Nature and Information Engineering Publishing

Journal of Language Service Studies

https://journals.niepublish.com/jlss

REVIEW

Quantifying Punctuation Patterns in Chinese Language for Language Service Applications

Jakub Dec [1] ***, Michał Dolina*** [1] ***, Stanisław Drożdż*** [1,2*] ***, Jarosław Kwapień*** [2] ***, Jin Liu*** [3] ***, Tomasz Stanisz*** [2]

[1] *Faculty of Computer Science and Mathematics, Cracow University of Technology, 31-155 Kraków, Poland*

[2] *Complex Systems Theory Department, Institute of Nuclear Physics, Polish Academy of Sciences, 31-342 Kraków, Poland*

[3] *School of Modern Languages, Georgia Institute of Technology, Atlanta, GA 30332, USA*

ABSTRACT

Punctuation is commonly treated as an auxiliary feature of writing, yet it encodes crucial information about syntactic organization, discourse rhythm, and narrative structure. In modern Chinese, punctuation forms a hybrid system shaped by traditional practices and imported Western conventions, making it especially relevant for quantitative analysis and language service applications. This review presents a concise summary of recent findings of a systematic investigation of punctuation in modern Chinese prose, with comparison to English translations, using a unified framework of the Zipf–Mandelbrot frequency statistics, the discrete Weibull spacing distributions, word-adjacency network analysis, and multifractal time-series methods. If punctuation marks are treated along with regular words as full linguistic tokens, Chinese punctuation follows robust power-law frequency scaling and its inclusion improves Zipfian behaviour. Inter-punctuation distances exhibit near-exponential Weibull distributions, reflecting frequent and regular segmentation, while English displays heavier tails indicative of longer punctuation-free spans. Network analyses reveal that punctuation functions as a central organizing hub in Chinese texts, reducing average shortest-path lengths and increasing clustering, whereas English networks are more lexically centered. Multifractal analysis further demonstrates that punctuation governs narrative dynamics across scales, with English translations exhibiting stronger multiscale variability. These results establish punctuation as a quantifiable

*CORRESPONDING AUTHOR:

Stanisław Drożdż, Faculty of Computer Science and Mathematics, Cracow University of Technology, 31-155 Kraków, Poland; Complex Systems Theory Department, Institute of Nuclear Physics, Polish Academy of Sciences, 31-342 Kraków, Poland; Email: stanislaw.drozdz@ifj.edu.pl

ARTICLE INFO

Received: 23 December 2025 | Revised: 19 May 2026 | Accepted: 27 May 2026 | Published Online: 4 June 2026

DOI: https://doi.org/10.63385/jlss.2i2.100997

CITATION

Dec, J., Dolina, M., Drożdż, S., et al., 2026. Quantifying Punctuation Patterns in Chinese Language for Language Service Applications. Journal of Language Service Studies. 2(2): 1–21. DOI: https://doi.org/10.63385/jlss.2i2.100997

COPYRIGHT

Copyright © 2026 by the author(s). Published by Nature and Information Engineering Publishing Sdn. Bhd. This is an open access article under the Creative Commons Attribution 4.0 International (CC BY 4.0) License (https://creativecommons.org/licenses/by/4.0/).

structural signal rather than a peripheral orthographic feature and highlight its practical relevance for machine translation, speech technology, subtitling, accessibility services, and data-driven quality assurance.

Keywords: Chinese Punctuation; Quantitative Linguistics; Language Services; Zipf's Law; Weibull Distribution; Machine Translation; Multifractality; Complex Networks

1. Background

The past decade has witnessed an accelerating transformation of the language service industry, propelled by artificial intelligence (AI), large language models (LLMs), and global digitalization. Translation, subtitling, localization, and speech-to-text applications are increasingly dependent on linguistic data and quantitative models. Within this context, understanding the structural regularities of the Chinese language—including its punctuation system—is vital for advancing intelligent language services. Unlike alphabetic languages, Chinese writing employs morphosyllabic characters rather than letters, and its punctuation marks were formalized only in the early twentieth century. The hybrid system that emerged combines Western marks such as commas, periods, and quotation marks with traditional symbols like the ideographic full stop (。). This complex mixture reflects the evolution of written communication in China and shapes how meaning and rhythm are conveyed.

Previous studies have demonstrated that statistical laws—such as Zipf's[1, 2] or Heaps and Herdan's[3, 4] laws—govern lexical distributions across languages. Recent research[5–7] extended this approach to punctuation, showing that these marks also exhibit scaling properties and contribute to the overall topology of linguistic networks. However, the practical implications of these findings for language service applications remain underexplored. Language services encompass all activities that facilitate multilingual communication—translation[8, 9], interpretation, localization, language technology[10], and accessibility support[11]. The field increasingly seeks empirical evidence to improve AI systems and service quality.

This review paper aims to bridge the gap between quantitative linguistics and applied language service research by analyzing punctuation patterns in modern Chinese texts and discussing their implications for intelligent translation and communication technologies. More specifically, the goal of this paper is to concisely synthesize the most important findings and conclusions from our previous works, published in more specialized journals, which addressed the issue of understanding punctuation marks as an integral part of natural language in its written form, on equal footing with regular words[6, 7, 12]. What we propose is that words and punctuation should be treated as equivalent, which originates from statistical arguments such as the Zipf-Mandelbrot law, topological and statistical properties of token co-occurrence networks involving both words and punctuation marks, as well as the fractal characteristics of sequences of functional text elements, such as sentences and clauses. In this context, particular emphasis is placed on comparing the properties of languages representing radically different grammatical and writing systems, i.e., Chinese and English. As a review paper aimed more at the social sciences and humanities community, it shall summarize what has been learned so far and put these findings into the context of language services, which, to the best of our knowledge, is the first approach of this kind reported in the literature.

We shall stress the fact that references to different kinds of language services in the latter part of this work are intended solely as general qualitative observations, aimed at indicating possible future applications of our results and conclusions. Description or implementation of methodologies specific to the field of language services is beyond the scope of this review. Rather, our intention is to offer a perspective that may encourage specialists in this domain to consider our findings and, potentially, also our methodological framework in their own research.

2. Methodology

2.1. Complex Systems and Quantitative Linguistics

The contemporary study of natural language increasingly adopts the perspective of complex systems science, which views language as a multilayered, hierarchically or-

ganized structure whose emergent properties cannot be reduced to its basic units alone[12, 13]. A text is not merely a sequence of discrete word-like elements; instead, it exhibits long-range correlations, fractal or even multifractal structure[14], and network-level organization arising from interactions among words, characters, punctuation marks, and discourse units[15, 16]. This view is extensively elaborated in modern linguistic complexity research, where natural language is shown to display hallmark features of complex systems: scale invariance[17–19], multiscaling[20], statistical regularities[21, 22], heavy-tailed distributions[23], and network small-world[24, 25] and scale-free properties[26, 27].

A cornerstone of quantitative linguistics is the Zipf–Mandelbrot law[28–31], which describes the rank frequency distribution:

$$f(r) = \frac{C}{(r+b)^{\alpha}} \tag{1}$$

capturing the empirical observation that linguistic tokens—traditionally words, but recently also punctuation marks—follow a power-law distribution (f—frequency, r—rank, α—the exponent of the Zipf-Mandelbrot law, close to 1, C and b—constants). Punctuation is now recognized as a crucial constituent stabilizing this scaling: its inclusion restores or markedly improves Zipfian behaviour for high-frequency ranks, where violations typically occur in standard word-only analyses. A second frequently cited linguistic relationship is expressed by the so-called Heaps' law, which captures how vocabulary size grows sublinearly with text length. This relation reflects the continual introduction of new lexical items as a text evolves, and is used in theoretical models of linguistic creativity and language growth[32].

Beyond frequency-based descriptions, temporal models of texts conceptualize them as symbolic sequences whose structural units (sentences, clauses, or segments between punctuation marks) form time series. Recent results[30, 33] demonstrate that the distances (waiting times) l between punctuation marks (i.e., distances between consecutive marks measured in words) follow a discrete Weibull distribution[34]—a discrete analogue of the widely known (continuous) Weibull distribution[35], encountered in modelling various phenomena, particularly in survival analysis[36, 37]. The probability mass function of the discrete Weibull distribution with parameters p and β is expressed as:

$$f(l) = (1-p)^{(l-1)^{\beta}} - (1-p)^{l^{\beta}}, p \in (0,1), \beta > 0. \tag{2}$$

The corresponding cumulative distribution function reads:

$$F(l) = 1 - (1-p)^{l^{\beta}} \tag{3}$$

and the hazard function[37]:

$$h(l) = 1 - (1-p)^{l^{\beta} - (l-1)^{\beta}}, \tag{4}$$

reflecting the probability of a punctuation mark occurring after a sequence of l words uninterrupted by punctuation. Thus, $h(1) = p$ determines the probability of a punctuation mark appearing already after the first word. The behaviour consistent with such functional forms appears universal across languages, including English, French, German, Italian, Polish, Russian, and also Chinese, as it has been confirmed independently in a dedicated Chinese punctuation study.

Finally, linguistic structures can be represented as word-adjacency networks, where tokens (words, characters, punctuation marks) form nodes and immediate co-occurrence relationships form edges[38]. These networks consistently exhibit heterogeneous degree distributions, short average shortest-path length, and strong clustering, properties reflecting the deep organization of language at syntactic and stylistic levels[39]. Network approaches have proven valuable in authorship attribution, stylometric analysis[40], and cross-linguistic comparison[41, 42]. Punctuation plays an important role in word-adjacency networks by reducing path lengths and enhancing network cohesion.

To summarize, complex systems and quantitative linguistics approaches provide a mathematical and statistical framework within which punctuation becomes analytically visible and theoretically indispensable.

2.2. Chinese Punctuation as a Linguistic System

Punctuation in Chinese has an original, developmental trajectory and functional profile distinct from those of Indo-European alphabetic writing systems. While the earliest preserved classical texts occasionally contained marks indicating pauses or syntactic junctures—attested as early as ca. 900 BCE[43]—Chinese punctuation remained largely non-standardized for more than two millennia. During the Han dynasty, around fifteen marks were in identifiable use, including explicit predecessors of the modern comma. By the late imperial period, Chinese manuscripts employed fifty

to sixty symbols—such as the pause mark "、", the circular sentence-final mark "。", and various annotation marks—but these were often placed in margins or interlinear spaces rather than integrated into the line of text. Classical Chinese writings, especially official documents and scholarly editions, typically lacked punctuation, transferring the task of segmentation to the reader.

A radical transformation occurred during the early twentieth century. Influenced by reform movements and the vernacular language campaign led by figures such as Hu Shih, Western-style punctuation—full stops, commas, quotation marks, exclamation marks, question marks—was introduced and gradually incorporated into the body of Chinese texts around the 1910s–1930s[44, 45]. Modern Chinese punctuation is thus a hybrid system: Western marks coexist with traditional sinographic features such as the ideographic full stop (。) and fullwidth punctuation.

