Comparative Analysis of Language Models for Linguistic Examination of Ancient Chinese Classics: A Case Study of Zuozhuan Corpus

Yiqin Zhang
Department of Information Management
Nanjing University
Nanjing, China
yiqin.zhang@smail.nju.edu.cn

Sanhong Deng
Department of Information Management
Nanjing University
Nanjing, China
sanhong@nju.edu.cn

Qi Zhang
Department of Information Management
Nanjing University
Nanjing, China
qi.zhang@smail.nju.edu.cn

Dongbo Wang
Department of Information Management
Nanjing Agricultural University
Nanjing, China
db.wang@njau.edu.cn

Hongcun Gong
Department of Information Management
Nanjing University
Nanjing, China
Hongcun_Gong@163.com

***Abstract*— Exploring the comparative analysis of translation styles across languages holds great significance for capturing the essence of ancient Chinese classics. This article presents a comprehensive analysis of language models for linguistic examination of ancient Chinese classics, using the cross-language Zuozhuan corpus as a focal point. We utilize the capabilities of five pre-trained language models to compare their effectiveness with the deep learning model Bi-LSTM-CRF in the areas of word segmentation and parts-of-speech tagging. Optimal training results obtained from the models facilitated the completion of word segmentation and parts-of-speech tagging across the entire corpus of ancient Chinese classics. Building on these advancements, this research delves into a meticulous lexical-level scrutiny of the linguistic style evident in ancient Chinese classics and their corresponding English translations. In contrast to the original Chinese text, the contemporary Chinese translation exhibits greater semantic clarity, manifesting a relatively singular phrase function and a heightened diversity in vocabulary combinations. Conversely, the English translation demonstrates a tendency toward simplification. The analysis encompasses facets such as parts of speech distribution, word length variation, lexical richness, and textual density, providing unprecedented insights into the cross-linguistic nuances of these venerable literary works.**

Keywords—ancient Chinese classics, cross-linguistic translation, quantitative linguistics, translation styles, corpus-driven methodology

I. INTRODUCTION

The profound cultural heritage embedded within ancient Chinese classics has perpetuated a timeless fascination, prompting renewed scholarly interest in linguistic exploration. In the pursuit of elucidating the intricate linguistic attributes of these texts, this study employs advanced language models to facilitate an in-depth analysis. In the intricate process of internationalizing ancient Chinese classics through translation, profound academic inquiries have meticulously addressed granular facets, encompassing content, methodologies, and the inherent subjects. Yet, a conspicuous oversight persists in examining the holistic stylistic nuances of such translations from an encompassing, macroscopic vantage point—a perspective quintessential to engendering a robust cross-cultural dialogue that seamlessly navigates and bridges linguistic stylistic dichotomies, ensuring resonance with global audiences. As the venerable tapestry of traditional Chinese literary masterpieces undergoes translation and broader global dissemination, a plethora of stylistic discourses has surfaced. These primarily seek to illuminate the intricate variances embedded in translators' stylistic predilections by orchestrating parallel critiques of contemporaneous Chinese or English interpretations [1, 2, 3, 4, 5]. However, within this academic tableau, the trajectory of cross-lingual translation appears somewhat stunted, primarily due to the stark absence of empirical computational stylistic analyses rooted in ancient Chinese [6, 7, 8]. This is further exacerbated by a paucity of exhaustive studies anchored in meticulously aligned translation corpora. In light of these observations, there emerges a pressing academic imperative to delve deeper into the stylistic disparities inherent in classical texts spanning multiple linguistic landscapes. Such endeavors should aim to systematically collate, analyze, and contrast these variances, while concurrently exploring innovative methodologies [9]. The objective is twofold: to both augment the fidelity and verisimilitude of ancient Chinese translations and to safeguard their intrinsic linguistic elegance and aesthetic nuance. Undertakings of this caliber are not mere academic exercises but resonate as crucial linchpins in the adept articulation of the Chinese narrative, thereby amplifying the resonance of China's cultural echo in the global symphony.

Within the ambit of the present investigation, we have meticulously curateda parallel corpus encompassing classical Chinese, modern Chinese, and English, pivoting around the archetypal pre-Qin masterpiece, 'Zuozhuan (Spring and autumn with Zuo's commentary).' This corpus serves as a foundation upon which we orchestrate targeted lexical excavations at the granular vocabulary tier, leveraging the prowess of pre-trained linguistic architectures. Empirical assessments underscore the paramount efficacy of these expansive pre-trained language paradigms when applied to the task of classical text mining. Drawing upon the most salient outcomes of these models for lexical identification, we subsequently undertake a comprehensive stylistic dissection, appraising parameters such as vocabulary breadth, lexical length, richness, and density, all encapsulated within the multifaceted prism of cross-lingual classical textual landscapes. This scholarly enterprise furnishes pivotal revelations, casting light upon arenas like the automated

translation of classical Chinese into English and cognate research trajectories.

