: (i) transforming the time series into a two-dimensional matrix representation; (ii) performing multi-scale structural analysis on the resulting matrix to quantify how complexity varies across scales.

Each character-frequency time series was first normalized to the range [0, 1] using the min-max method. This normalization eliminates the influence of absolute frequency differences between texts, ensuring that the calculated complexity values reflect structural characteristics rather than variations in overall frequency magnitude.

**Step 1. Transformation of the time series into matrix $T_0$**

Let $X = \{X(1), X(2), \ldots, X(S)\}$ be a scalar time series of length $S$, and let $\tau$ denote a fixed delay, typically set to 1 unless otherwise specified. We construct two derived sequences: the truncated time series $X_1 = \{X(1), X(2), \ldots, X(S-\tau)\}$ and the delayed time series $X_2 = \{X(1+\tau), X(2+\tau), \ldots, X(S)\}$. These two sequences are jointly embedded in a two-dimensional coordinate system, where the horizontal axis corresponds to $X_1$ and the vertical axis to $X_2$. Each axis is partitioned into $L$ equal segments, yielding an $L \times L$ grid. For each cell in the grid, we compute the mean of all data points falling within it. The resulting matrix $T_0(i, j)$ records the average value for the $(i, j)$th cell. This matrix captures the underlying spatial distribution of the scalar time series in a two-dimensional form.

**Step 2. Multi-scale analysis of matrix $T_0$**

To capture structure at different resolutions, we perform a multi-scale decomposition of matrix $T_0$ through coarse-graining. Specifically, for each scale $k$, we generate a new matrix $T_k$ by averaging over non-overlapping square patches of increasing size.

At the first scale ($k = 1$), the matrix $T_0$ is sequentially covered by non-overlapping modules of size $\xi \times \xi$ (with $\xi = 2$ in this study), and each module is replaced by the average of its internal elements to generate the matrix $T_1$. This operation preserves the original dimensions of the matrix ($L \times L$) but reduces its structural detail. The process is applied recursively: at scale $k$ ($1 \le k \le \log_\xi L$), each element $T_k(i, j)$ is determined

by averaging the values within the corresponding $\xi^k \times \xi^k$ region in the matrix $T_{k-1}$. The general recursive formula is

$$T_k(i,j) = \frac{1}{\xi^k \times \xi^k} \sum_{m=1}^{\xi^k} \sum_{l=1}^{\xi^k} T_{k-1} \times \left( \left\lfloor \frac{i-1}{\xi^k} \right\rfloor \cdot \xi^k + m, \left\lfloor \frac{j-1}{\xi^k} \right\rfloor \cdot \xi^k + l \right), \quad (1)$$

where "$\lfloor \cdot \rfloor$" denotes the floor function. This process continues until the entire matrix is averaged into a single block. The resulting matrices $\{T_1, T_2, \ldots, T_n\}$ reflect the system's structure at successively coarser resolutions.

The coarsening process is governed by a scaling parameter $\xi$. Here, $\xi^n$ represents the module edge length used to cover the matrix during the $n$th renormalization step, with values greater than 1 and restricted to integers. For computational convenience, we selected $\xi = 2$ in this study. This choice preserves clear inflection points in $C_k$ and is sufficient for the SCN system. For other systems, we recommend starting with $\xi = 2$; if no inflection points are observed, the value of $\xi$ can be increased.

**Step 3. Quantification of complexity across scales**

To evaluate structural changes across scales, we define the local structural complexity $C_k$ at scale $k$ as the difference between $T_{k-1}$ and $T_k$. Specifically, we define a local operator

$$O_k = -\frac{1}{2}(T_k - T_{k-1}) \cdot (T_k - T_{k-1}) = T_k \cdot T_{k-1} - \frac{1}{2}(T_k \cdot T_k - T_{k-1} \cdot T_{k-1}),$$

where "$\cdot$" denotes element-wise multiplication. Using this, the local complexity at scale $k$ is computed as

