Modeling the Law of Abbreviation in Classical, Modern, and ChatGPT-Generated Chinese: A Power-Law Analysis of Structural Economy

Jianwei Yan[1], **Heng Chen**[2]

[1] Department of Linguistics, Zhejiang University;
[2] Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies.
jwyan@zju.edu.cn; chenheng@gdufs.edu.cn

Abstract

This study investigates the Law of Abbreviation—the inverse relationship between word length and frequency—across Classical, Modern, and ChatGPT-generated Chinese. Using a tri-partite parallel corpus and a power-law model $y = a * x^{-b}$, we analyze the relationship between word length and the average usage frequency of words within a given word length category to assess structural economy. Results confirm consistent Zipfian distribution across all text types, with high R^2 values indicating strong model fit. However, the parameter b varies significantly: Classical Chinese shows the steepest decline, suggesting strong pressure for brevity; Modern Chinese exhibits a moderated pattern; ChatGPT-generated texts display the weakest pressure, prioritizing fluency over compression. These differences reflect evolving communicative priorities and reveal that while AI models can mimic statistical distributions, they underrepresent deeper structural pressures found in natural language evolution. This study offers insights into lexical optimization and the parameter b offers a useful metric for comparing structural efficiency across modalities. Implications are discussed in relation to language modeling, cognitive economy, and the evolution of linguistic structure.

1 Introduction

One of the most enduring empirical patterns in quantitative linguistics is the inverse relationship between word length and word frequency, commonly known as the Law of Abbreviation. Zipf (1935: 25) famously hypothesized that "the magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences," and also asserted that "the larger a word is in length, the less likely it is to be used" (Zipf, 1935: 22). This observation has given rise to a long-standing discussion regarding the appropriate directionality of modeling: should word length be treated as a function of frequency, or should frequency be modeled as a function of word length (Strauss et al., 2007: 277)?

Researchers who model word length as a function of frequency often draw on Zipf's (1949) "principle of least effort," which suggests that frequently used linguistic forms tend to be shorter for communicative efficiency (e.g., Köhler, 1986). For instance, Breiter (1994) found that higher-frequency words tend to be shorter based on frequency dictionaries. Wang (2014) used corpus data and confirmed a negative correlation between frequency and length in Chinese, consistent with a power-law distribution. Moreover, Bentz and Ferrer-i-Cancho (2016), in a large-scale cross-linguistic study covering 1,262 texts in 986 languages, found robust negative correlations between frequency and word length, attributing the universal pattern to fundamental principles of information processing: "Words that are used more frequently tend to be shorter" (p. 1).

In contrast, other scholars have argued for the reverse modeling direction, treating word frequency as a function of word length. This approach aligns with Zipf's (1935: 22) assertion that longer words are inherently less frequent,

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 56–62
August 29, 2025

reflecting structural constraints on usage. It has been adopted by Miller et al. (1958), Chen et al. (2015), Linders and Louwerse (2023), and Li and Lei (2025), etc. For example, Linders and Louwerse (2023) demonstrated that the Law of Abbreviation holds in natural spoken dialogues, extending prior findings based on written corpora. Li and Lei (2025) validated the law in Chinese texts across four genres, revealing that while the inverse relationship persists, genre-specific factors and character polysemy may modulate the strength of the effect.

While both approaches may be mathematically equivalent under parameter transformation (Strauss et al., 2007: 279-280), they embody distinct theoretical assumptions, and entail different perspectives on how form and usage interact. Modeling word length as a function of frequency suggests that linguistic structure is shaped by language use—frequent forms tend to become shorter over time. Conversely, modeling frequency as a function of word length assumes that structural features of language constrain how often a form is used, with shorter or simpler forms being more cognitively efficient and therefore more likely to recur.

Notably, most prior studies have focused on languages using alphabetic writing systems, particularly those using Latin scripts such as English, Dutch, or German. Research on Chinese remains limited, with only a handful of studies examining the length–frequency relationship (e.g., Breiter, 1994; Wang, 2014; Chen et al., 2015; Li and Lei, 2025). Moreover, few studies utilize parallel corpora or consider how emerging forms of language generation—such as AI-generated text—may reflect or deviate from natural linguistic patterns. This study addresses both gaps by focusing on Chinese and incorporating AI-generated texts as a comparison.