II. Related work

A. Computational Stylistic Analysis of Classical Literature

In the realm of historiography, the Zuozhuan has been venerated as a magnum opus that serves as a cornerstone for understanding ancient Chinese literary paradigms [10]. Recognized not only for its substantive content but also its distinct pre-Qin prose, the text represents a veritable confluence of historical narratives and stylistic elegance unique to its epoch. When subjected to cross-language translation, the meticulous characterization and nuances inherent in the Zuozhuan present both a challenge and opportunity for translators. The examination of stylistic disparities inherent in such translation endeavors holds paramount significance for several compelling reasons [11, 12, 13]. Firstly, this scrutiny facilitates a deeper appreciation of the intrinsic textual subtleties, thereby revealing the intricate interplay between linguistic features and the socio-historical contexts from which they originate. Secondly, by evaluating the variations in style across translations, scholars are able to extract insights into broader metalinguistic choices, thereby shedding light on potential shifts in interpretive nuances and cultural resonances that manifest within translated texts. Consequently, this affords an invaluable framework for comprehending the reception, interpretation, and perception of the Zuozhuan across diverse linguistic landscapes. Lastly, the exploration of stylistic aspects in cross-language translations of the Zuozhuan can pave the way for enhanced translation methodologies, attuned to preserving both the semantic integrity and stylistic essence of the sourcetext. This endeavor ensures that the magnificence of this seminal work resonates profoundly across linguistic boundaries and is aptly conveyed to a global audience.

Numerous scholarly inquiries have delved into stylistic divergences present in modern Chinese texts or literary compositions as translated by distinct translators[14, 15]. However, these investigations predominantly remain circumscribed to singular layers of translation, neglecting a comprehensive, macroscopic scrutiny of overarching linguistic attributes inherent in disparate language translations. There is apalpable dearth in profound examinations of cross- linguistic translation styles and the underlying genesis of their stylistic nuances. Both qualitative assessments and empirical methodologies in this domain conspicuously lag in their depth and breadth [16, 17, 18]. Consequently, a rigorous exploration elucidating the stylistic dichotomies amongst classical Chinese, contemporary Chinese, and English texts emerges as not merely an avenue to bridge the existing scholarly lacuna but also as an indispensable fulcrum. This foundation becomes pivotal for ensuing endeavors that aim to accentuate the global resonance of Chinese culture, such as the curation of Sino-foreign classical lexicons, automated translation of canonical works, and pedagogical support in translation instruction.

B. Cross-Language Translation in Computational Linguistics

In the realm of computational linguistics, the exploration of cross-language translation holds profound significance, encapsulating the convergence of linguistic analysis, machine learning, and cultural interchange [19]. The scholarly investigation of how linguistic expression traverses linguistic boundaries has not only revolutionized translation methodologies but has also unearthed intricate layers of linguistic nuance embedded within texts [20] .

Advancements in machine translation, bolstered by the capabilities of large-scale language models, have reshaped the landscape of cross-language translation. Traditional translation methods, reliant on rule-based and statistical approaches, have given way to neural machine translation (NMT) systems that harness the power of deep learning techniques [21, 22]. This paradigm shift has facilitated the development of highly sophisticated models that can discern context,syntactic structures, and cultural subtleties to generate translations of remarkable quality.The role of foreign translations in disseminating traditional Chinese culture is pivotal, with language style disparities arising from cultural variances such as socio-historical context, literary genre, and lexical connotation significantly impacting the translation quality of literary works [23]. At present, scholarly inquiries into the linguistic style of classical works predominantly concentrate on single-language texts. Scholars primarily trace their literary and historical underpinnings through the lens of literature and linguistics, discerning the literary language's character and creative prowess by means of rhetorical devices, linguistic usage, extensive rhetorical devices, and character dialogue [24, 25, 26]. For instance, some researchers have juxtaposed and analyzed the divergence in word frequency levels between original literature and other comparable literature pertaining to specific word types [27, 28]. Subsequently, they have calculated the characteristic coefficient of the word type in specific literature, identifying and scrutinizing representative high-frequency feature words, culminating in the quantification of linguistic style similarities amidst the various schools of the pre-Qin dynasty.

While numerous explorations have delved into stylistic disparities among modern Chinese texts or literary works rendered by distinct translators [29], these investigations have been confined to a single stratum of translation, neglecting comprehensive macro-level linguistic traits in diverse language translations. Furthermore, there has been a dearth of in-depth analysis concerning the stylistic features of cross- language translations and their formative rationales, leaving qualitative analysis and empirical research notably deficient. Thus, the exploration and analysis of language style disparities between ancient Chinese, modern Chinese, and English texts not only bridge extant research lacunae but also furnish a potent stepping stone for subsequent endeavors in promoting Chinese culture globally—such as compiling Chinese and foreign classic dictionaries, automating the translation of classic literature, and facilitating auxiliary translation instruction.