$$C_k = \left| \sum_{i=1}^{L} \sum_{j=1}^{L} O_k(i,j) \right|. \quad (2)$$

Here, $C_k = 0$ indicates a completely regular segment with no structural complexity, while $C_k > 0$ reflects the presence of local structural complexity, with larger values corresponding to higher complexity. Finally, the total complexity $C$ of the system is obtained by summing over all local complexities

$$C = \sum_{k=1}^{\log_\xi L} C_k. \quad (3)$$

## 3. Results

### 3.1. The results for the full text

We analyzed the time series derived from the full texts of ADRM and AMAB, whose structural complexities are shown in Fig. 3. The figure depicts the variation of structural complexity $C_k$ with scale $k$ for character-frequency time series constructed by counting the occurrences of each Chinese character in the complete texts. As $k$ increases, representing a higher degree of coarse-graining, both works exhibit a roughly parabolic trend in $C_k$, with complexity initially rising and then declining. However, clear differences emerge between the two novels: the peak complexity of ADRM occurs at $k = 5$, whereas that of AMAB appears at $k = 4$. Furthermore, at all scales, AMAB consistently exhibits higher complexity values than ADRM.

These differences likely reflect underlying variations in textual structure and linguistic style. The shift in peak position indicates that the characteristic repetition cycles of semantic or structural units differ between the texts: a coarse-graining scale of $k = 5$ captures the longer recurring patterns in ADRM, whereas $k = 4$ is more suited to AMAB. Additionally, the consistently higher complexity observed in AMAB may result from greater diversity and uneven distribution in narrative pacing, character appearances, and word usage. This variability enables more structural information to be preserved after coarse-graining, leading to higher complexity values across all scales.

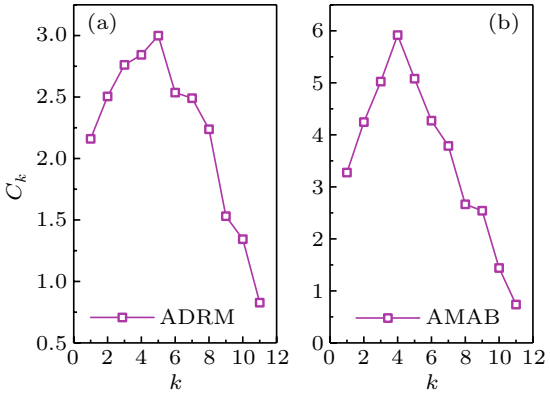



**Fig. 3.** The local complexity $C_k$ of the complete texts of ADRM and AMAB at different scales $k$. Panel (a) shows the local complexity distribution of ADRM, while panel (b) shows that of AMAB. The parameter $k$ represents the analysis scale, with each value corresponding to a distinct level of structural organization in the text.

### 3.2. The results for the text divided into two parts

We further divided the time series corresponding to ADRM and AMAB into two sequential parts. For ADRM, the text is divided into two parts based on both structural and historical considerations: the first 80 chapters, generally attributed to Xueqin Cao, and the remaining 40 chapters, likely completed by E Gao.[16] Similarly, for AMAB, the first 70 chapters are generally attributed to Nai'an Shi, while the latter 30 chapters may have been added by Guanzhong Luo.[17] This segmentation allows us to analyze the local complexity of each part separately, capturing potential differences in writing style, structural features, and linguistic patterns between the original and later sections. This segmentation aligns with standard practices in time series and complex network analysis, where long sequences are partitioned to reveal local variations that might be obscured in the overall analysis.[36] Figure 4 presents

the variation of structural complexity $C_k$ with scale $k$ for each part.