In this study, we adopt the modeling perspective that treats frequency as a function of word length, based on three key considerations. First, this direction emphasizes the structural constraints that word form imposes on usage, aligning with the linguistic principle that shorter forms are more cognitively efficient and thus more likely to recur. Second, in diachronic and cross-system comparisons, word length is more stable than frequency, making it a more reliable independent variable; and treating frequency as a response variable enables us to assess whether different language production systems—including large language models—adhere to the same efficiency principles observed in human language. Finally, this approach is empirically grounded in Chen et al. (2015), who modeled the length–frequency relationship in Chinese using the power-law function and demonstrated that the parameter b captures the rate at which average usage frequency decreases with increasing word length, reflecting the evolutionary dynamics of the Chinese lexicon: A larger value of b indicates a steeper decline in frequency as length increases, signaling a stronger pressure for efficiency and simplification.

Specifically, we apply the power-law function to the corpus comprises three parallel versions: (1) Classical Chinese texts, (2) their Modern Chinese equivalents, and (3) their Modern Chinese translations generated by ChatGPT from the same Classical Chinese input. This study aims to address three research questions:

Question 1: Does the inverse relationship between word length and word frequency hold consistently across Classical Chinese, Modern Chinese, and ChatGPT-generated Chinese?

Question 2: How do the fitted power-law parameters, namely the parameter b, differ across these text types, and what do they reveal about structural pressures toward lexical economy?

Question 3: To what extent does ChatGPT-generated language replicate or diverge from the natural patterns of lexical economy observed in natural languages?

This diachronic and cross-modal design allows us to examine how structural features such as word length influence language usage across ancient, modern, and AI-generated language.

2 Material and Method

2.1 Material

The data for this study were drawn from the Classical-Modern Chinese parallel corpus,[1] which provides sentence-aligned pairs of Classical Chinese texts and their Modern Chinese

1 https://github.com/NiuTrans/Classical-Modern

equivalents. From this corpus, we randomly extracted ten Classical Chinese texts, each with 200 sentences, and their sentence-aligned Modern Chinese equivalents, yielding ten parallel pairs and a total of 2,000 aligned sentence pairs.

To generate the AI-translated dataset, we prompted ChatGPT-4o using batches of 100 Classical Chinese sentences with the following instruction (original in Chinese, with an academic English translation provided below):

Chinese prompt: "以下是 100 个文言文句子，请将这些句子翻译为流畅、自然的现代汉语。翻译时不必拘泥于文言原文的句式结构，重点在于准确传达原意，使现代读者易于理解。请仅输出现代汉语译文，保持语义准确、语言通顺，避免逐字直译。"

English translation: "The following are 100 Classical Chinese sentences. Please translate them into fluent and idiomatic Modern Chinese. You are not required to adhere strictly to the syntactic structures of the source text; instead, prioritize conveying the intended meaning clearly and naturally for a contemporary readership. Provide only the translated Modern Chinese sentences. Ensure semantic fidelity and linguistic fluency, and avoid literal, word-for-word translation."

All ChatGPT translations were generated in separate sessions, each using the same prompt. This procedure yielded ChatGPT-generated translations for each Classical Chinese sentence, resulting in aligned triplets for every source sentence: (1) the Classical Chinese, (2) the Modern Chinese version, and (3) the ChatGPT-generated Modern Chinese version. Each text type comprised 10 files, with 200 sentences per file.

Then, text segmentation was conducted using language-specific tools. The Classical Chinese texts were segmented with *udkanbun* (Yasuoka, 2019),[2] a syntactic parser based on Universal Dependencies and specifically designed for Classical Chinese (漢文/文言文). For both human and ChatGPT-generated Modern Chinese texts, segmentation was performed with *stanza* 1.10.1 (Qi et al., 2020),[3] a Python-based NLP toolkit supporting multiple languages including Chinese. The segmentations were manually checked to ensure accuracy. An overview of token and type counts for the three versions across the ten files is presented in Table 1.

Table 1: Word counts of the 30 text files.