Despite standardization, Chinese writing still exhibits distinctive structural properties that bear directly on quantitative modeling:

1. Absence of spaces: words are not explicitly delimited, so segmentation requires algorithmic tokenization and interacts with punctuation-based parsing.
2. Flexible sentence segmentation: modern Chinese authors frequently use extended comma-chained clauses, a tendency historically rooted in the lack of strict period-comma distinctions[46]. This leads to greater variability in sentence length distributions, especially at sentence-terminating punctuation marks, which deviate more strongly from Weibull-type behaviour than mid-sentence marks—a pattern confirmed by recent statistical studies.
3. Multiple types of sentence-final punctuation: the ideographic full stop, fullwidth question and exclamation marks, semicolon, and ellipsis all function as sentence delimiters, but with differing stylistic and pragmatic distributions.
4. Punctuation embedded as structural cues: marks play critical syntactic, semantic, and prosodic roles in the comprehension of morphosyllabic text, particularly given the high density of homophony and polysemy in Chinese.

These features generate measurable consequences for quantitative models. For instance, spacing distributions between punctuation marks in Chinese follow discrete Weibull patterns but typically display shorter tails than in English translations, reflecting fewer extremely long clauses or sentences. Likewise, the interplay between character-based writing and punctuation leads to complex, multifractal segmentation structures, especially visible in narrative texts such as *Soul Mountain*, where multifractal signatures arise from highly variable punctuation-mediated pacing. Understanding Chinese punctuation, therefore, requires a synthesis of historical evolution, orthographic structure, and quantitative regularities—an approach that aligns with the broader view of language as a complex system.

2.3. Language Services and Quantitative Modeling

Language services—encompassing translation, localization, subtitling, editing, accessibility engineering, and increasingly AI-mediated communication—depend fundamentally on accurate segmentation, rhythm reconstruction, and discourse interpretation. In this domain, punctuation is not merely a stylistic ornament but a structural signal with direct computational and service-level implications.

First, in machine translation (MT) and large-language-model (LLM)-based translation systems, punctuation is essential for sentence boundary detection, clause segmentation, and discourse coherence[10]. Errors in punctuation recognition lead to cascading translation errors, particularly in Chinese, where boundaries are less overtly marked[8, 47]. Quantitative models of punctuation—such as its Zipf–Mandelbrot frequency behaviour and Weibull spacing distributions—provide statistically grounded priors for improved segmentation algorithms.

Second, in speech-to-text (automatic speech recognition—ASR) and prosody restoration, punctuation determines pause structure, intonational contours, and phrase breaks[48]. Because punctuation frequency and spacing exhibit predictable statistical regularities across genres and languages, incorporating these regularities into ASR post-processing might enhance fluency, readability, and alignment with human expectations.

Third, in subtitling and captioning, punctuation is one of the determinants of readability and cognitive load[49]. The discrete Weibull structure of inter-punctuation distances of-

fers a basis for modeling optimal subtitle chunk lengths and for distinguishing natural breakpoints in Chinese text—an important challenge in languages lacking explicit word boundaries.

Fourth, network-based language services, such as text simplification and author-style adaptation, can also benefit from the proper inclusion of punctuation [50]. Recent evidence indicates that when punctuation tokens are included, Chinese and English word-adjacency networks become more compactified in the average shortest path length sense, whereas excluding punctuation produces paths that are more inflated in Chinese than in English [51].

Fifth, in accessibility services, including screen-reader optimization and plain-language rewriting, punctuation signals cognitive segmentation and listening rhythm [52].

Finally, quantitative modeling of punctuation participates directly in the broader trend of data-driven quality assurance in language services [53]. By treating punctuation as a feature tractable with statistics and network theory, service providers can potentially implement objective metrics that flag anomalous punctuation patterns or stylistic drift across translated or localized content.

In summary, far from being a peripheral or auxiliary component of writing, punctuation constitutes a linguistic signal whose behaviour underpins technological, communicative, and accessibility functions in modern language services. Its modeling—particularly in the structurally distinctive environment of Chinese—has the potential of supporting the development of more reliable and linguistically informed language service technologies.

3. Network Representations

3.1. Corpus Design

To illustrate punctuation behaviour, a corpus of modern Chinese prose was compiled, totaling approximately 3 million characters. It included three representative novels: Gao Xingjian's *Soul Mountain* [54, 55], Din Ling's *The Sun Shines over the Sanggan River* [56, 57], and Liu Yichang's *The Drunkard* [58, 59], and their published English translations. The reasons why the corpus of Chinese texts was created from these three novels are as follows: (1) each of these novels was written by a recognized author (including a Nobel Prize laureate), (2) the length of the texts allows for reliable statistics of words and punctuation marks, (3) published English translations are available for all of them, (4) they were written at a time when the Chinese punctuation had already reached maturity, and (5) they differ in terms of stylistic properties (including fractal characteristics). It should be stressed, however, that in no case was the selection criterion based on their typicality for Chinese literature (which seems natural, given that each of these books received above-average critical acclaim and therefore could not be considered typical). Nevertheless, from the perspective of the goal of the present review stated above, the corpus created from the selected texts does not differ significantly in its lexical and punctuation properties from other Chinese corpora studied by the same and other methods [12, 51].

The workflow consists of four sequential stages. Data curation converts raw text files into enriched, standardized inputs suitable for analysis. These are processed by the Data Engine, which uses a language-specific description set to tokenize the texts, producing tokenized text files. The system then branches into two analytical modules: the Token Processor, which extracts statistical and distributional properties, and the Network Processor, which constructs graph-based linguistic representations. Their outputs—Experiment-Data Files and Network-Data Files—serve as intermediate files for the Visual Processor, which generates the final graphical outputs used for quantitative and structural analysis of the texts. The entire related procedure is schematically depicted in **Figure 1**.

Custom Python scripts using the Jieba segmentation library [60, 61] extracted punctuation symbols and neighboring words. Each punctuation mark was assigned an index, and its distance to the next punctuation mark was measured in characters. Non-lexical symbols such as decorative brackets were removed for uniformity.

3.2. Network Representation of *Soul Mountain*: Chinese vs. English

Figures 2 and **3** present parallel network visualizations of *Soul Mountain* in its original Chinese version and in the English translation. Both networks were constructed using the same token-adjacency method: every word and every punctuation mark is treated as a node, and an edge connects tokens that appear consecutively in the running text.

Although the narrative content is parallel, the two figures reveal clear structural differences produced by the linguistic systems themselves.

The Chinese network displays a dense, highly interconnected structure, dominated by a small number of extremely frequent hubs. These hubs include not only common function words but also punctuation marks, which in Chinese assume a noticeably central structural role. Several of the top-degree nodes are highlighted in the figure to emphasize their connectivity.

A green polyline overlays the network, tracing the sequential reading order through the first 1,000 tokens. The path repeatedly loops through the same punctuation and function-word hubs, illustrating how Chinese prose relies on punctuation to manage clause boundaries and narrative rhythm. Because Chinese lacks inflectional marking and uses fewer explicit syntactic connectors, punctuation marks—especially the comma (，) and ideographic full stop (。)—function as major organizational nodes that link semantically distinct parts of the text.

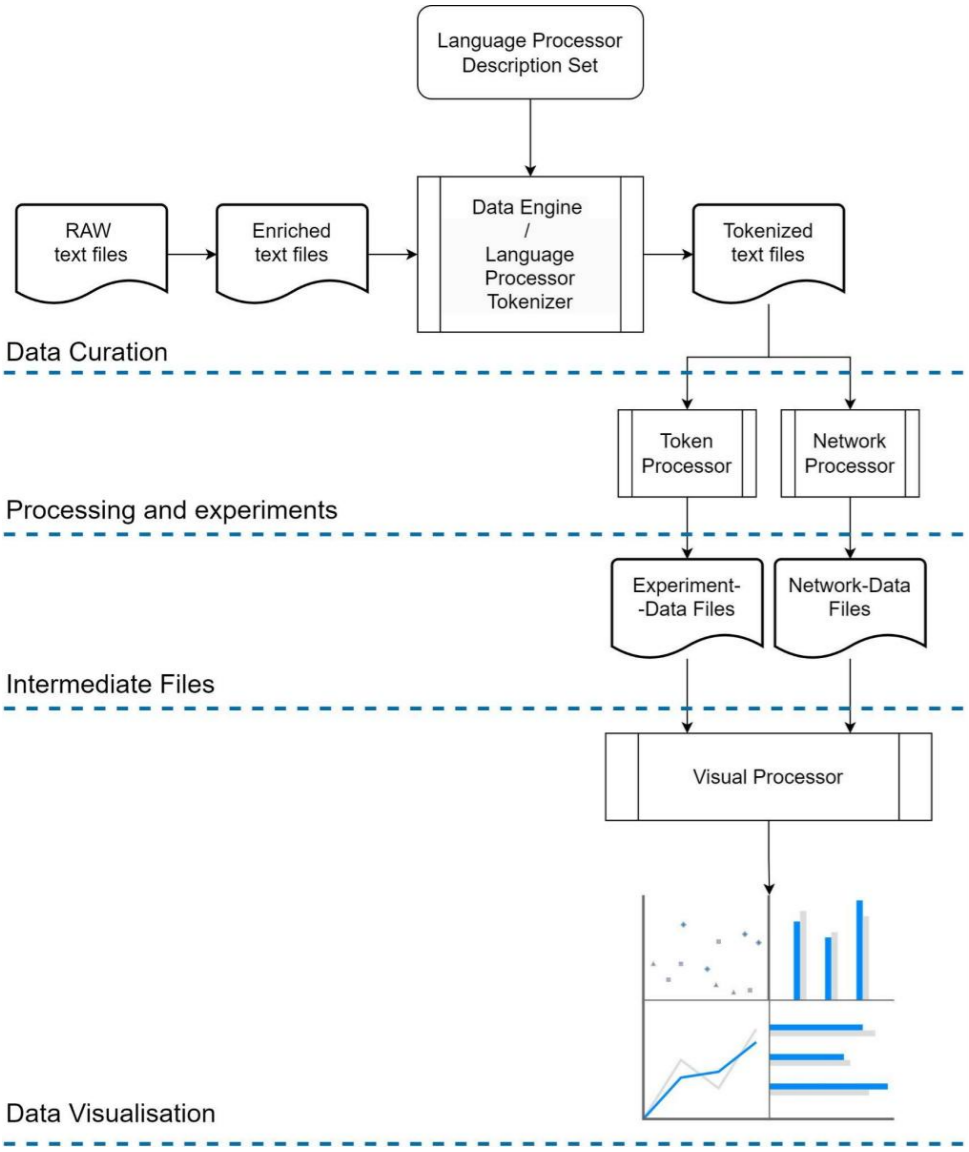


Figure 1. Overview of the processing pipeline: from raw file acquisition, through cleaning and tokenization, to experimentation with both token-based and network-based analyses, culminating in visualization.