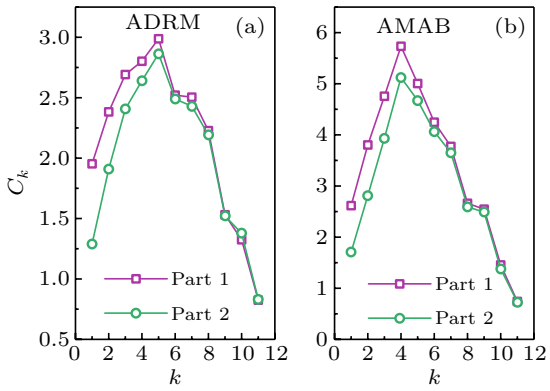



**Fig. 4.** The variation of local complexity $C_k$ for ADRM and AMAB when the texts are divided into two sequential parts. Panel (a) shows $C_k$ for ADRM, with the purple curve representing the first 80 chapters and the green curve representing the last 40 chapters. Panel (b) shows $C_k$ for AMAB, with the purple curve representing the first 70 chapters and the green curve representing the last 30 chapters.

The results reveal that: (i) For both ADRM and AMAB, $C_k$ exhibits a parabolic-like pattern with increasing $k$, first rising and then falling, consistent with the whole-text results in Fig. 3. The peak complexity appears at $k = 5$ for ADRM and at $k = 4$ for AMAB. (ii) Across all scales, AMAB consistently shows higher complexity values than ADRM. (iii) Noticeable differences exist between part 1 and part 2 for both novels, especially at smaller scales, with the difference being especially pronounced in ADRM.

These findings may be linked to differences in writing style caused by changes in authorship. Historical records indicate that the first 80 chapters of ADRM were written by Xueqin Cao, while the final 40 chapters were completed by E Gao, whose narrative rhythm and patterns of lexical repetition, as reflected in structural complexity, differ considerably from those of Cao. In contrast, the first 70 chapters of AMAB were authored by Nai'an Shi and the last 30 by Guanzhong Luo, who is believed to have been Shi's student, likely resulting in greater stylistic consistency. Such differences are particularly evident at smaller scales, reflecting differences in local word usage and structural patterns.

### 3.3. The results for the text divided into three parts

To further verify the robustness of the previous observations, we divided the time series of ADRM and AMAB into three sequential parts: for ADRM, part 1 corresponds to the first 40 chapters, part 2 to the middle 40 chapters, and part 3 to the final 40 chapters; for AMAB, part 1 includes the first 35 chapters, part 2 the middle 35 chapters, and part 3 the final 30 chapters. Figure 5 presents the variation of structural complexity $C_k$ with scale $k$ for each part.

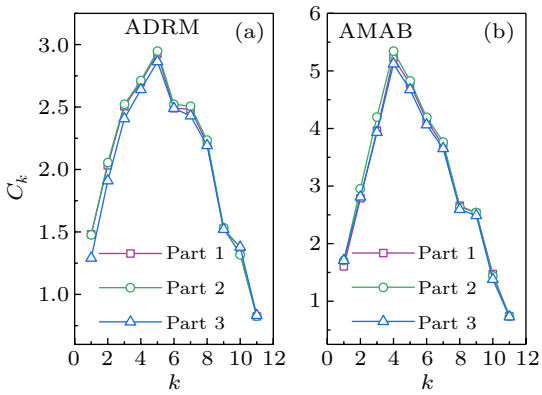



**Fig. 5.** The variation of local complexity $C_k$ for ADRM and AMAB when the texts are divided into three sequential parts. Panel (a) shows $C_k$ for ADRM, with the purple, green, and blue curves representing the first 40, middle 40, and last 40 chapters, respectively. Panel (b) shows $C_k$ for AMAB, with the purple, green, and blue curves representing the first 35, middle 35, and last 30 chapters, respectively.

### 3.4. The results for the text divided into parts of 10 chapters each

The results reveal that: (i) For both ADRM and AMAB, $C_k$ exhibits a parabolic-like pattern with increasing $k$ — first rising and then declining — consistent with the overall trends in Figs. 3 and 4, with peak positions remaining at $k = 5$ for ADRM and $k = 4$ for AMAB. (ii) At all scales, AMAB maintains higher complexity values than ADRM, in agreement with the findings from the earlier figures. (iii) In both works, the curves for part 1 and part 2 nearly overlap, indicating similar structural complexity characteristics, which can be attributed to having been written by the same author. In contrast, part 3 differs noticeably from the first two parts, particularly at smaller scales, with this difference being slightly larger in ADRM than in AMAB. This likely reflects changes in narrative rhythm, character occurrence frequency, and lexical patterns in the later chapters. Specifically, the final 40 chapters of ADRM were completed by E Gao, whose writing style differs from that of the original author Xueqin Cao, whereas the final 30 chapters of AMAB were completed by Guanzhong Luo, who had a close literary relationship with Nai'an Shi, likely resulting in smaller stylistic differences, and consequently, weaker complexity changes across parts in AMAB.