Classical		Modern		ChatGPT	
token	**type**	**token**	**type**	**token**	**type**
3,234	1,163	3,468	1,929	2,860	1,681
3,181	1,160	3,325	1,899	2,784	1,724
3,172	1,184	3,356	1,852	2,797	1,673
3,202	1,185	3,466	1,940	2,953	1,726
3,062	1,112	3,173	1,788	2,749	1,647
3,137	1,146	3,417	1,860	2,919	1,763
3,111	1,170	3,265	1,853	2,913	1,708
3,031	1,136	3,265	1,844	2,665	1,653
3,331	1,231	3,535	1,993	2,913	1,721
3,309	1,222	3,774	2,061	2,899	1,739

2.2 Method

There are various approaches to evaluating the relationship between word length and word frequency, including non-parametric methods, linear mixed-effects regression models, and power-law formulations. For example, Bentz and Ferrer-i-Cancho (2016) employed a non-parametric approach using Kendall's τ, avoiding any specific functional form. Li and Lei (2025) predicted log-transformed frequency from character length using linear mixed-effects models. In addition, numerous equations describing the relationship between word length and frequency (or frequency rank) have been theoretically developed and employed in empirical studies (Ferrer-i-Cancho, 2025).

Informed by prior studies and for consistency with diachronic studies on Chinese, the present study adopts the modeling approach of Chen et al. (2015). Specifically, we fit the data to a power-law function of the form: $y = a * x^{-b}$, where x is word length (in characters), y is the mean word ratio (MWR), calculated as the token count divided by the type count for each word length class, indicating the average usage frequency of words within a given word length. Parameters a and b are estimated from the data.

[2] https://github.com/KoichiYasuoka/UD-Kanbun

[3] https://github.com/stanfordnlp/stanza

For each segmented text file, we recorded: (1) Tokens—the total number of word occurrences of a given length; (2) Types—the number of unique words of that length; and (3) MWR—the ratio of tokens to types. For example, in the segmented Classical Chinese sentence "余 / 啖 / 林檎 / 一 / 枚 / ， / 梨 / 二 / 枚 / ， / 山胡桃 / 五 / 枚 (I ate one apple, two pears, and five hickory nuts), words of one character appear 9 times (tokens, 余, 啖, 一, 枚, 梨, 二, 枚, 五, 枚) across 7 unique items (types, 余, 啖, 一, 枚, 梨, 二, 五), yielding an MWR of 1.29. Two-character (林檎) and three-character (山胡桃) words each occur once, with 1 token and 1 type, resulting in an MWR of 1.00.

We then fit a power-law model to the data, and computed the coefficient of determination (R^2) to assess the goodness of fit. The parameter *b* serves as an index of structural economy in the lexicon and is used to trace diachronic trends and production modality effects on the length–frequency relationship.

3 Results and Discussion

3.1 Regularity of the Inverse Length–Frequency Relationship

To empirically test the universality of the inverse relationship between word length and frequency in Chinese, we fitted power-law functions to all 30 texts. Table 2 reports the goodness of fit, and Appendix A presents the fitted curves and observed data for each text.[4]

Traditionally, an R^2 value greater than 0.9 (Mačutek and Wimmer, 2013: 233) or 0.8 (Eom, 2006: 121) is considered satisfactory. As shown in Table 2 and Appendix A, all three text types demonstrate excellent model fit, with most texts achieving R^2 exceeding 0.9, indicating that the power-law relationship holds robustly. The results consistently support the hypothesis of structural economy: as word length increases, average frequency usage sharply declines. For Classical Chinese, the R^2 values range from 0.8532 to 0.9921 (M = 0.9655, SD = 0.0426). For Modern Chinese, R^2 values range from 0.8288 to 0.9533 (M = 0.9195, SD = 0.0373). For ChatGPT-generated Chinese, the results are similarly robust, with R^2 values between 0.8355 and 0.9550 (M = 0.9210, SD = 0.0342).

These findings empirically confirm that Zipf's Law of Abbreviation is consistently observed across all three modalities, demonstrating the robustness of the inverse length–frequency relationship in Chinese, providing evidence that this relationship reflects a pervasive regularity of lexical systems. These results echo the cross-linguistic patterns reported by Bentz and Ferrer-i-Cancho (2016), and further suggest that even large language models like ChatGPT reproduce this statistical regularity—possibly as a byproduct of optimizing communicative efficiency.

Table 2: Power-law modeling results of word length-frequency distributions in 30 text files.