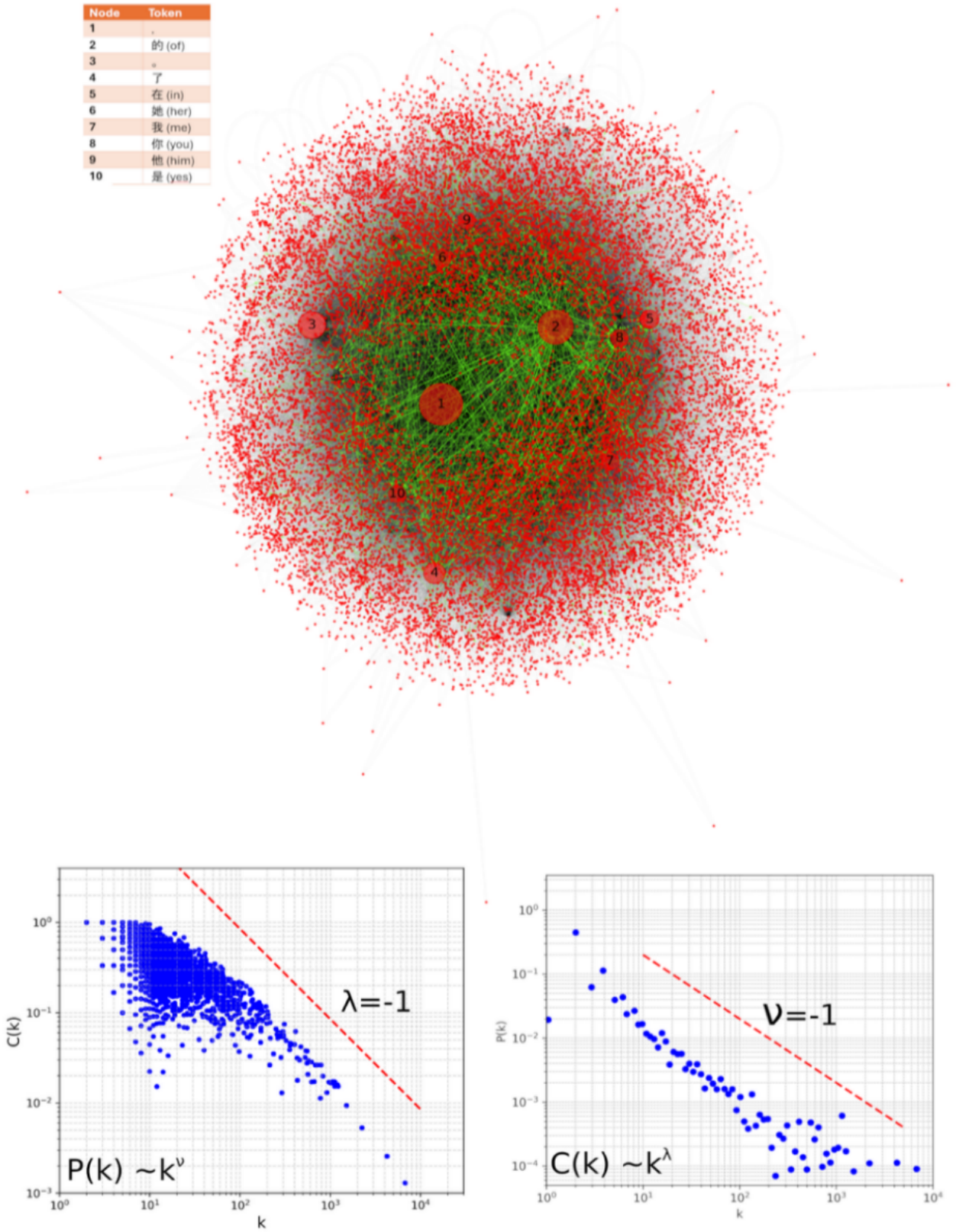


Figure 2. *Soul Mountain* by Gao Xingjian: (Top) A network representation of the Chinese text; a green subnetwork represents the first 1,000 tokens as they occur in the text. (Bottom) Local clustering coefficient C (left) and node degree distribution P (right) as functions of node degree k (Bottom). Power-law dependencies with specific exponents are also shown as visual guides.

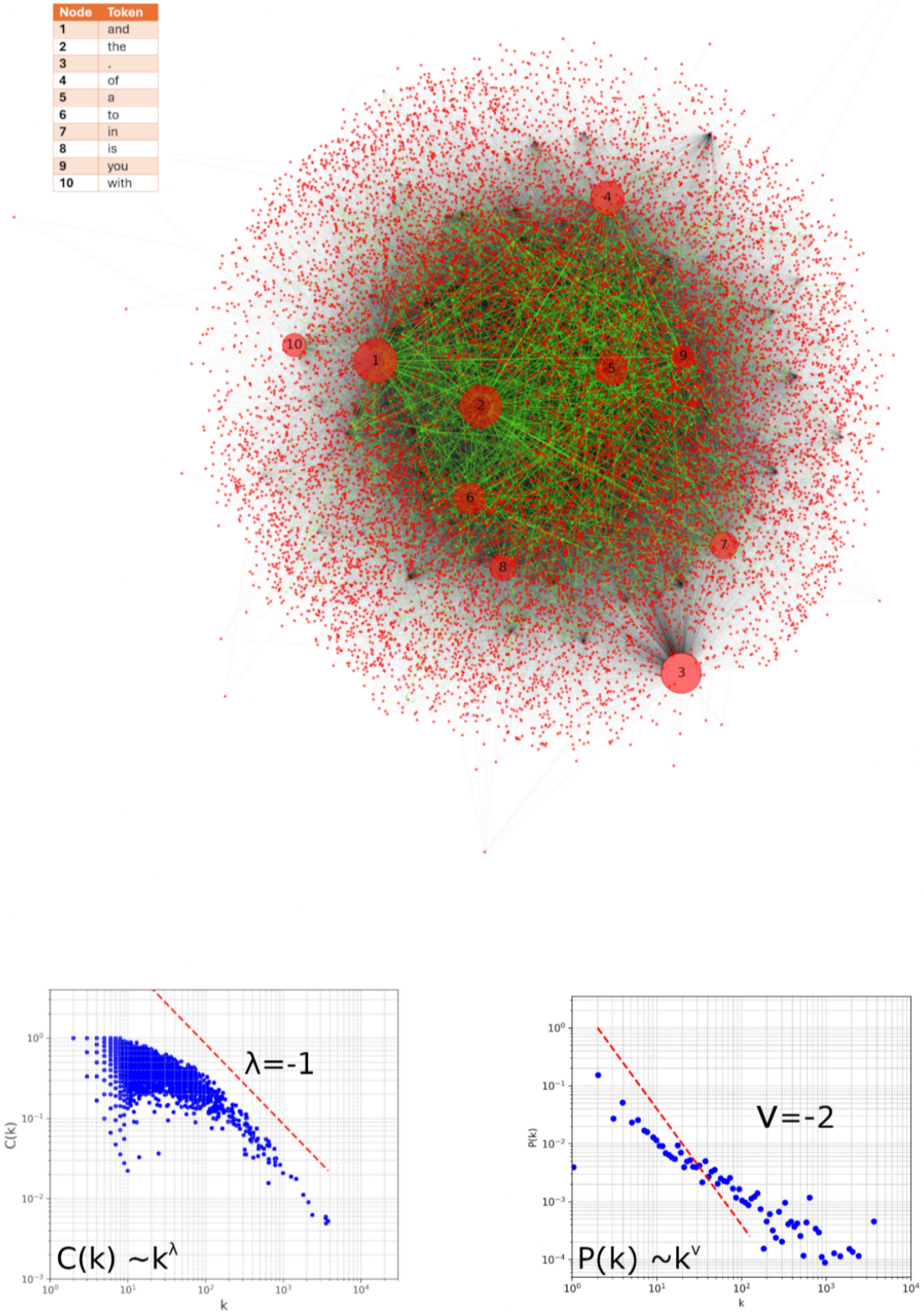


Figure 3. *Soul Mountain* (English translation): The same quantities as in **Figure 2**.

The two inset plots quantify these structural properties.

1. The local clustering-coefficient distribution: $C(i) = \frac{2e_i}{k_i(k_i-1)'}$; the ratio between the number k of links actually present among a node's neighbors and the total number of possible links between those neighbors, shows that many lower-degree nodes participate in tightly bound local groupings, characteristic of repeated multi-character expressions and formulaic phrase structures.
2. The node degree distribution (shown as a cumulative distribution) follows a heavy-tailed form, indicating the presence of hubs with substantially greater connectivity than the average node.

Together, these patterns show that the Chinese network has strong local cohesion and a small-world character. Earlier published measurements[6] confirm this impression: including punctuation reduces the average shortest-path length from 3.84 to 3.25, and increases clustering from 0.19 to 0.27, meaning punctuation strengthens both global and local connectivity.

The English network uses the same construction procedure but organizes itself differently due to the syntactic and stylistic conventions of English. While hubs are still present, they are more evenly distributed across high-frequency lexical items, and punctuation marks play a less dominant role than in the Chinese network.

Several structural differences are visible: (i) English employs punctuation more sparingly and relies heavily on function words (such as and, the, to, of) to encode syntactic relations. As a result, punctuation nodes in **Figure 3** appear less central than in **Figure 2**, producing a more lexically centered hub structure. (ii) Because punctuation appears at longer intervals in English, the reading-order polyline in **Figure 3** (analogous to that in **Figure 2**) spreads over a wider set of lexical nodes before returning to punctuation hubs. This produces a network with slightly longer paths through lexical clusters and fewer punctuation-anchored shortcuts. (iii) In contrast with the dense pockets of closely connected tokens in the Chinese network, English displays a more moderate clustering pattern. English multi-word expressions are often less rigidly formulaic than Chinese multi-character sequences; thus, local neighborhoods tend to be less compact. (iv) The node degree distribution in **Figure 3** still exhibits a heavy tail, but the leading hubs are more evenly shared between lexical words and punctuation marks. This reflects English's reliance on function words as syntactic glue, rather than punctuation.

Viewed together, **Figures 2** and **3** highlight how linguistic structure shapes network topology even when the underlying narrative is the same. In Chinese, punctuation is highly central and acts as a major connector, clusters around idiomatic or formulaic sequences are dense, and hubs produce strong shortcuts, creating a compact small-world structure. In English, translation of *Soul Mountain*, lexical function words, not punctuation, become the dominant hubs. Clustering is more diffuse, with fewer dense subgraphs and paths through the network are longer, reflecting the reduced structural role of punctuation. In sum, the Chinese network is more punctuation-anchored, while the English network is more lexically anchored. Similar observations can be made for other Chinese and English versions of the same texts[6, 51]. These differences validate the methodological decision to treat punctuation as a full linguistic token: its structural impact varies by language and is essential for accurate modeling of textual organization.