Building on the preceding analyses of the whole texts, the two-part division, and the three-part division, we further applied a finer segmentation scheme. Specifically, ADRM and AMAB were partitioned into 10-chapter segments: 12 segments for ADRM and 10 segments for AMAB. For each segment, we computed the local complexity $C_k$, and then averaged contiguous segments to form three parts: for ADRM, the averages over segments 1–4, 5–8, and 9–12 define part 1, part 2, and part 3, respectively; for AMAB, the averages over segments 1–3, 4–7, and 8–10 define part 1, part 2, and part 3, respectively.

For ADRM (Fig. 6), the curves for part 1 and part 2 are

nearly identical across most scales, with a marked deviation of part 3 from the first two parts only at $k = 7$ and $k = 10$. This suggests that the first eight segments share similar structural complexity patterns, while the last four segments display distinct structural features at certain scales, likely related to the fact that the final 40 chapters were authored by a different writer.

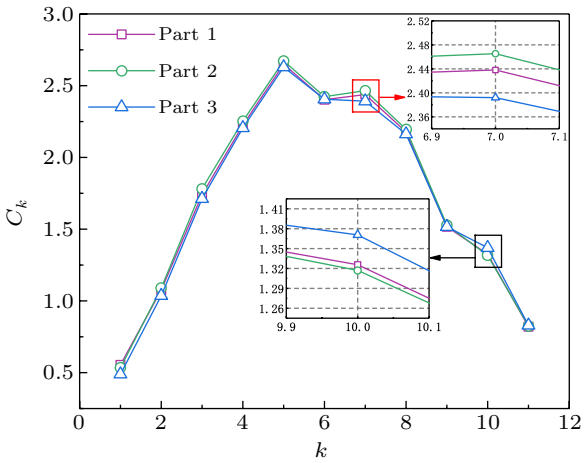



**Fig. 6.** The variation of local complexity $C_k$ with scale $k$ for ADRM, when the text is divided sequentially into 10-chapter segments. The purple, green, and blue curves represent the average $C_k$ for the first 40, middle 40, and final 40 chapters, respectively.

For AMAB (Fig. 7), the three parts exhibit some variation but without a clear or consistent trend. In light of the results from Figs. 4(b) and 5(b), this lack of pronounced divergence may reflect the apprentice–master relationship between the two authors of AMAB, leading to greater stylistic similarity as reflected in structural complexity. Moreover, the reduced length of each segment after division may limit the ability to detect differences between part 3 and the other two parts at finer scales.

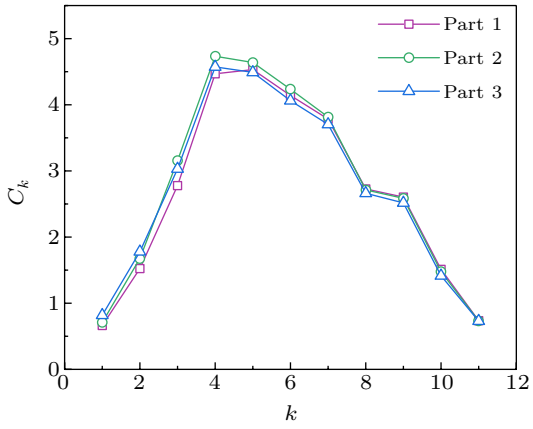



**Fig. 7.** The variation of local complexity $C_k$ with scale $k$ for AMAB, when the text is divided sequentially into 10-chapter segments. The purple, green, and blue curves represent the average $C_k$ for the first 30, middle 31 (chapters 40–70), and final 40 chapters, respectively.