Classical		**Modern**		**ChatGPT**	
R^2	***b***	R^2	***b***	R^2	***b***
0.9779	1.1008	0.8288	0.7749	0.8355	0.7173
0.9358	0.8996	0.9194	0.9122	0.9411	0.9566
0.9829	1.0290	0.9524	1.0296	0.9164	0.8451
0.9873	1.0235	0.9237	0.9085	0.9207	0.8930
0.9921	1.0254	0.9243	0.9059	0.9550	0.9437
0.9808	0.9130	0.9179	0.9406	0.9165	0.8159
0.9716	1.0593	0.8911	0.8288	0.9509	1.0032
0.9863	1.0102	0.9533	1.0560	0.9105	0.7463
0.9872	1.0174	0.9524	1.0959	0.9453	0.9716
0.8532	0.8205	0.9318	0.9948	0.9178	0.8231

3.2 Variation in the Power-Law Parameter *b* and Lexical Economy

As demonstrated by Chen et al. (2015), the parameter *b* in the power-law model reflects the rate at which average word frequency decreases with increasing word length, thus serving as a quantitative index of lexical economy. A higher *b*-value indicates a steeper decline, suggesting a stronger systemic preference for brevity through more frequent use of shorter words.

As shown in Table 2 and Appendix A, the average *b*-values across the three text types reveal meaningful distinctions in lexical economy. Classical Chinese exhibits the highest *b*-value,

[4] Although the consistently high R^2 values and visualizations confirm the Zipfian patterns across the three text types, a modeling-related limitation remains—namely, the small number of word length categories available for power-law fitting, particularly in some Classical Chinese texts. Future research may adopt alternative models and complementary approaches to triangulate the results and enhance generalizability.

demonstrating the steepest inverse relationship (*M* = 0.9899, *SD* = 0.0850), with *b*-values ranging from 0.8205 to 1.1008. This indicates the strongest structural pressure for brevity, consistent with the compactness and density characteristic of traditional literary forms. Modern Chinese shows a slightly lower mean (*M* = 0.9447, *SD* = 0.1006; range: 0.7749 to 1.0959), suggesting a somewhat weaker pressure for brevity and greater tolerance for longer word forms. In contrast, ChatGPT-generated Chinese presents the smallest *b*-value (*M* = 0.8716, *SD* = 0.0977), with a range from 0.7173 to 1.0032. This comparatively flatter decline implies that the model prioritizes fluency and plausibility over compression, consistent with its generative objective to optimize for readability rather than structural economy.

These findings are supported by lexical diversity patterns (see Table 1). Classical Chinese has the highest token/type ratio (*M* = 2.7136, *SD* = 0.0394), suggesting low lexical diversity and high repetition. Modern Chinese follows with a moderate ratio (*M* = 1.7897, *SD* = 0.0292), while ChatGPT-generated texts show the lowest ratio (*M* = 1.6701, *SD* = 0.0350), indicating greater lexical variety and less repetition. These patterns align with the *b*-value trends: the steeper slope in Classical Chinese reflects tighter lexical economy, whereas the flatter slope in ChatGPT texts suggests weaker brevity constraints.

To statistically assess these differences, a Kruskal–Wallis H test was conducted on the *b*-values, revealing a significant difference among the three text types, $H(2) = 6.68$, $p = 0.036$. Follow-up Mann–Whitney U tests showed a significant difference between Classical Chinese and ChatGPT-generated texts ($p = 0.009$, *Cliff's delta* = 0.70), while comparisons between Classical and Modern ($p = 0.385$, *delta* = 0.24) and between Modern and ChatGPT ($p = 0.162$, *delta* = 0.38) were not statistically significant, though they yielded small-to-moderate effect sizes. Figure 1 visualizes these distributional differences.

Overall, the results reveal a progressive attenuation in the pressure for lexical economy across both historical and generative dimensions. While all three modalities conform to Zipfian scaling, the magnitude of *b* highlights systematic variation in structural optimization: Classical Chinese reflects strong efficiency-driven constraints, Modern Chinese represents a moderated form of such pressure, and ChatGPT-generated language prioritizes fluency, coherence, and accessibility. This diachronic and modality-based divergence likely mirrors evolving communicative priorities. Notably, while ChatGPT reproduces surface-level statistical regularities, it may not fully internalize the deeper structural pressures that govern naturally evolved human language.

The parameter *b*, therefore, serves not only as an indicator of Zipfian adherence but also as a sensitive metric for comparing structural economy across modalities.

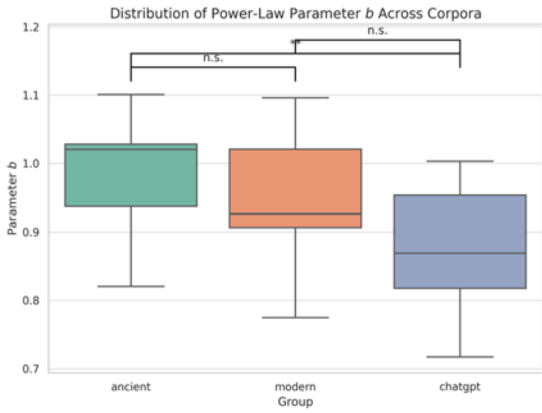


Figure 1: Distribution of power-law *b* values across text types.

