The Historian's Fingerprint
A Computational Stylometric Study of the Zuo Commentary and Discourses of the States

Wenjie Hua

School of Chinese Language and Literature, Wuhan University, China

huawenjie@whu.edu.cn

Abstract

Previous studies suggest that authorship can be inferred through stylistic features like function word usage and grammatical patterns, yet such analyses remain limited for Old Chinese texts with disputed authorship. Computational methods enable a more nuanced exploration of these texts. This study applies stylometric analysis to examine the authorship controversy between the *Zuo Commentary* and the *Discourses of the States*. Using PoS 4-grams, Kullback-Leibler divergence, and multidimensional scaling (MDS), we systematically compare their stylistic profiles. Results show that the *Zuo Commentary* exhibits high internal consistency, especially in the later eight Dukes chapters, supporting its integration by a single scholarly tradition. In contrast, the *Discourses of the States* displays greater stylistic diversity, aligning with the multiple-source compilation theory. Further analysis reveals partial stylistic similarities among the Lu, Jin, and Chu-related chapters, suggesting shared influences. These findings provide quantitative support for Tong Shuye's arguments and extend statistical validation of Bernhard Karlgren's assertion on the textual unity of the *Zuo Commentary*.

1 Background

Stylometry, which also known as authorship identification, is the process of analyzing textual features to determine uncertain authorship (authorship attribution) or verify an author's identity (authorship verification). The fundamental premise of stylometry is that "authors have an unconscious aspect to their style, an aspect which cannot consciously be manipulated but which possesses features that are quantifiable and may be distinctive." This study applies stylometry to old Chinese texts, focusing on two historical works with contentious authorship: *Zuo zhuan* 左傳 (Zuo Commentary) and *Guo yu* 國語 (Discourses of the States).

The *Zuo Commentary* is a chronicle-style historical record documenting events in Central Plains states during the Spring and Autumn period. Regarding its authorship, figures such as Sima Qian 司馬遷 (145-86? BCE), Ban Gu 班固 (32-92 AD), and Du Yu 杜預 (233-285 AD), along with early records from *The Analects*, identify Zuo Qiu Ming 左丘明, the Tai shi 太史 (historian) of the state of Lu 魯, as the author of the work. Relevant sources include:

> The gentleman of Lu (魯君子), Zuo Qiu Ming, compiled the *Zuo Commentary* to the Spring and Autumn Annals. (Sima Qian, *Records of the Grand Historian - Yearly Chronicle of the Feudal Lords*)

However, started in the Tang dynasty (618-907), historians raised doubts about whether Zuo Qiu Ming mentioned in *Records of the Grand Historian* and *The Analects* were the same person. In the late Qing period, New Text scholars argued more broadly that the *Zuo Commentary* was a forgery by Liu Xin 劉歆 (50?-23 BCE). Due to the absence of further archaeological evidence, contemporary views generally accept the reliability of Sima Qian's records.

The *Discourses of the States* is a state-specific historical record. It primarily focuses on events during the Spring and Autumn period, with some content overlapping with the *Zuo Commentary*. The text comprises 21 chapters, each consisting of independent speeches or dialogues.

Concerning its authorship, Sima Qian, followed by Ban Gu, attributed it to the same author as the *Zuo Commentary*:

> Zuo Qiu lost his sight and finished the *Discourses of the States*. (Sima Qian, *Records of the Grand Historian* - Autobiographical Afterword of the Grand Historian)

Confucius composed the *Spring and Autumn Annals* based on Lu's historical records, and Zuo Qiu Ming organized their accounts as commentary and further compiled divergences into the *Discourses of the States*. (Ban Gu, *Book of Han* - Sima Qian)

Karlgren (1968) conducted the earliest stylometric analysis of the *Zuo Commentary* and the *Discourses of the States*. By manually selecting seven sets of function words from the *Zuo Commentary*, Karlgren compared its linguistic style with *The Analects* and *Mencius*, which represent the linguistic style of the Lu region. He reached two significant conclusions. First, the *Zuo Commentary* is either the work of a single author or represents a specific school, as it exhibits a high degree of internal consistency. However, it does not reflect the style of Lu's "gentlemen of Lu." Second, the grammar of the *Zuo Commentary* is very similar to that of the *Discourses of the States*. Hirase (1998) affirmed Karlgren's judgment and used subsequent research on calendar records to demonstrate a connection between the author of the *Zuo Commentary* and the court of the State of Han 韓.