3.3. Network Topology

Including punctuation as nodes significantly alters network metrics as compared with their analogues without punctuation. For the Chinese version of *Soul Mountain*, the network contains $N = 20,980$ nodes and $E = 87,929$ edges. Average shortest-path length dropped from $L = 3.84$ (words-only) to $L = 3.25$ (words and punctuation). Clustering coefficient increased from $C = 0.19$ to $C = 0.27$. These shifts indicate enhanced connectivity and local cohesion, confirming that punctuation operates as an organizing principle rather than a boundary delimiter. While both the Chinese and English networks were scale-free with degree exponent $v < 2$, differences arise from punctuation convention: commas in Chinese often replace conjunctions, generating hub-like nodes with higher degree centrality that connect semantically distinct phrases.

4. Zipf's Statistics for Chinese and English Texts

In order to quantify how punctuation contributes to the statistical organization of Chinese prose and its English

translation, this section presents a focused Zipfian analysis of two datasets: (a) *Soul Mountain* in the original Chinese and in its English translation, and (b) a mini-corpus consisting of three Chinese novels (*The Drunkard*, *The Sun Shines over the Sanggan River*, and *Soul Mountain*) together with their English translations. All analyses follow the methodological framework established by Dolina et al.[6] in their study of Zipf's law and Weibull spacing for modern Chinese punctuation.

Figures 4 and **5** display the rank–frequency distribution of tokens in two cases: when punctuation marks are treated as full linguistic units and when they are excluded from the analysis. Punctuation is shown to improve Zipfian (power-law) behaviour by reducing deviations at high ranks. However, due to the fact that the slope behaviour is decided by the high-rank tokens where punctuation marks are mingled with the words, there is no sizable difference between the two cases.

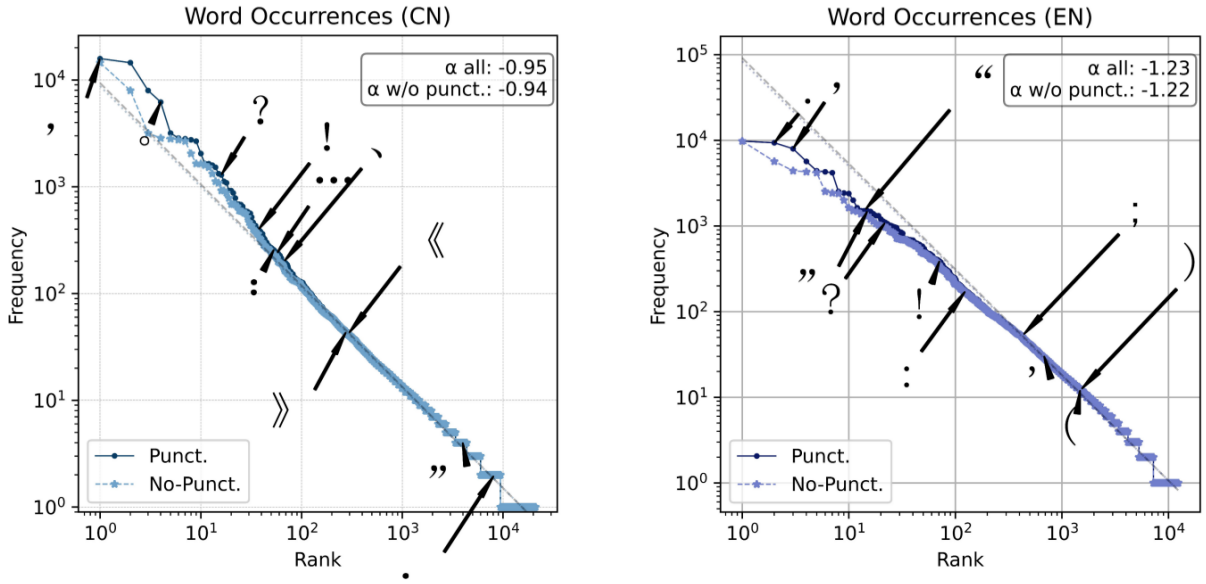


Figure 4. Zipf plots for the Chinese original *Soul Mountain* (left panel) and its English translation (right panel) including and excluding punctuation marks.

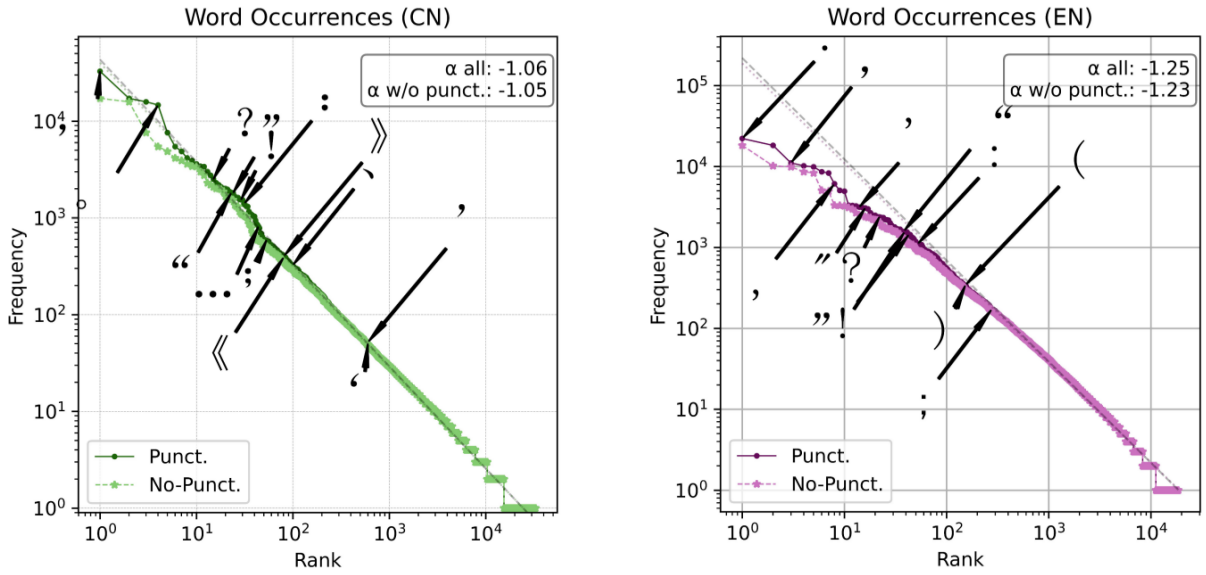


Figure 5. Zipf plots for the mini-corpora composed of three Chinese books (left) and their English translations (right).

The comma (，) and ideographic full stop (。) dominate the high-frequency regime and collectively account for more than half of all punctuation occurrences, reflecting their syntactic centrality in Chinese narrative structure—especially the preference for comma-chained clauses (**Figure 4**).

The slope index α for the Chinese text of *Soul Mountain* is lower (less negative) than its counterpart for the English texts, implying a flatter hierarchy of token frequencies. Overall, the results obtained for this book confirm the universality of the Zipf–Mandelbrot description while demonstrating language-specific adjustments in scaling exponent.

For the mini-corpora shown in **Figure 5**, the Zipf distributions exhibit: (i) robust power-law scaling, with $\alpha \approx 1.06$ when punctuation is included; (ii) a clear improvement in Zipfian fit for the all-token distributions relative to word-only distributions; (iii) a heavier usage of intermediate-frequency punctuation, especially semicolons and ellipses, visible as small inflection points in the mid-rank region.

A key methodological insight is that the relative difference between Chinese and English exponents remains visible across all texts. This indicates that the Zipf exponent is primarily shaped by structural linguistic factors, not narrative content or authorial style. The figures thus provide quantitative support for the claim that punctuation forms an integral component of the statistical "signature" of Chinese prose. Through the lens of Zipf's original interpretation of the scale-free structure of word frequencies as a manifestation of least-effort principle[2], which can in turn be viewed as a result of language self-organization, the power-law form of token frequency distribution combining words and punctuation marks may be considered as an emergent phenomenon related to the complexity of natural language.

5. Inter-Mark Distances in Chinese and English Mini-Corpora

Another statistical characteristic that can be related to the complexity of texts is the distance l between two successive punctuation marks. It is defined as the number of tokens (characters or words) occurring between them. Punctuation placement is modeled as a stochastic waiting-time process whose probability mass function follows a discrete Weibull form given by Equation (2). The shape parameter β controls the tail behaviour of the distribution: values $\beta < 1$ correspond to decreasing hazard and heavy-tailed spacing, $\beta \approx 1$ yields near-exponential decay, and $\beta > 1$ indicates increasingly frequent segmentation.

5.1. Analysis of Distances between Punctuation Marks

The four panels in **Figure 6** show the results of the analysis of distances between punctuation marks, separately for (i) all punctuation marks and for (ii) sentence-ending punctuation only, for both Chinese and English corpora (**Figure 7**).

5.2. Results for Different Corpora and Distance Types

(a) Inter-punctuation distances (all punctuation marks). The top-left panel (CN, all punctuation) shows a distribution strongly dominated by short distances. The fitted Weibull exponents assume values $1.5 < \beta < 1.8$ indicating super-exponential decay with relatively limited tail thickness. This behaviour reflects the dense use of internal punctuation—especially commas and enumeration marks—in Chinese prose, which promotes frequent segmentation and suppresses long punctuation-free stretches. Short inter-punctuation distances are therefore highly probable, while very long gaps occur only rarely, resulting in comparatively compact spacing distributions (**Figure 6**, top left). Consistently, the hazard function increases steeply with l (**Figure 7**, top left).

(b) Sentence-ending punctuation distances. The top-right panel (CN, sentence-ending punctuation) exhibits broader distributions than those for all punctuation marks, but the tails remain moderate (**Figure 6**, top right). Weibull shape parameters are typically $1.0 < \beta < 1.5$, indicating controlled variability in sentence length. Interestingly, the hazard function $h(l)$ is slowly increasing as compared with the case discussed above (**Figure 7**, top right).

(c) Inter-punctuation distances (all punctuation marks). The lower-left panel (EN, all punctuation) displays a substantially heavier-tailed distribution than its Chinese counterpart (**Figure 6**, bottom left). The Weibull fits yield shape parameters $1.2 < \beta < 1.5$, corresponding to a more slowly increasing hazard (**Figure 7**, bottom left) and elevated probability of long inter-punctuation gaps. This indicates that

English permits extended stretches of text without punctuation, relying more strongly on syntactic function words rather than frequent punctuation for local structuring. Although Chinese narrative prose allows stylistic flexibility, sentence boundaries remain relatively well-regulated, producing fewer extreme sentence lengths than observed in English. As a consequence, English curves lie consistently above the Chinese ones in the tail region of the distribution. It is worthwhile to note that, in general, the case of $\beta < 1.0$ is rare yet observed in some remarkable pieces of English literature like *Finnegans Wake* and *Ulysses* by James Joyce. In that case, the hazard function decreases with consecutive words and, unusually, the probability of encountering a punctuation mark decreases with distance l [62].