## 4. Discussion and conclusion

Complexity analysis of textual data has been widely applied in literature, historical documents, and linguistic studies to uncover structural features and patterns of information distribution.[37] As the saying goes, culture is the soul of a country and a nation,[38] providing a value orientation for our research on classical literary works such as ADRM and AMAB, and enabling a deeper appreciation of the unique charm and profound heritage of China's traditional culture when analyzing the structural and semantic complexity of these works. Previous approaches, such as word-frequency statistics, entropy measures, and network-based models, have achieved significant success in quantifying textual properties and distinguishing between different genres or authorial styles.[28,29,39] However, these methods face limitations when applied to lengthy Chinese classical novels, particularly in capturing structural features across multiple scales and detecting recurring patterns at different hierarchical levels. Such constraints hinder a precise characterization of structural variation within long literary works.

To address these limitations, this study applies a multiscale structural complexity framework to two of China's Four Great Classical Novels: ADRM and AMAB. Using character-frequency time series and analyzing their structural complexity $C_k$ across different coarse-graining scales $k$, we systematically examine differences in both global and local structures. To ensure a comprehensive assessment, we investigate not only the full texts but also progressively finer segmentations, including a two-part division, a three-part division, and a 10-chapter segment-based division, thereby validating the stability and patterns of the complexity features from multiple perspectives.

The whole-text analysis shows that the structural complexity of both novels exhibits a parabolic-like trend with increasing scale $k$, yet their peak positions and overall complexity levels differ markedly: ADRM peaks at $k = 5$, while AMAB peaks at $k = 4$, and AMAB consistently shows higher complexity values at all scales. These differences suggest variations in the characteristic repetition cycles of semantic or structural units, with AMAB demonstrating greater diversity in narrative rhythm, character distribution, and vocabulary usage, thereby retaining more structural information across scales.

Local analyses further reveal structural contrasts. In both the two-part and three-part divisions, AMAB exhibits smaller inter-part differences than ADRM, consistent with historical accounts indicating that the two authors of AMAB had a master–apprentice relationship. In contrast, ADRM shows more pronounced differences between parts, likely reflecting the change in authorship between its first 80 and final 40 chapters. The finer 10-chapter segmentation further confirms these observations: ADRM's part 3 diverges notably from the first two parts at specific scales (e.g., $k = 7$ and $k = 10$), whereas AMAB's three parts differ only slightly and without a clear

pattern, again supporting the hypothesis of greater stylistic consistency in AMAB.

This study offers a novel methodological perspective for literary analysis by employing multi-scale structural complexity, providing empirical insights into the stylistic and structural diversity of two classical masterpieces. These findings contribute to quantitative stylistics, authorship attribution, and the study of textual organization in classical literature. However, the current analysis relies solely on character-frequency sequences and does not yet incorporate higher-level semantic information. Future work could integrate semantic networks, topic modeling, or syntactic analysis with multi-scale complexity measures. Such integration would help uncover deeper structural regularities and extend the applicability of this approach to a wider range of texts and languages.

## Acknowledgments

Project supported by the National Natural Science Foundation of China (Grant Nos. 12275179, 11875042, and 12150410309) and the Natural Science Foundation of Shanghai (Grant No. 21ZR1443900).