III. Data and Method

This paper selects the "Chinese Philosophy E-Texts Project" online open-access digital library(http://ctext.org/zhs/) and the "Ancient Poetry and Literature" website (https://www.gushiwen.org/) as the sources of the corpus. The "Chinese Philosophy E-Texts Project" website offers a repository of Chinese classical texts, encompassing works primarily from the Pre-Qin and Han periods that span various fields such as philosophy, history, and linguistics. As an open community, the website permits researchers to annotate and supplement texts during the digitization process of classical Chinese texts. This collaborative approach significantly enhances the efficiency of proofreading and

correction, ensuring the continuous updating and iteration of ancient Chinese language resources and maintaining the precision and accessibility of the texts. Since its launch, the online open-access digital resource repository has made significant contributions to the digitization of classical texts, with its data being widely embraced by scholars in the field and employed in digital humanities research. As a result of its reputation, we have chosen this resource repository as our corpus source.

A. Corpus Construction and Annotation

The meticulously constructed Zuozhuan corpus forms the bedrock of this research. Upon obtaining optimal training outcomes from the pre-trained models, the entire corpus undergoes rigorous word segmentation and parts-of-speech tagging processes, augmenting the availability of linguistically annotated ancient Chinese classics.

TABLE I. SAMPLE SENTENCE ALIGNMENT IN MANDARIN CHINESE, ENGLISH AND VERNACULAR

Table Head	Classical Chinese	Modern Chinese	English
1	八月，紀人伐夷。 夷不告，故不書。	八月，纪国人讨伐夷国。 夷国没有前来报告鲁国，所以《春秋》不加记载。	In the eighth month the state of Ji made a punitive expedition against the state of Yi. There was no record about it in the Annals for no report coming from this state.
1.01	八月，紀人伐夷。	八月，纪国人讨伐夷国。	In the eighth month the state of Ji made a punitive expedition against the state of Yi.
1.02	夷不告，故不書。	夷国没有前来报告鲁国，所以《春秋》不加记载。	There was no record about it in the Annals for no report coming from this state.

B. The synthesis of the part-of-speech tagging sets

Due to the considerable disparities between Classical Chinese and Modern Chinese in various aspects such as inflected words, historical and contemporary ambiguities, homophonic characters, function words (such as "而" (ér), "乎" (hū), "焉" (yān), "何" (hé), "也" (yě), "以" (yǐ), etc.), as well as in terms of syntactic structures, and grammar, within the corpora of classical texts, this research employs distinct part-of-speech tagging methodologies. In the analysis of part-of-speech tagging for Classical Chinese, this tag set is constructed based on the inherent linguistic characteristics of pre-Qin classical texts, in conjunction with the foundational set of word classes annotated for pre-Qin Chinese language by Nanjing Normal University, the part-of-speech annotation set from the Institute of Computing at the Chinese Academy of Sciences (NLPIR), and the part-of-speech annotation set from the LTP project at Harbin Institute of Technology. For the purpose of Modern Chinese, the ICTPOS3.0 Chinese Part-of-Speech Tagging Set is employed.

In the realm of English language tagging, the methodology outlined in the Penn Treebank is utilized. This tagging set encompasses a comprehensive array of features, including part-of-speech tags, phrase structure tags, and grammatical tags, rendering it notably comprehensive and systematically structured. The synthesis of the part-of-speech tagging sets for classical texts, Modern Chinese, and English is ultimately presented in the tabular format depicted below.

TABLE II. SET OF ENGLISH PART-OF-SPEECH MARKERS

Abbreviate	Full name	Abbreviate	Full name
a	Adjective	n	General noun
c	Conjunction	nr	Personal name
d	Adverb	p	Preposition
dt	Determiner	r	Pronoun
EX	Existential there	RP	Particle
FW	Foreign word	v	Verb
m	Cardinal number	w	Punctuation
MD	Modal	y	Interjection

TABLE III. SET OF ANCIENT CHINESE PART-OF-SPEECH MARKERS

Abbreviate	Full name	Abbreviate	Full name
a	Adjective	ns	Location noun
c	Conjunction	p	Preposition
d	Adverb	q	Quantity
f	Direction noun	r	Pronoun
gv	Ancient Chinese verb	t	Temporal noun
j	Concurrently kind of word	u	Auxiliary
m	Number	v	Verb
n	General noun	w	Punctuation
nr	Personal name	y	Interjection

TABLE IV. SET OF MODERN CHINESE PART-OF-SPEECH MARKERS

Abbreviate	Full name	Abbreviate	Full name
a	Adjective	nz	Proper noun
c	Conjunction	p	Preposition
d	Adverb	q	Quantity
f	Direction noun	r	Pronoun
g	Morpheme	t	Temporal noun
i	Idiom	u	Auxiliary
m	Number	v	Verb
n	General noun	x	Non-morphemic words
nr	Personal name	y	Interjection
ns	Location noun	z	Status