(d) Sentence-ending punctuation distances. The lower-right panel of **Figure 6** (EN, sentence-ending punctuation) reveals the least evident cross-language difference. The fitted Weibull exponents cluster near $\beta \approx 1.4$, producing moderate tails for words and heavier tails for characters. Long English sentences—often exceeding 60–80 characters or 30–40 words—occur with markedly higher probability than in the Chinese corpora. This indicates that English prose tolerates, and frequently employs, extended sentence constructions with substantial internal complexity (see also, e.g., Sichel, Williams, Xiao and Yule [63–66]). Interestingly, unlike Chinese, here the hazard function increases more strongly than in the case of all punctuation marks considered (**Figure 7**, bottom right).

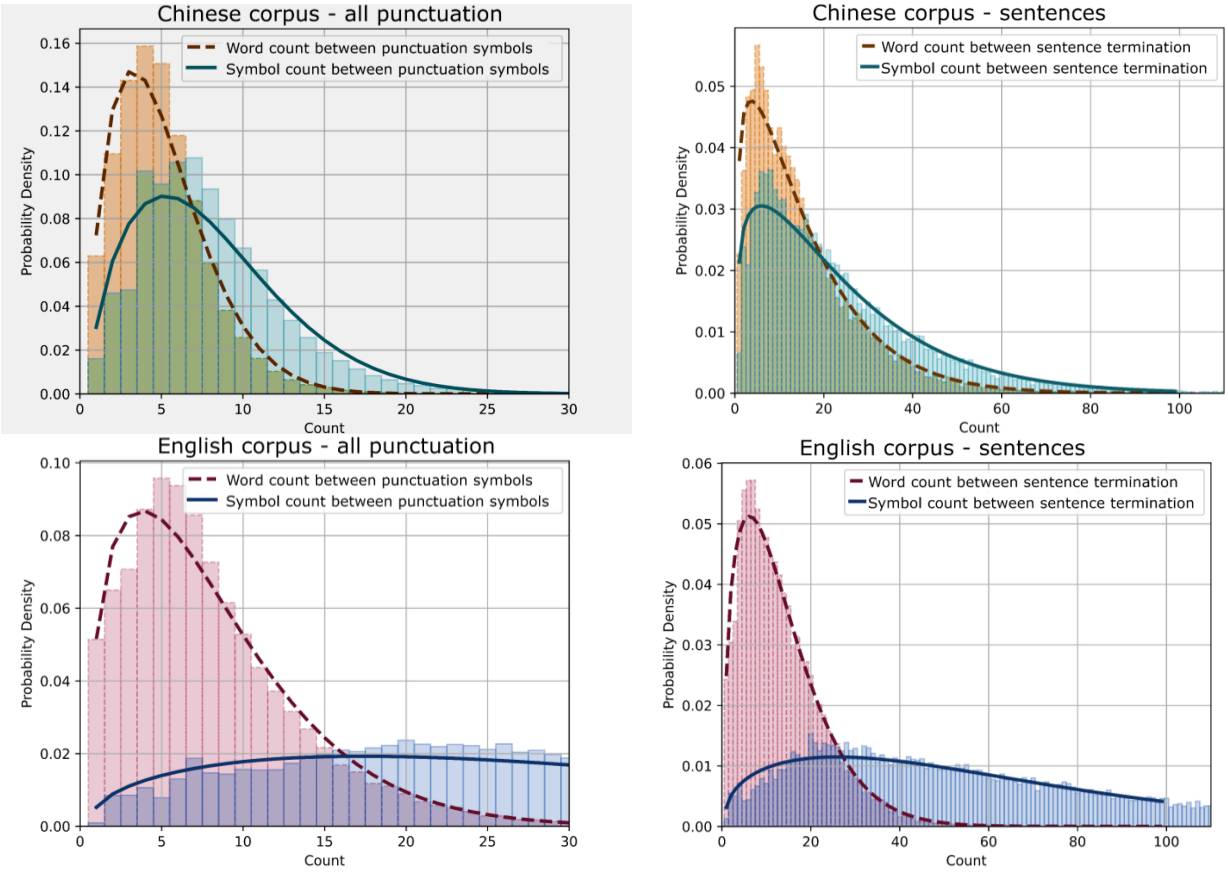


Figure 6. Distributions (histograms) of intervals between punctuation marks for the Chinese micro-corpora (top) and their English translations (bottom).

Note: The left-side panels correspond to taking into account all the punctuation marks and measuring the corresponding distances in terms of the number of words (dashed fit) and, independently, in terms of the number of characters (solid fit). The right-side panels correspond to the same characteristics for the distances between the sentence-ending punctuation marks only. The resulting distributions are presented in terms of the best-fitted discrete Weibull distributions with p and β as free parameters. For the Chinese micro-corpus: $p = 0.07$, $\beta = 1.57$ (all marks, words), $p = 0.03$, $\beta = 1.61$ (all marks, characters), $p = 0.04$, $\beta = 1.17$ (full stops, words), and $p = 0.02$, $\beta = 1.19$ (full stops, characters); for the English micro-corpus: $p = 0.05$, $\beta = 1.40$ (all marks, words), $p = 0.01$, $\beta = 1.42$ (all marks, characters), $p = 0.03$, $\beta = 1.37$ (full stops, words), and $p = 0.01$, $\beta = 1.17$ (full stops, characters).

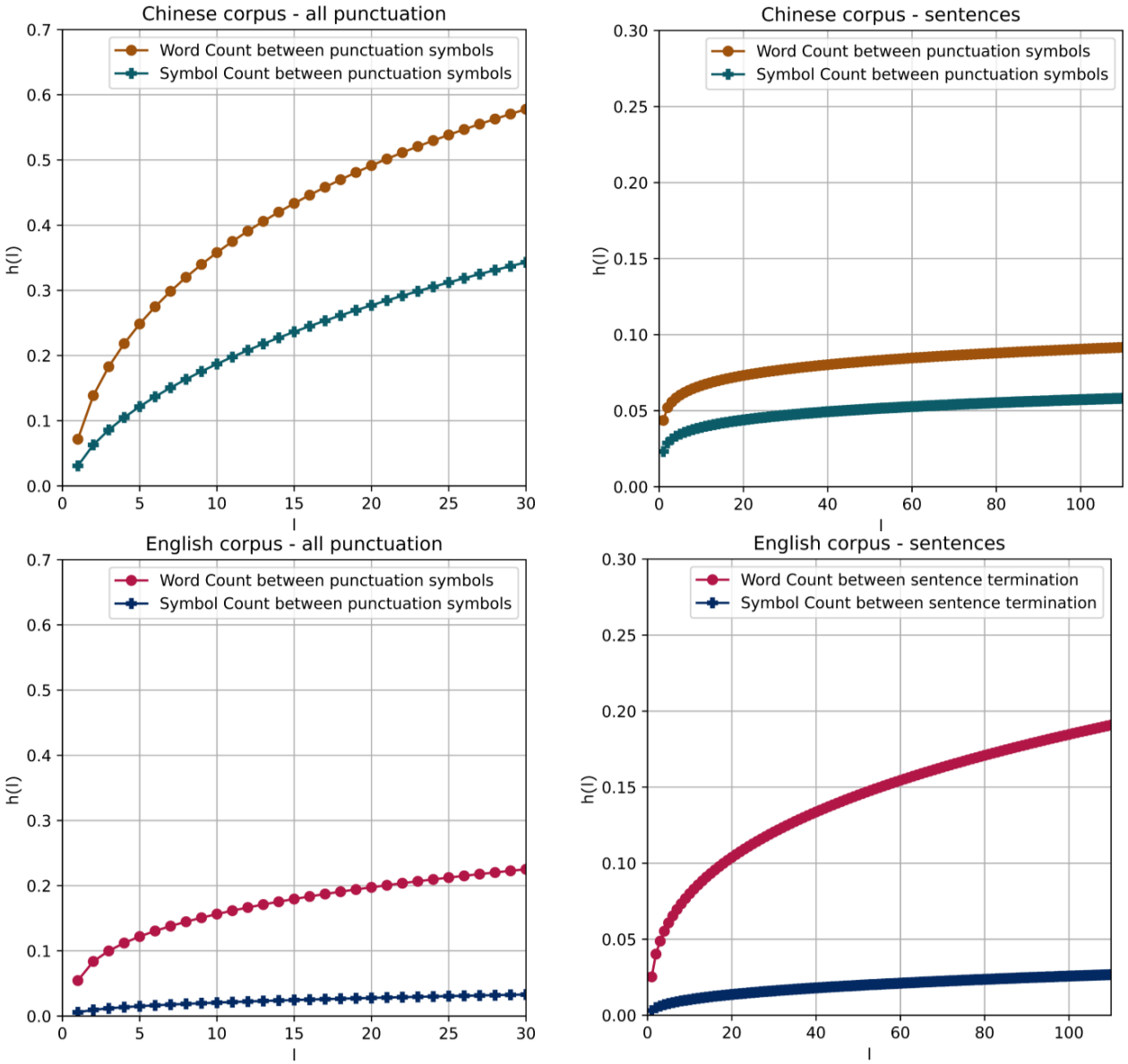


Figure 7. Hazard functions $h(l)$ corresponding to the inter-punctuation intervals for the Chinese micro-corpora (top) and their English translations (bottom).

Note: The left-side panels consider all punctuation marks and quantify the intervals both in terms of word counts and, independently, character counts between consecutive punctuation marks. The right-side panels restrict the analysis to sentence-ending punctuation only, capturing the corresponding distances measured in words and characters. The curves represent the hazard functions derived from the estimated parameters of the discrete Weibull distributions.

5.3. Cross-Linguistic Interpretation

Taken together, the four Weibull plots demonstrate that: (i) English exhibits heavier tails for both all-punctuation and sentence-ending distributions, reflecting greater variability in clause and sentence length. (ii) Chinese displays near-exponential spacing, indicating tighter punctuation-driven segmentation. (iii) Sentence-ending punctuation amplifies these contrasts most strongly, with English showing a pronounced propensity for very long sentences. (iv) Frequent punctuation in Chinese produces compact spacing distributions, whereas sparser punctuation in English permits broader temporal dispersion. These results identify punctuation spacing as a quantitative marker of language-specific syntactic organization.

6. Multifractal Characteristics of Punctuation Patterns

While the Zipf-Mandelbrot structures may be related to the phenomenon of complexity, if a certain interpreta-

tion mentioned in Section 4 is assumed, a multifractal organization is a property inherently and inextricably built into a wide class of complex systems. Multifractal structure of fluctuations of measured quantities related to a complex system, revealed through the Multifractal Detrended Fluctuation Analysis (MFDFA) framework[67, 68], captures how these fluctuations of different amplitudes behave across scales and how the narrative rhythm oscillates between concentrated bursts of short clauses and extended descriptive passages[14, 69, 70].