2 Tools and Corpus

Distinguishing from earlier studies reliant on manual annotation and subjective judgment, since the 1980s, the introduction of digital tools and mathematical methods has made stylometry more feasible for processing vast corpora, providing repeatable experimental procedures and objective metrics. Building upon the methodology of Karlgren, who conducted full-text statistics on the *Zuo Commentary* and *Discourses of the States*, we expanded the sample size to uncover linguistic style differences more comprehensively, which aims to conduct a mega-size computational analysis of the *Zuo Commentary* and the *Discourses of the States*, presenting the results visually and intuitively. For word segmentation (WS) and part-of-speech tagging (POS), Jiayan, a professional Python-based NLP tool for old Chinese, Jiayan[1] was utilized for this step. After statistics and calculation processing, we employed visualization tools to output the results.

The version of the *Zuo Commentary* commonly used today originates from the Western Jin dynasty,

where Du Yu reorganized the single spread text (單篇別行) by integrating it into the *Chun Qiu* 春秋 (the Spring and Autumn Annals), aligning the commentary years with the corresponding years in the Annals, a process described as "attaching the commentary years to the Annals years" (分經之年與傳之年相附) (Ma, 1992). This version, referred to as *Chun Qiu Zuo zhuan* 春秋左傳 (*Zuo Commentary to the Spring and Autumn Annals*), serves as part of the corpus for our study. In order to facilitate analysis, we divided the *Zuo Commentary* into twelve parts, corresponding to the twelve rulers. Similarly, the *Discourses of the States* was divided into twenty-one parts, based on its chapters.

3 Methodology

3.1 POS n-grams

POS n-grams, defined as sequences of n consecutive part-of-speech tags, offer significant advantages as higher-order POS features. They capture subtle stylistic differences that can be indicative of authorship. POS n-grams have demonstrated strong performance in previous studies (Martinc et al., 2017; Siagian and Aritsugi, 2019). Similar methods have also been applied to post-Classical Chinese literature. For example, Liu and Xiao (2015) provides valuable insights; however, there are currently no precedents for applying such methods to Old Chinese.

Compared to analyzing individual POS tags or simple POS elements, n-grams demonstrate greater robustness, particularly for short, stylistically diverse texts. This makes them highly suitable for analyzing Old Chinese texts.

After comparing p-values, mean differences, and contrastive analysis for n = 2, 3, and 4, we found that 4-grams offer the best overall performance in terms of statistical significance, contextual coverage, and text differentiation power. Therefore, we computed POS 4-gram statistics for individual chapters and constructed cosine similarity matrices. Retaining only the most frequent n-grams helped reduce noise and sparsity.

If a text exhibited clear clustering patterns among chapters, we computed the group centroid of the J set in the original feature space to represent central stylistic features and identified outlier chapters based on their Euclidean distance to this centroid. Subsequently, we applied multidimensional scaling (MDS) to visualize the distance matrices in low-dimensional space.

3.2 Kullback-Leibler Divergence

Function words reveal writing style (Pennebaker, 2011). Damerau (1975) was the first to propose an authorship identification method based on the frequency of function words. Subsequent studies (Halvani et al., 2020; Zhao and Zobel, 2007) have justified the effectiveness of this approach for alphabetic languages. This stylistic phenomenon is also evident in Old Chinese, where function words are particularly effective in distinguishing writing styles based on authorship, regional characteristics, and temporal context.

Based on previous research, we exhaustively compiled all function words in the two texts and calculated their unary probabilities. Subsequently, we introduced the concept of relative entropy for the same function word across texts and computed its Kullback-Leibler (KL) divergence using the following formula:

$$D(p \parallel q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)} \qquad (1)$$

To assess the disparity between two probability distributions, relative entropy is zero when $p = q$ and increases as their difference grows. KL divergence quantifies this disparity by summing the relative differences across elements. To prevent invalid operations (e.g., division by zero), we replaced zero frequencies with a small constant ϵ for numerical stability.