3.3 Structural Divergence in Lexical Economy between AI-Generated and Human Texts

Although ChatGPT outputs exhibit high R^2 values—indicating conformity to Zipfian distributions—they consistently yield lower *b*-values than both Classical and Modern Chinese. This suggests that while the model captures the general pattern of frequency decline with word length, the intensity of this relationship is weaker.

This divergence implies that ChatGPT approximates, but may do not fully internalize, the structural constraints underlying natural lexical distributions. Human-authored texts—especially Classical Chinese—reflect strong pressures for brevity and efficiency. In contrast, ChatGPT-generated language appears driven more by statistical plausibility than by structural economy, associating shorter words with higher frequency without being governed by communicative constraints.

The contrast with Classical Chinese is particularly notable. Its historical evolution favored compression and information density—qualities not explicitly optimized in neural models. Instead, ChatGPT is trained to maximize coherence and fluency based on probabilistic exposure, often resulting in less disciplined lexical structures despite surface Zipfian regularity.

Taken together, these findings affirm the generalizability of Zipf's Law of Abbreviation across modalities, while also revealing graded differences in lexical economy.

4 Conclusion

This study examined whether the inverse relationship between word length and frequency—commonly known as the Law of Abbreviation—holds across three language modalities: Classical Chinese, Modern Chinese, and ChatGPT-generated Chinese. Using a power-law model, we analyzed the relationship between word length and average usage frequency measured by MWR across three text types. All texts showed strong Zipfian patterns, with high R^2 values indicating good model fit.

However, significant differences emerged in the fitted b parameter, which reflects lexical economy. Classical Chinese exhibited the largest b-values, indicating the strongest preference for short, frequent words. Modern Chinese showed moderate brevity pressure, while ChatGPT-generated texts had the smallest b-values, suggesting weaker structural constraints. A Kruskal–Wallis H test confirmed significant group differences, and post hoc analysis found a significant contrast between Classical Chinese and ChatGPT, while differences involving Modern Chinese were not significant.

These findings suggest that while large language models like ChatGPT can replicate surface-level Zipfian distributions, they do not fully reproduce the deeper efficiency pressures observed in human-authored language, particularly in highly compressed systems like Classical Chinese. The b parameter thus serves as a useful indicator of structural economy across production modalities.

There are some limitations for this study. First, the corpus was limited to 2,000 sentences per text type. Although balanced and systematically sampled, the dataset may not capture the full lexical variability of each modality. Second, Chinese segmentations relied on the NLP tools which, despite manual check, may overlook certain morphological subtleties. Third, ChatGPT outputs were generated using a fixed prompt under a single condition, which may have limited stylistic variation. Repeating the generation process under varied prompts and conditions may offer greater lexical and stylistic diversity.

Future research could extend the corpus to include larger and more genre-diverse datasets, compare across different LLMs (e.g., DeepSeek, Claude, Gemini, etc.), and incorporate additional complexity metrics such as syntactic depth, semantic density, or information-theoretic entropy. Longitudinal tracking of AI-generated texts across training iterations may also reveal whether structural economy emerges or erodes as model architectures evolve.

Acknowledgments

This work was partly supported by the National Social Science Foundation of China (Grant No. 24CYY064) and the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities in China (Grant No. 22JJD740018). The authors would like to sincerely thank the three anonymous reviewers for their valuable comments and constructive suggestions.

References

Christian Bentz and Ramon Ferrer-i-Cancho. 2016. Zipf's law of abbreviation as a language universal. In Christian Bentz, Gerhard Jäger, and Igor Yanovich (eds.), *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. University of Tübingen, Tübingen, pages 1–4. https://doi.org/10.15496/publikation-10057.

Maria A. Breiter. 1994. Length of Chinese words in relation to their other systemic features. *Journal of Quantitative Linguistics*, 1(3):224–231. https://doi.org/10.1080/09296179408590020.

Heng Chen, Junying Liang, and Haitao Liu. 2015. How does word length evolve in written Chinese? *PLOS ONE*, 10(9): e0138567. https://doi.org/10.1371/journal.pone.0138567.