6.1. Multifractal Detrended Fluctuation Analysis

For a series u_j the main steps of the MFDFA algorithm comprise[67]:

1. Construction of an integrated signal: $x_i = \sum_{j=1}^{i} u_j$.
2. Detrending of this signal by subtracting a properly adjusted polynomial trend $p^{(m)}(i)$ in a window of size s.
3. Calculating the variance of the residual signal as:

$$f^2(v,s) = \frac{1}{s}\sum_{i=1}^{s}(x_i - p^{(m)}(i))^2. \quad (5)$$

4. Calculating a family of fluctuation functions $F_q(s)$ of order q by using the average variance computed across all M_s non-overlapping windows:

$$F_q(s) = \{\frac{1}{2M_s}\sum_{v=0}^{2M_s-1}[f^2(v,s)]^{q/2}\}^{1/q}. \quad (6)$$

5. Repeating the points 2–4 for different values of the scale s and the Rényi index q.
6. If $F_q(s)$ functions obey power-law of type

$$F_q(s) \sim s^{H(q)}, \quad (7)$$

a time series under study is fractal. If, additionally, $H(q)$ is q-dependent, it indicates that the time series is multifractal or monofractal otherwise. The underlying scaling is typically studied on the double logarithmic scale, where—when it holds—results in straight lines[71]. The functions $H(q)$ are considered the generalized Hurst exponents. In particular, $H(2)$ is the conventional Hurst exponent[44, 72, 73].

Multifractality can thus be quantified directly in terms of $H(q)$-dependence, but another commonly adopted measure is the singularity spectrum[74, 75], which can be derived from $H(q)$ by the Legendre transform:

$$\begin{aligned} f(\alpha) &= q[\alpha - H(q)] + 1, \\ \alpha &= H(q) + qH'(q). \end{aligned} \quad (8)$$

6.2. Multifractality of Sentence-Length Dynamics in *Soul Mountain*

Punctuation-mediated segmentation in *Soul Mountain* displays a markedly multifractal organization in the temporal evolution of sentence lengths and inter-punctuation distances[51, 76]. However, because punctuation in Chinese and English often performs different tasks, the Chinese and English versions of *Soul Mountain* differ systematically in their fluctuation exponents, singularity spectra, and long-range correlation patterns (**Figures 8** and **9**). To examine these differences rigorously, two complementary time-series representations were evaluated. In this part, we deal with time series corresponding to the number of characters (Chinese) or words (English) between consecutive sentence-ending punctuation marks. This representation captures sentence-level temporal organization and long-range narrative pacing.

The Chinese original exhibits clear multifractal scaling ($\Delta\alpha \approx 0.5$–0.55) with the fluctuation functions obeying power-law behaviour over a significant range of scales $s <$ 1,000. The generalized Hurst exponent $H(q)$ decreases with increasing q, indicating different scaling between small and large fluctuations. The English translation displays roughly similar strength of multifractality at both the sentence and punctuation-mark level. All singularity spectra display a left-sided asymmetry ($A\alpha > 0$) indicating that these are the large fluctuations that are characterized by a richer multifractal property[77].

6.3. Multifractality of Distances between All Punctuation Marks in *Soul Mountain*

Here, the time series of distances between any two consecutive punctuation marks: periods, commas, semicolons, quotation marks, parentheses, and dashes are considered. In the Chinese text, the high density of commas and enumeration marks produces a relatively regular alternation of short

punctuation-delimited segments. The resulting time series exhibits multifractal behaviour but with limited intermittency. The singularity spectrum width typically lies near $\Delta\alpha \approx 0.5$ and the asymmetry of $f(\alpha)$ is larger than in the sentence-ending case. This pattern indicates that, while the Chinese punctuation supports stylistic modulation, frequent segmentation suppresses extreme fluctuations at the micro-structural level.

The English translation shows somewhat richer multifractality in this case. Reduced punctuation density allows longer intra-sentence spans, increasing the contrast between short and long fragments. The MFDFA results reveal a broader singularity spectrum with $\Delta\alpha \approx 0.58$ and pronounced asymmetry associated with extremely large-scale fluctuations. These features indicate a highly heterogeneous segmentation structure, consistent with the English-language reliance on structural-segmentation mechanisms other than frequent punctuation.

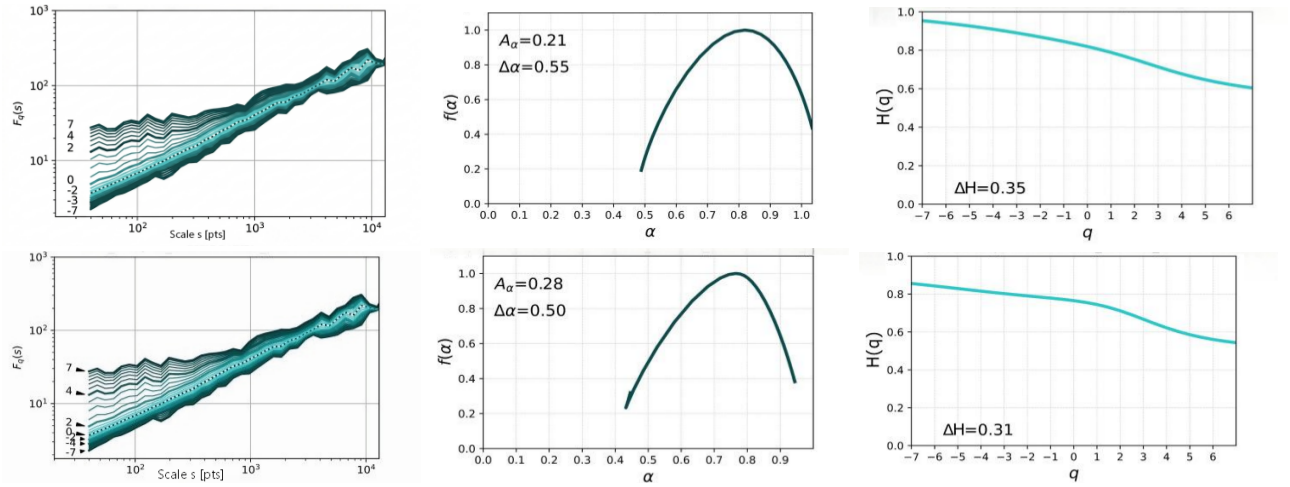


Figure 8. Results of MFDFA for *Soul Mountain* in its Chinese original version for the distances between the sentence ending punctuation marks (top) and all the punctuation marks (bottom). Fluctuation functions (left, with values of the index -7 $\leq q \leq$ 7 denoted), singularity spectra (middle, with the asymmetry parameter).

Note: $A\alpha$ and the spectrum width $\Delta\alpha$, and the generalized Hurst exponents (right, with their variability ΔH) calculated from the approximate scaling regions of s are shown.

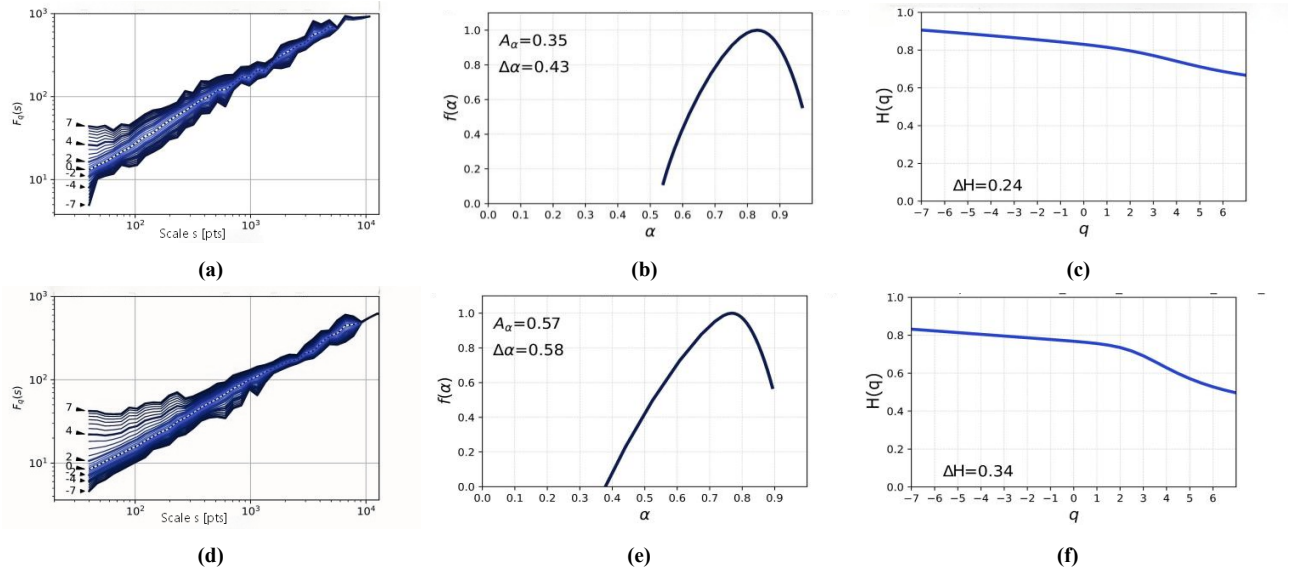


Figure 9. The same quantities as in **Figure 8** but here for the English translation of *Soul Mountain*.

7. Applications to Language Services

The quantitative characterization of punctuation presented in this study may have implications for contemporary language services. Punctuation emerges as a structurally meaningful signal rather than a secondary orthographic feature. The Zipfian, Weibull, network-based, and multifractal results collectively demonstrate that punctuation in Chinese encodes regularities that have the potential of being operationalized in practical service-oriented technologies.

7.1. Machine Translation and Large Language Models

Machine translation (MT), especially neural and LLM-based systems, depends critically on accurate segmentation [78, 79]. In Chinese, where explicit word boundaries are absent and syntactic relations are often underspecified lexically, punctuation serves as a primary indicator of clause and sentence structure [80]. The Zipf–Mandelbrot scaling observed for punctuation frequencies implies that these marks follow stable hierarchical patterns comparable to those of high-frequency lexical tokens. Incorporating punctuation-aware priors into tokenization and alignment stages can therefore stabilize translation outputs, particularly for long, comma-chained sentences that are characteristic of modern Chinese prose. The Weibull-distributed spacing further suggests that MT systems should expect tighter and more regular punctuation-driven segmentation in Chinese than in English. Deviations from the expected spacing distributions, such as excessively long punctuation-free spans in translated Chinese output, can be interpreted as indicators of translation distortion, over-literal sentence fusion, or loss of a rhythm. Conversely, overly frequent sentence termination in English translations may signal excessive fragmentation. Thus, Weibull parameters provide a quantitative diagnostic tool for translation quality estimation and post-editing support.