Using the computed KL divergence as a distance metric, we compiled function word statistics for individual chapters and constructed distance matrices for within-group and between-group comparisons of the *Zuo Commentary* and the *Discourses of the States*. To analyze clustering patterns, we first calculated group centroids in the original feature space and identified outliers via Euclidean distance to these centroids. Then, we used MDS to visualize as well.

4 Results

We calculated and visualized POS 4-grams cosine similarity matrices for the chapters of the *Zuo Commentary* and *Discourses of the States*. The *Zuo Commentary* shows high internal similarity, forming distinct clusters (the J set) with minor variations. In contrast, the *Discourses of the States* displays lower internal similarity. MDS provides us a clear view of internal consistency and divergence.

The *Discourses of the States* chapters show wide dispersion in feature patterns, while the *Zuo Commentary* exhibits strong clustering, especially among the last eight dukes recorded in the *Zuo Commentary* — *Dukes Xi, Wen, Xuan, Cheng, Xiang, Zhao, Ding, and Ai* — forming a cohesive group. In contrast, the first four dukes recorded, *Dukes Yin, Huan, Zhuang, and Min*, display more dispersed patterns (see Figure 1). Comparing independent chapters to the *Zuo Commentary*'s group centroid (calculated from its coordinate positions in MDS space) reveals that *Discourse of Lu-1, Lu-2, Chu-1, Jin-4*, and *Jin-8* exhibit similar stylistic proximities, suggesting overlaps with the *Zuo Commentary* (see Figure 2).
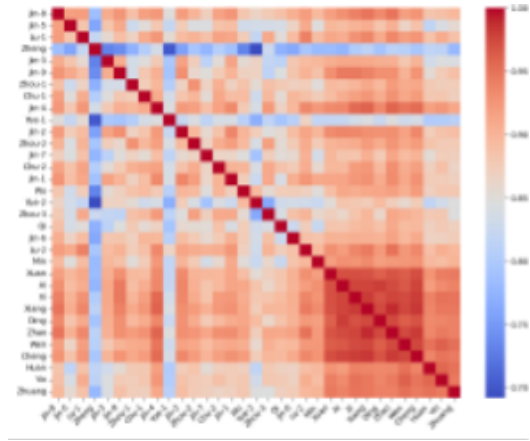

Figure 1: Cosine Similarity of PoS 4-grams in the Zuo Commentary and the Discourses of the States
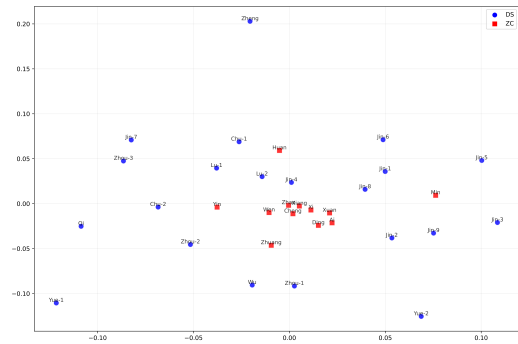

Figure 2: MDS Projection of 4-grams with Chapter Distances

KL divergence values near 0 indicate greater similarity, and its heat-map mirrors the patterns in the prior heat-map (see Figure 3). MDS projection based on KL divergence also aligns with POS 4-grams results, showing clustering for the *Zuo Commentary* and dispersion for the *Discourses of the States*. Notably, the stylistic differences be-

tween the first four and last eight dukes are also evident within the *Zuo Commentary*. Furthermore, the *Discourses of the States* chapters with proximity to the *Zuo Commentary*'s centroid align with the previous results (see Figure 4).
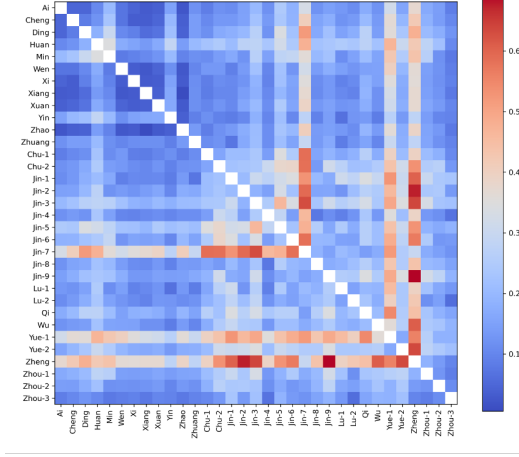