## References

[1] Kwapień J and Drożdż S 2012 *Phys. Rep.* **515** 115
[2] Grabska-Gradzińska I, Kulig A, Kwapień J, Oświecimka P and Drożdż S 2012 arXiv:1212.3171
[3] Zipf G K 1949 *Human Behavior and The Principle of Least Effort* (Cambridge Mass: Addison-Wesley) p. 573
[4] Mehri A and Jamaati M 2017 *Phys. Lett. A* **381** 2470
[5] Nabeshima T and Gunji Y P 2004 *Biosystems* **73** 131
[6] Lü L Y, Zhang Z K and Zhou T 2013 *Sci. Rep.* **3** 1082
[7] Deng W B and Pato M P 2017 *Physica A* **481** 167
[8] Altmann E G, Cristadoro G and Esposti M D 2012 *Proc. Natl. Acad. Sci. USA* **109** 11582
[9] Montemurro M A and Pury P A 2002 *Fractals* **10** 451
[10] Bhan J, Kim S, Kim J, Kwon Y, Yang S and Lee K 2006 *Chaos Solitons Fractals* **29** 69
[11] Cong J and Liu H T 2014 *Phys. Life Rev.* **11** 598
[12] Cancho R F I and Sole R V 2001 *Proc. R. Soc. Lond. B* **268** 2261
[13] Kulig A, Drożdż S, Kwapień J and Oświecimka P 2015 *Phys. Rev. E* **91** 032810
[14] Abramov O and Mehler A 2011 *J. Quant. Linguist.* **18** 291
[15] Zhou R C 2009 *Between Noble and Humble: Cao Xueqin and the Dream of the Red Chamber* edited by R G Ronald and S F Mark (New York: Peter Lang)
[16] Yang Y, Gu C G, Xiao Q and Yang H J 2017 *PLoS One* **12** e0171776
[17] Lü N Y 2008 *J. Peking Univ. (Philos. Soc. Sci.)* **2** 68
[18] Zheng W W and Jin M Z 2023 *WIREs Comput. Stat.* **15** e1584
[19] Jafariakinabad F, Tarnpradab S and Hua K A 2020 *Syntactic Neural Model for Authorship Attribution* In *Proceedings of the Thirty Third International Flairs Conference*, North Miami Beach, FL, USA pp. 234–239
[20] Lagutina K, Lagutina N, Boychuk E, Larionov V and Paramonov I 2021 *Authorship Verification of Literary Texts with Rhythm Features* In *Proceedings of the 2021 28th Conference of Open Innovations Association (FRUCT)*, Moscow, Russia
[21] Arefin A S, Vimieiro R, Riveros C, Craig H and Moscato P 2014 *PLoS ONE* **9** e111445
[22] Estevez-Rams E, Mesa-Rodriguez A and Estevez-Moya D 2019 *PLoS ONE* **14** e0214863
[23] Kulig A, Kwapień J, Stanisz T and Drożdż S 2017 *Inf. Sci.* **375** 98
[24] Kumar R and Singh P 2021 *Physica A* **573** 125994
[25] Sigman M and Cecchi G A 2002 *Proc. Natl. Acad. Sci. USA* **99** 1742
[26] Masucci A and Rodgers G 2006 *Phys. Rev. E* **74** 026102
[27] Liang W, Shi Y M, Tse C K and Wang Y L 2012 *Sci. China Inf. Sci.* **55** 2417
[28] Liu H T 2008 *Physica A* **387** 3048
[29] Liu Z, Zheng Y and Sun M 2008 *Complexity* **5** 1
[30] Tang X 2024 *Heliyon* **10** e25464
[31] Zhao L M, Shi J J, Zhang C K and Liu Z X 2025 *Appl. Sci.* **15** 1677
[32] Van Houdt G, Mosquera C and Nápoles G 2020 *Artif. Intell. Rev.* **53** 5929
[33] Bagrov A A, Iakovlev I A, Iliasov A A, Katsnelson M I and Mazurenko V V 2020 *Proc. Natl. Acad. Sci. USA* **117** 30241
[34] Wang P, Gu C G, Yang H J, Wang H Y and Moore J M 2023 *Physica A* **609** 128358
[35] Wang P, Gu C G, Yang H J and Wang H Y 2022 *Chaos Solitons Fractals* **164** 112721
[36] Wang P, Gu C G, Yang H J and Wang H Y 2022 *Electron. Res. Arch.* **30** 3660
[37] Febres G and Jaffé K 2014 *Quantifying Literature Quality Using Complexity Criteria* ArXiv (Cornell University)
[38] Xi J P 2019 *Qiushi* **12**
[39] Febres G and Jaffé K 2016 *J. Quant. Linguist.* **24** 16