Jee-Hyeon Eom. 2006. *Rhythmus im Akzent: Zur Modellierung der Akzentverteilung als einer Grundlage des Sprachrhythmus im Russischen*. Otto Sagner, München.

Ramon Ferrer-i-Cancho. 2025 On the class of coding optimality of human languages and the origins of Zipf's law. 1–7. https://doi.org/10.48550/arXiv.2505.20015.

Reinhard Köhler. 1986. *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Brockmeyer, Bochum.

Zhuolun Li and Lei Lei. 2025. Deciphering cross-genre dynamics: Testing the Law of Abbreviation and the Meaning-Frequency Law in Chinese across genres. *Lingua*, 320:1–18. https://doi.org/10.1016/j.lingua.2025.103934.

Guido M. Linders and Max M. Louwerse. 2023. Zipf's law revisited: Spoken dialog, linguistic units, parameters, and the principle of least effort. *Psychonomic Bulletin & Review*, 30:77–101. https://doi.org/10.3758/s13423-022-02142-9.

Ján Mačutek and Gejza Wimmer. 2013. Evaluating Goodness-of-Fit of Discrete Distribution Models in Quantitative Linguistics. *Journal of Quantitative Linguistics*, 20 (3):227–240. https://doi.org/10.1080/09296174.2013.799912.

George A. Miller, Edwin B. Newman, and Elizabeth A. Friedman. 1958. Length-frequency statistics for written English. *Information and Control*, 1(4):370–389. https://doi.org/10.1016/S0019-9958(58)90229-8.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.14.

Udo Strauss, Peter Grzybek, and Gabriel Altmann. 2007. Word length and word frequency. In Peter Grzybek (ed.), *Contributions to the Science of Text and Language,* pages 277–294. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-4068-9_13.

Lu Wang. 2014. Synergetic studies on some properties of lexical structures in Chinese. *Journal of Quantitative Linguistics*, 21(2):177–197. https://doi.org/10.1080/09296174.2014.882186.

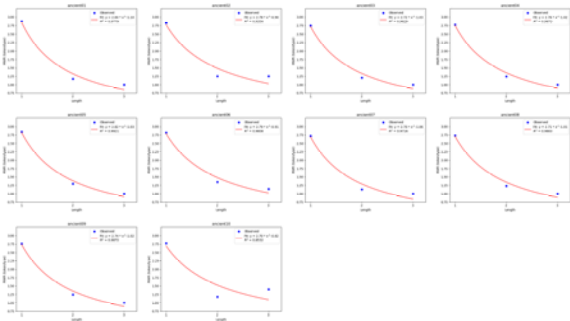
Koichi Yasuoka. 2019. Universal Dependencies Treebank of the Four Books in Classical Chinese. In *Proceedings of the 10th International Conference of Digital Archives and Digital Humanities (DADH2019),* pages 20–28.

George Kingsley Zipf. 1935. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin, Boston.

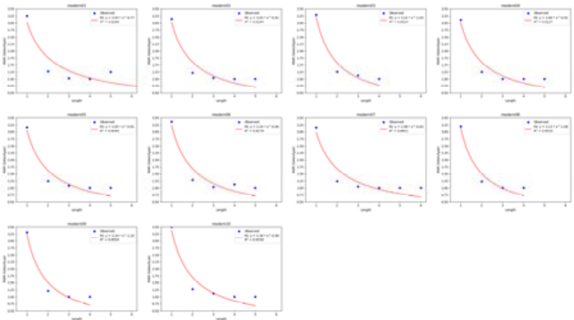
George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Hafner, New York.

Appendix A

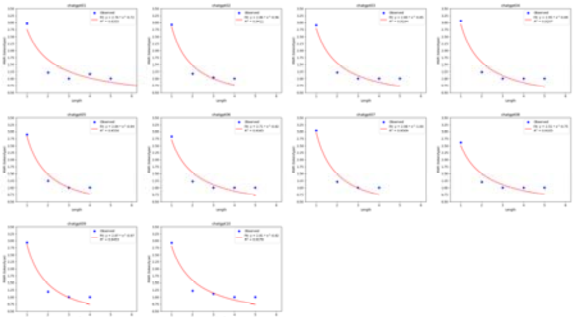
Power-law fitting results for the Classical, Modern and ChatGPT-Generated Chinese texts.



(a) Power-law fitting results for the Classical Chinese texts.



(b) Power-law fitting results for the Modern Chinese texts.



(c) Power-law fitting results for the ChatGPT-Generated Chinese texts.