7.2. Automatic Speech Recognition and Prosody Restoration

In ASR and text-to-speech systems, punctuation plays a decisive role in reconstructing pauses, intonation contours, and phrase boundaries [81]. The multifractal analysis of punctuation-mediated sentence dynamics can be exploited to improve prosody restoration by aligning pause durations and intonational breaks with statistically expected punctuation patterns rather than relying solely on acoustic cues. For ASR post-processing, punctuation restoration remains a major challenge, especially for Mandarin Chinese. The results presented here indicate that punctuation placement is not random but governed by universal yet language-specific statistical laws. Integrating Zipfian rank expectations and Weibull-distributed spacing constraints into punctuation prediction modules can improve both readability and naturalness, leading to outputs that better match human-written standards.

7.3. Subtitling and Captioning

Subtitling and captioning systems must balance temporal constraints, readability, and cognitive load. In Chinese, frequent punctuation enables fine-grained segmentation, which is advantageous for subtitle chunking but can also lead to over-segmentation if not properly calibrated. The near-exponential Weibull spacing observed for Chinese punctuation provides a principled basis for defining optimal subtitle unit lengths, ensuring that breaks align with natural linguistic boundaries rather than arbitrary character limits.

7.4. Network-Based Language Services and Style Control

The network analysis demonstrates that punctuation significantly reduces average shortest-path lengths and increases clustering in linguistic graphs, particularly in Chinese. From a service perspective, this finding supports the inclusion of punctuation tokens in semantic and stylistic modeling. Applications such as text simplification, style transfer, and authorial adaptation can redefine punctuation-centered network hubs as anchors for restructuring content without altering semantic core meaning. In multilingual and localization contexts, punctuation-aware networks also facilitate cross-linguistic alignment. While Chinese networks are punctuation-anchored more than the English ones, recognizing this asymmetry allows service systems to adjust similarity metrics and avoid penalizing structurally legitimate differences. This is especially relevant for translation memory systems and automated consistency checks across

language versions.

7.5. Quality Assurance and Data-Driven Evaluation

Finally, the quantitative regularities identified in this study enable objective quality assurance metrics for language services. Anomalies in Zipf slopes, Weibull tail behaviour, or multifractal spectra can be used to detect stylistic drift, machine-generated artifacts, or inconsistencies across large document collections. Treating punctuation as structured data rather than superficial markup thus contributes to the broader trend toward evidence-based evaluation in the language service industry.

8. Conclusions

The principal aim of this review is to synthesize a few recent findings regarding the fundamental importance of punctuation in written texts (both the modern Chinese and the Western ones) from the perspective of statistical properties of natural language, and to sketch their potential relevance to the field of language service studies. More specifically, these findings demonstrate that punctuation in modern Chinese prose constitutes a structurally rich and quantitatively tractable component of language, rather than a merely auxiliary orthographic device. By applying a unified framework of quantitative linguistics and complex systems analysis—encompassing the Zipf–Mandelbrot statistics, the discrete Weibull distribution, network topology, and multifractal analysis—it was shown that punctuation encodes stable regularities that are deeply intertwined with syntactic organization, narrative rhythm, and message cohesion.

Chinese punctuation was found to obey power-law rank-frequency distributions comparable to those of lexical tokens, with the inclusion of punctuation consistently improving Zipfian scaling behaviour for small ranks. This result reinforces the view that punctuation must be treated as a full-fledged linguistic unit in statistical modeling. The systematic difference between Chinese and English in their Zipf exponents further indicates that punctuation reflects typological distinctions in clause construction and information packaging, rather than idiosyncratic authorial style. These findings align with and extend earlier work by situating punctuation at the core of cross-linguistic statistical signatures.

The analysis of inter-punctuation distances revealed that punctuation placement in Chinese follows super-exponential Weibull distributions with thin tails, reflecting frequent and relatively regular segmentation, while English exhibits heavier-tailed behaviour indicative of longer punctuation-free spans and greater sentence-length variability. This contrast was shown to be particularly pronounced for complete punctuation, underscoring fundamental differences in syntactic tolerance and narrative pacing between the two languages. Importantly, these spacing regularities provide interpretable parameters that can be applied to segmentation, translation alignment, and readability modeling.

From a structural perspective, word-adjacency network analysis demonstrated that punctuation plays a crucial connective role in Chinese texts. When punctuation tokens are included, linguistic networks exhibit reduced average shortest-path lengths and increased clustering, confirming that punctuation acts here principally as an organizing hub. The comparison with English networks further revealed a slight shift from punctuation-centered to lexically centered connectivity, validating the language-specific functional role of punctuation within network representations of text.

The multifractal analysis added a complementary dynamical dimension to these findings. Punctuation-mediated sentence-length and inter-punctuation time series were shown to exhibit genuine multifractality, with English translations displaying slightly broader singularity spectra than the Chinese originals. These results indicate that punctuation governs not only local segmentation but also long-range narrative dynamics, offering a quantitative handle on rhythm, pacing, and cognitive variability in texts.

Taken together, the results of this study support a central conclusion: punctuation constitutes a computable structural signal that bridges linguistic theory and practical language service applications. By quantifying punctuation behaviour across multiple analytical levels, this work provides empirical foundations for punctuation-aware machine translation, speech technology, subtitling, accessibility services, and data-driven quality assurance. In the context of AI-mediated communication and LLMs, such insights are especially timely, as they point toward lightweight, interpretable features that can enhance fluency, coherence, and cultural adequacy without relying solely on opaque model scaling.

Our analyses discussed in this paper, like any research,

are subject to certain limitations. We have presented results obtained from a relatively small corpus of texts, both in Chinese and in English. Although, in terms of lexicon and punctuation, these texts do not qualitatively differ from other books, either in their original versions or in English translation, as was demonstrated by our previous analyses [13, 51, 73], some quantitative differences may nevertheless be present. Moreover, our analysis does not account for the influence of literary genre, individual features of the author's style, or the target translation language. Therefore, the results presented in this review should be regarded as an indication of possible directions for further analysis and for applying our methods in a completely different field, namely, language services.

Looking forward, several directions for future research emerge. Consistent with what has been stated above, extending the present framework to additional genres, such as news, social media, and spoken transcripts, would clarify the robustness of punctuation statistics across communicative contexts. Incorporating other languages with non-alphabetic or hybrid writing systems would further test the universality and system-specificity of the observed patterns. Finally, integrating quantitative punctuation features directly into operational language service pipelines offers a promising direction for bridging theoretical linguistics, complexity science, and applied language technology. In conclusion, this study stresses that punctuation is an important object of quantitative linguistic inquiry and a valuable resource for modern language services. It contributes to a more subtle and structurally informed understanding of language in the age of intelligent communication technologies. We encourage specialists in these domains to consider our findings and our methodological framework in their own research.

Author Contributions

Conceptualization, J.D., M.D., S.D., J.K, J.L, and T.S.; methodology, S.D. J.K. and T.S.; software, J.D., M.D. T.S.; validation, J.D., M.D., S.D., J.K., J.L. and T.S.; formal analysis, J.D., M.D., S.D., J.K., J.L. and T.S.; investigation, J.D., M.D., S.D., J.K., J.L. and T.S.; resources, J.D., M.D. and J.L.; data curation, J.D. and M.D.; writing, S.D. and J.K.; original draft preparation, S.D. and J.K.; writing—review and editing, J.D., M.D., S.D., J.K., J.L. and T.S.; visualization, J.D. and M.D.; supervision, S.D., J.K. and J.L.; project administration, S.D.; funding acquisition, J.D., M.D., S.D., J.K., J.L. and T.S. All authors have read and agreed to the published version of the manuscript.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflict of interest.

AI Use Statement

During the preparation of this work, ChatGPT was used to create a summary of selected previous publications by the authors; the results and figures presented are the authors' own work. The authors reviewed and edited the content as necessary and take full responsibility for the final content of the published article.

References

[1] Zipf, G.K., 1935. The Psychobiology of Language: An Introduction to Human Ecology. Houghton-Mifflin: Boston, MA, USA.

[2] Zipf, G.K., 1949. Human Behavior and the Principle of Least Effort. Addison-Wesley Press: Cambridge, MA, USA.

[3] Heaps, H.S., 1978. Information Retrieval: Computational and Theoretical Aspects. Academic Press, Inc: Orlando, FL, USA.

[4] Herdan, G., 1960. Type-Token Mathematics: A Text-

book of Mathematical Linguistics. Mouton: The Hague, The Netherlands.

[5] Dec, J., Dolina, M., Drożdż, S., et al., 2024. Multifractal hopscotch in Hopscotch by Julio Cortázar. Entropy. 26(8), 716.

[6] Dolina, M., Dec, J., Drożdż, S., et al., 2025. Quantifying patterns of punctuation in modern Chinese prose. Chaos. 35, 023155.

[7] Kulig, A., Kwapień, J., Stanisz, T., et al., 2017. In narrative texts punctuation marks obey the same statistics as words. Information Sciences. 375, 98–113.

[8] Zong, H., Wang, Y., Tsang, E.C.C., et al., 2025. Improving translation accuracy for long sentences via rule-based and LLM-driven segmentation. International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), Bali, Indonesia, 12–15 July 2025; pp. 7–14.

[9] Xiao, R., Hu, X., 2015. Corpus-Based Studies of Translational Chinese in English-Chinese Translation. Shanghai Jiao Tong University Press: Shanghai, China; Springer-Verlag: Berlin/Heidelberg, Germany.

[10] Li, S., Hu, R., Wang, L., 2025. Efficiently building a domain-specific large language model from scratch: A case study of a classical Chinese large language model. arXiv preprint. arXiv:2505.11810. DOI: https://doi.org/10.48550/arXiv.2505.11810

[11] O'Hagan, M., McDonough Dolmaya, J., 2023. Introduction to Digital Translation: International Journal of Translation and Localization: Positioning translation and localization in a changing digital world. Digital Translation. 10, 1–15.

[12] Stanisz, T., Drożdż, S., Kwapień, J., 2024. Complex systems approach to natural language. Physics Reports. 1053, 1–84.

[13] Kwapień, J., Drożdż, S., 2012. Physical approach to complex systems. Physics Reports. 515(3–4), 115–226.

[14] Drożdż, S., Oświęcimka, P., Kulig, A., et al., 2016. Quantifying origin and character of long-range correlations in narrative texts. Information Sciences. 331, 32–44.

[15] Liu, H., 2009. Statistical properties of Chinese semantic networks. Chinese Science Bulletin. 54, 2781–2785.

[16] Cong, J., Liu, H., 2014. Approaching human language with complex networks. Physics of Life Reviews. 11(4), 598–618.

[17] Feng, J., Wang, P., Gu, C., 2025. Multiscale structural complexity analysis of the Chinese classics A Dream of Red Mansions and All Men Are Brothers. Chinese Physics B. 35(1), 010506.

[18] Gu, Q.-C., Qin, G.-Q., Wang, Y.-Q., et al., 2019. Scale-invariance exists in the series of character intervals in the four great Chinese novels. Communications in Theoretical Physics. 71, 1139.