Figure 3: KL Divergence between the Zuo Commentary and the Discourses of the States
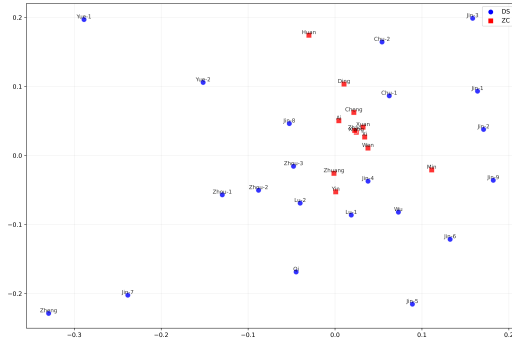

Figure 4: MDS Projection of KL Divergence with Chapter Distances

5 Related Work

Revisiting the history of scholarship, Yao Nai was the first to claim that "(Zuo Commentary) has accumulated additional elements, especially influenced by followers of Wu Qi (累有附益, 而由吳起之徒為之者蓋尤多)" Jiao and Shen (2016), but the evidence remains insufficient. Later, Zhang (1982), based on the statement in *Han Feizi* that "Wu Ch'i was a native of Tso-shih in Wei"（吳起, 衛左氏中人也）(Liao, 1959), inferred that Wu Qi was the author of *Zuo Commentary*. Liu (2008) held a similar view. Subsequently, Tong (2006) listed four pieces of evidence from the perspective of national affairs and more clearly delineated the author's affiliation as "persons related to the states of Lu, Jin, and Chu, along with their disciples." Now, we should admit that Tong (2006)'s perspective is well-founded.

6 Conclusions

This study provides a new empirical perspective on this academic debate through experiments using part-of-speech 4-grams and KL divergence. First, the significant stylistic differences between the *Zuo Commentary* and the *Discourses of the States* quantitatively refute the most traditional view—that both texts originated from Zuo Qiu Ming. Notably, the internal heterogeneity of the *Discourses of the States* aligns with the "multi-source compilation theory" proposed by Zhang (1939) and Wang (1986), which suggests a layered integration of historical materials from multiple states.

In contrast, while the first four chapters exhibit a relatively scattered stylistic pattern, the latter eight chapters of the *Zuo Commentary*, spanning from *Duke Xi* to *Duke Ai*, demonstrate a high degree of homogeneity. Overall, the *Zuo Commentary* maintains strong internal consistency, lending support to Karlgren (1968)'s conclusion that it was authored by either a single individual or, more precisely, a cohesive scholarly school. Furthermore, combining four scattered chapters, we may hypothesize that the diachronic evolution of authorship followed a convergent trajectory—namely, the emergence of this very school likely occurred after the period of *Duke Xi*. Notably, *Discourse of Lu*, *Jin*, and *Chu* exhibit high stylistic similarity to the *Zuo Commentary* in both experiments. These localized similarities strongly align with Tong (2006)'s proposition that the text was influenced by scholarly circles associated with the states of Lu, Jin, and Chu.

In conclusion, computational stylometric not only validates the intuition of earlier scholars that the *Zuo Commentary* was not the work of a single author or period, but also reveals, through quantitative evidence, that its textual unity is more likely derived from a school that integrated elements from Lu, Jin, and Chu. This finding echoes the hypotheses of Zhang (1982) and Liu (2008) regarding Wu Qi's involvement in its authorship while also concretizing Tong (2006)'s school-based theory, offering a new approach to the study of classical text formation.

7 Limitations

This study demonstrates the potential of stylometric methods in analyzing Old Chinese texts and intuitively presenting abstract linguistic features but acknowledges certain limitations. Our methodology assumes stylistic consistency across an author's works regardless of textual content or temporal variation. This assumption, while foundational to stylometry, remains theoretically contested. An author's style may indeed evolve due to genre adaptation or diachronic linguistic changes. Also, the accuracy of the Jiayan NLP toolkit for old Chinese POS tagging has been challenged by models based on BERT(Devlin et al., 2019) and RoBERTa(Liu et al., 2019) in recent years. We should use newer tools to improve accuracy and avoid passing errors to downstream analysis. We believe that future research should explore additional state-level texts to establish indicators beyond the *Discourses of the States*, thereby determining the regional origin of the *Zuo Commentary*. Furthermore, developing benchmark datasets and establishing standardized evaluation frameworks will advance stylometry as a robust discipline.

References

Fred J. Damerau. 1975. The use of function word frequencies as indicators of style. *Computers and the Humanities*, 9(6):271–280.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding.

Oren Halvani, Lukas Graner, and Roey Regev. 2020. Cross-domain authorship verification based on topic agnostic features. In *Conference and Labs of the Evaluation Forum (CLEF)*, Thessaloniki, Greece. European Language Resources Association (ELRA).

Takao Hirase. 1998. *Saden No Shiryō Hihanteki Kenkyū*. Tōkyō Daigaku Tōyō Bunka Kenkyūjo. 101998.

Xun Jiao and Qinhan Shen. 2016. *Supplementary Commentaries on the Zuo Commentary and Additional Annotations to the Zuo Tradition of the Spring and Autumn Annals*. Shanghai Chinese Classics Publishing House, Shanghai.

Bernhard Karlgren. 1968. *On the authenticity and nature of the Tso chuan*. Chengwen Pub. Co., Taipei. Print.

W. K. Liao. 1959. *The Complete Works of Han Fei Tzŭ: A Classic of Chinese Political Science*, volume 2. Arthur Probsthain, 41 Great Russell Street, London, W.C. 1.

Shipei Liu. 2008. *Zuo'an Ji [Collected Works of Liu Shipei]*. China Bookstore.

Ying Liu and Tianjiu Xiao. 2015. Words and n-gram models analysis for "a dream of red mansions". *Data Analysis and Knowledge Discovery*, 31(4):50–57.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Su Ma. 1992. *Zuo Zhuan Shi Wei (Chronological Account of the Zuo Commentary)*. Qilu Press, Jinan.

Matej Martinc, Iza Škrjanec, Katja Zupan, and Senja Pollak. 2017. Pan 2017: Author profiling - gender and language variety prediction. In *Conference and Labs of the Evaluation Forum (CLEF)*, Dublin, Ireland. European Language Resources Association (ELRA).

James W. Pennebaker. 2011. The secret life of pronouns. *New Scientist*, 211(2828):42–45.

Al Hafiz Akbar Maulana Siagian and Masayoshi Aritsugi. 2019. DBMS-KU approach for author profiling and deception detection in arabic. In *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation*, volume 2517 of *CEUR Workshop Proceedings*, pages 115–121, Kolkata, India. CEUR-WS.org.

Shuye Tong. 2006. *Research on Zuo Zhuan of Spring and Autumn Annals (Revised Edition)*. Zhonghua Book Company.

Shu-Min Wang. 1986. The author and compiler of guo yu. *Studies in Chinese Classics*.

Taiyan Zhang. 1982. Article on "qiuming" in volume one of *Reading of Zuo Zhuan of Spring and Autumn Annals*. In Yihua Jiang, editor, *The Complete Works of Zhang Taiyan (Volume 2)*, pages 58–59. Shanghai People's Publishing House.

Xin-Cheng Zhang. 1939. *General Examination of Forged Books, Vol. 1 (Wei Shu Tong Kao, Vol. 1)*, volume 1 of *Studies in Chinese Forgery*. Commercial Press, Shanghai.

Ying Zhao and Justin Zobel. 2007. Searching with style: Authorship attribution in classic literature. In *Proceedings of the 30th Australasian Conference on Computer Science*, ACSC '07, pages 59–68, Ballarat, Australia. Australian Computer Society.