[19] Yang, Y., Gu, C., Xiao, Q., et al., 2017. Evolution of scaling behaviors embedded in sentence series from A Story of the Stone. PLOS One. 12(2), e0171776.

[20] Liu, Y., Zhuo, X., Zhou, X., 2024. Multifractal analysis of Chinese literary and web novels. Physica A. 641, 129749.

[21] Chen, H., Liu, H., 2018. Quantifying evolution of short and long-range correlations in Chinese narrative texts across 2000 years. Complexity. 2018(1), 9362468.

[22] Yang, T., Gu, C., Yang, H., 2016. Long-range correlations in sentence series from A Story of the Stone. PLOS One. 11(9), e0162423.

[23] Baek, S.K., Bernhardsson, S., Minnhagen, P., 2011. Zipf's law unzipped. New Journal of Physics. 13, 043004.

[24] Yang, L., Liu, J., Liu, Z., et al., 2026. From text to network: A framework for identifying causal factors and risk propagation paths in maritime accidents. Reliability Engineering & System Safety. 271, 112282.

[25] Ferrer i Cancho, R., Solé, R.V., 2001. The small world of human language. Proceedings of the Royal Society B. 268(1482), 2261–2265.

[26] Ferrer i Cancho, R., Solé, R.V., Köhler, R., 2004. Patterns in syntactic dependency networks. Physical Review E. 69, 051915.

[27] Steyvers, M., Tenenbaum, J.B., 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. Cognitive Science. 29(1), 41–78.

[28] Kwapień, J., Drożdż, S., Orczyk, A., 2010. Linguistic complexity: English vs. Polish, text vs. corpus. Acta Physica Polonica A. 117(4), 716–720.

[29] Mandelbrot, B., 1953. An informational theory of the statistical structure of language. In: Jackson, W. (Ed.). Communication Theory. Academic Press: Princeton, NJ, USA. pp. 486–502.

[30] Mandelbrot, B., 1954. Simple games of strategy occurring in communication through natural languages. Transactions of the IRE Professional Group on Information Theory. 3(3), 124–137.

[31] Piantadosi, S.T., 2014. Zipf's word frequency law in natural language: A critical review and future directions. Psychonomic Bulletin & Review. 21, 1112–1130.

[32] Kulig, A., Drożdż, S., Kwapień, J., et al., 2015. Modeling the average shortest-path length in growth of word-adjacency networks. Physical Review E. 91, 032810.

[33] Stanisz, T., Drożdż, S., Kwapień, J., 2023. Universal versus system-specific features of punctuation usage patterns in major western languages. Chaos, Solitons & Fractals. 168, 113183.

[34] Nakagawa, T., Osaki, S., 1975. The discrete Weibull distribution. IEEE Transactions on Reliability. R-24(5), 300–301.

[35] Weibull, W., 1951. A statistical distribution function of wide applicability. ASME Journal of Applied Mechanics. 18(3), 293–297.

[36] Johnson, N.L., Kotz, S., Balakrishnan N., 1994. Continuous Univariate Distributions, Vol. 1. Wiley-

Interscience: New York, NY, USA.
[37] Miller Jr., R.G., 1998. Survival Analysis. John Wiley & Sons: New York, NY, USA.
[38] Chen, H., Chen, X., Liu, H., 2018. How does language change as a lexical network? An investigation based on written Chinese word co-occurrence networks. PLOS One. 13, e0192545.
[39] Solé, R.V., Corominas-Murtra, B., Valverde, S., et al., 2010. Language networks: Their structure, function, and evolution. Complexity. 15, 20–26.
[40] Kurzynski, M., 2026. Cognitive stylometry: A computational study of defamiliarization in modern Chinese. Computational Humanities Research. 2, E1.
[41] Amancio, D.R., Nunes, M.G.V., Oliveira Jr., O.N., et al., 2011. Using metrics from complex networks to evaluate machine translation. Physica A. 390, 131–142.
[42] Stanisz T., Kwapień J., Drożdż S., 2019. Linguistic data mining with complex networks: A stylometric-oriented approach. Information Sciences. 482, 301–320.
[43] Guan, X., 2002. The History of Punctuation in Ancient China. Bashu Shushe: Chengdu, China. (in Chinese)
[44] Forester, E., 2018. 1919—The Year that Changed China: A New History of the New Culture Movement. De Gruyter Oldenbourg: Berlin, Germany; Boston, MA, USA.
[45] Mullaney, T.S., 2017. Quote unquote language reform: New-style punctuation and the horizontalization of Chinese. Modern Chinese Literature and Culture. 29(2), 206–250.
[46] Zeng, Y., Cui, J., Li, L., 2025. Fractal illusions: An experimental study of long-range sentence-length correlations in randomly generated natural language texts. arXiv preprint. arXiv:2508.19782. DOI: https://doi.org/10.48550/arXiv.2508.19782
[47] He, S., Ibrahim, N.A., Kang, M.S., 2025. Bridging website communication and language: Translation quality assessment of Chinese medical texts on Malaysian medical tourism websites. Theory and Practice in Language Studies. 15(6), 1938–1948.
[48] Chafe, W., 1988. Punctuation and the prosody of written language. Written Communication. 5(4), 395–426.
[49] Guerreiro, N.M., Rei, R., Batista, F., 2021. Towards better subtitles: A multilingual approach for punctuation restoration of speech transcripts. Expert Systems with Applications. 186, 115740.
[50] Han, Y., 2025. Normalization or creation? A corpus-based study of normalization in the Chinese translation of English children's literature. Humanities and Social Sciences Communications. 12, 1051.
[51] Dec, J., Dolina, M., Drożdż, S., et al., 2025. Average shortest-path length in word-adjacency networks: Chinese versus English. Physical Review E. 112, 064318.
[52] Burger, J.M., 2018. The Effect of Iteratively Applying Plain Language Techniques in Forms and Their Terms and Conditions [Master's Thesis]. Stellenbosch University: Stellenbosch, South Africa.
[53] Szymaszek, P., 2021. Data Driven Methods for Analysis and Improvement of Academic English Writing Exercises [Master's Thesis]. Aalto University School of Science: Espoo, Finland.
[54] Gao, X., 2000. Soul Mountain. Cosmos Books: Beijing, China. (in Chinese)
[55] Gao, X., 2001. Soul Mountain. Lee, M. (Trans.). Harper Perennial: New York, NY, USA.
[56] Ling, D., 1956. The Sun Shines over the Sanggan River. People's Literature Publishing House: Beijing, China. (in Chinese)
[57] Ling, D., 1984. The Sun Shines over the Sanggan River. Jenner, W.J.F.(Trans.). Foreign Languages Press: Beijing, China.
[58] Liu, Y., 2018. The Drunkard. People's Literature Publishing House: Beijing, China. (in Chinese)
[59] Liu, Y., 2020. The Drunkard. Yiu, C.C.-L. (Trans.). Chinese University of Hong Kong Press: Hong Kong, China.
[60] Ha, L., Sicilia-Garcia, E.I., Ming, J., et al., 2003. Extension of Zipf's law to word and character n-grams for English and Chinese. International Journal of Computational Linguistics & Chinese Language Processing. 8(1), 77–102.
[61] Sun, J., 2012. "Jieba" (Chinese for "to stutter") Chinese text segmentation: Built to be the best Python Chinese word segmentation module. Available from: https://github.com/fxsjy/jieba (cited 3 January 2026).
[62] Stanisz, T., Drożdż, S., Kwapień, J., 2024. Statistics of punctuation in experimental literature—The remarkable case of Finnegans Wake by James Joyce. Chaos. 34, 083124.
[63] Sichel, H., 1974. On a distribution representing sentence-length in written prose. Journal of the Royal Statistical Society A. 137(1), 25–34.
[64] Williams, C.B., 1940. A note on the statistical analysis of sentence-length as a criterion of literary style. Biometrika. 31(3–4), 356–361.
[65] Xiao, H., 2008. On the applicability of Zipf's law in Chinese word frequency distribution. Journal of Chinese Language and Computing. 18(1), 33–46.
[66] Yule, G.U., 1939. On sentence-length as a statistical characteristic of style in prose: With applications to two cases of disputed authorship. Biometrika. 30(3–4), 363–390.
[67] Kantelhardt, J.W., Zschiegner, S.A., Koscielny-Bunde, E., et al., 2002. Multifractal detrended fluctuation analysis of nonstationary time series. Physica A: Statistical Mechanics and its Applications. 316(1–4), 87–114.
[68] Oświęcimka, P., Kwapień, J., Drożdż, S., 2006. Wavelet versus detrended fluctuation analysis of multifractal structures. Physical Review E. 74, 016103.
[69] Barabási, A.-L., Vicsek, T., 1991. Multifractality of self-affine fractals. Physical Review A. 44, 2730.

[70] Kwapień, J., Blasiak, P., Drożdż, S., et al., 2023. Genuine multifractality in time series is due to temporal correlations. Physical Review E. 107, 034139.
[71] Drożdż, S., Kwapień, J., Oświęcimka, P., et al., 2009. Quantitative features of multifractal subtleties in time series. EPL: Europhysics Letters. 88, 60003.
[72] Ausloos, M., 2012. Generalized Hurst exponent and multifractal function of original and translated texts mapped into frequency and length time series. Physical Review E. 86, 031108.
[73] Hurst, H.E., 1951. Long-term storage capacity of reservoirs. Transactions of the American Society of Civil Engineers. 116(1), 770–799.
[74] Halsey, T.C., Jensen, M.H., Kadanoff, L.P., et al., 1986. Fractal measures and their singularities: The characterization of strange sets. Physical Review A. 34, 1601.
[75] Nakao, H., 2000. Multi-scaling properties of truncated Lévy flights. Physics Letters A. 266(4–6), 282–289.
[76] Liu, J., Gunn, E., Youssef, F., et al., 2023. Fractality in Chinese prose. Digital Scholarship in the Humanities. 38(2), 604–620.
[77] Drożdż, S., Oświęcimka, P., 2015. Detecting and interpreting distortions in hierarchical organization of complex time series. Physical Review E. 91, 030902(R).
[78] Shanahan, M., McDonell, K., Reynolds, L., 2023. Role play with large language models. Nature. 623, 493–498.
[79] Wang, Z., Xu, G., Ren, M., 2026. "Language is the dress of thought": A new method for automatic detection of AI-generated text. Decision Support Systems. 201, 114578.
[80] Guo, Z., Deng, J., Zou, Y., et al., 2024. A hybrid method of combination probability and machine learning for Chinese geological text segmentation. Computers & Geosciences. 183, 105512.
[81] Wang, B., Tang, F., 2020. Corpus-based interpreting studies in China: Overview and prospects. In: Hu, K., Kim, H.K. (Eds.). Corpus-based Translation and Interpreting Studies in Chinese Contexts. Palgrave Studies in Translating and Interpreting. Palgrave Macmillan: Cham, Switzerland. pp.61–87.