A Appendix

A.1 Corpus Classification

The *Zuo Commentary*

Duke Yin 隱, Duke Huan 桓, Duke Zhuang 莊, Duke Min 閔, Duke Xi 僖, Duke Wen 文, Duke Xuan 宣, Duke Cheng 成, Duke Xiang 襄, Duke Zhao 昭, Duke Ding 定, Duke Ai 哀.

The *Discourses of the States*

the Discourse of Zhou 1 - 3 周語, the Discourse of Lu 1 and 2 魯語, the Discourse of Qi 齊語, the Discourse of Jin 1 - 9 晉語, the Discourse of Zheng 鄭語, the Discourse of Chu 1 and 2 楚語, the Discourse of Wu 吳語, the Discourse of Yue 1 and 2 越語.

A.2 Selection of n (Example)

Text Pair	Statistic	p-value
Xi & Ai	1.4920	0.1373
Xi & Ding	7.6496	4.77E-12
Xi & Xuan	7.2402	3.02E-11
Xi & Zhuang	10.0681	2.28E-18
Xi & Cheng	1.4624	0.1452
Xi & Wen	6.7607	3.27E-10

Table 1: Results of 2-grams on Different Chapters

Text Pair	Statistic	p-value
Xi & Ai	2.7994	0.00565
Xi & Ding	6.2389	4.19E-09
Xi & Xuan	6.4660	1.28E-09
Xi & Zhuang	7.8409	3.39E-13
Xi & Cheng	1.7063	0.0896
Xi & Wen	6.3221	2.60E-09

Table 2: Results of 3-grams on Different Chaptsers

Text Pair	Statistic	p-value
Xi & Ai	4.3110	2.58E-05
Xi & Ding	7.8073	3.69E-13
Xi & Xuan	8.0324	8.73E-14
Xi & Zhuang	13.6326	4.74E-30
Xi & Cheng	0.2850	0.775972457
Xi & Wen	5.9907	1.06E-08

Table 3: Results of 4-grams on Different Chapters

A.3 POS 4-grams Statistics (Example)

POS 4-grams	Count
nt_nt_wp_nh	8
nt_wp_d_v	4
wp_d_v_n	18
d_v_n_wp	24
v_n_wp_v	26
n_wp_v_u	8
wp_v_u_wp	5
v_u_wp_nt	5
u_wp_nt_wp	9
wp_nt_wp_n	18

Table 4: POS 4-grams statistics for Lord Yin

A.4 Function Words Categories

Tag	Description	Example
a	adjective	幽明
b	other noun - modifier	男，女
c	conjunction	與，而
d	adverb	皆
e	exclamation	嗚呼
g	morpheme	甥
h	prefix	非
i	idiom	發憤忘食
j	abbreviation	五帝
k	suffix	者
m	number	一，百
n	general noun	鬼神，山川
nd	direction noun	東，南
nh	person name	軒轅
ni	organization name	遼隊
nl	location noun	城北
ns	geographical name	襄平縣
nt	temporal noun	春，夏
nz	other proper noun	山海經
o	onomatopoeia	嗚嗚
p	preposition	以，為
q	quantity	年，歲
r	pronoun	其，斯
u	auxiliary	之，所
v	verb	賜
wp	punctuation	，。！
ws	foreign words	CPU
x	non - lexeme	萏, 翱
z	descriptive words	默然，區區

Table 5: Function Words Categories

A.5 Function Words Statistics (Example)

Word	PoS	Count	Unary Probability
之	u	119	0.030861
以	p	97	0.025156
不	d	94	0.024378
之	r	71	0.018413
其	r	69	0.017894
而	c	66	0.017116
也	u	61	0.015820
于	p	53	0.013745

Table 6: Function Words in Chu-1

A.6 KL Divergence Statistics (Example)

Comparison	KL Divergence
Yin & Ai	0.4561
Yin & Cheng	0.4393
Yin & Ding	0.5129
Yin & Huan	0.6547
Yin & Min	1.5357
Yin & Wen	0.3379
Yin & Xi	0.2747
Yin & Xiang	0.2994
Yin & Xuan	0.2358
Yin & Zhao	0.3308
Yin & Zhuang	0.5614

Table 7: KL Divergence for Lord Yin