C. Experimental model

In the dynamic landscape of natural language processing and computational linguistics, the pursuit of extracting intricate language styles from textual data has led to the advent of innovative methodologies. Among these, the utilization of Bi-LSTM-CRF [30], BERT (Bidirectional Encoder Representations from Transformers) [31], and Large Language Models (LLMs) stands as a testament to the fusion of advanced machine learning techniques with linguistic analysis [32]. This discourse presents a comprehensive exploration of the principles underlying these models and their applications within the realm of language style mining research.

1) Bi-LSTM-CRF: The Bi-LSTM-CRF architecture, an amalgamation of Bidirectional Long Short-Term Memory (Bi-LSTM) networks and Conditional Random Fields (CRF), embodies the essence of sequential labeling tasks. Bi-LSTM networks excel in capturing contextual dependencies through bidirectional traversal of sequences. This is complemented by the CRF layer, which models the transition probabilities between sequential labels, fostering coherence in labeling sequences. Applied in language style mining research, Bi- LSTM-CRF can delineate linguistic attributes such as part- of-speech tagging and named entity recognition. By capturing contextual intricacies and sequential patterns, Bi- LSTM-CRF unveils language nuances, enabling comprehensive analyses of style variations across diverse textual corpora.

2) Pre-trained language models: BERT, a revolutionary pre-trained language model, harnesses the transformative power of transformer architecture. Pre-trained on massive text corpora, BERT exhibits unparalleled proficiency in understanding contextual relationships within sentences. This "masked language model" learns bidirectional representations by predicting masked words in sentences, enabling it to grasp intricate language contexts. In language style mining research, BERT's contextual embeddings empower the extraction of semantic nuances, syntactic structures, and even sentiment fluctuations across texts [33]. This model's fine-tuning capability tailors it to specific tasks, facilitating tasks such as sentiment analysis, style transfer, and even the detection of authorship shifts.

3) Large Language Models: LLMs(Large Language Models), epitomized by GPT (Generative Pre-trained Transformer) models, epitomize the synergy of neural networks and vast text corpora. Trained on diverse linguistic patterns, LLMs transcend conventional boundaries by generating coherent and contextually plausible text. Their unsupervised learning paradigm allows for the acquisition of linguistic features at unprecedented scales. Within language style mining research, LLMs contribute to stylistic analysis by unraveling intricate sentence structures, syntactic variations, and even latent semantic nuances. Additionally, LLMs can generate text in specific styles or emulate literary voices, facilitating tasks such as style imitation or creative writing augmentation.

IV. Experiments and Results

A. Word Segmentation and Part-of-Speech Tagging Tasks

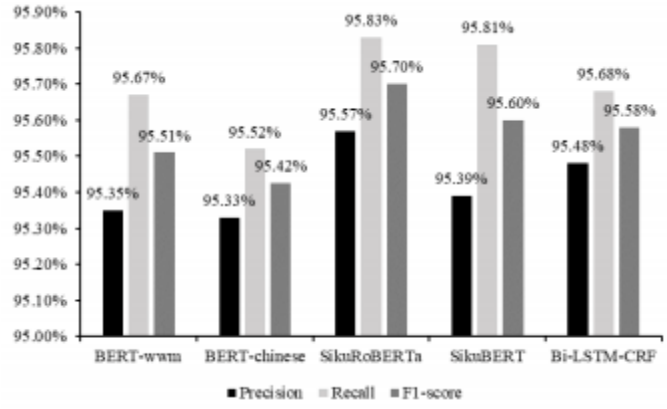


Fig. 1. Model performance comparison chart

The provided table presents the performance metrics of various models in a part-of-speech tagging task, measured in terms of Precision, Recall, and F1-score. Among the models, SikuRoBERTa attained the highest Precision, Recall, and F1-score, achieving 95.57%, 95.83%, and 95.70% respectively. This indicates its exceptional ability to accurately identify and classify parts of speech. Following closely, SikuBERT exhibited competitive results with a Precision of 95.39%, Recall of 95.81%, and F1-score of 95.60%. Both SikuRoBERTa and SikuBERT seem to excel in capturing linguistic nuances. BERT-wwm and BERT-chinese models displayed almost identical performance, with a F1-score of 95.51% and 95.42% respectively, suggesting their effectiveness in part-of-speech tagging. Bi-LSTM-CRF demonstrated comparable results, showcasing a F1-score of 95.58%, signifying its capability to capture sequential patterns. In essence, while all models demonstrated high accuracy, SikuRoBERTa and SikuBERT slightly outperformed the others, indicating their aptitude in handling intricate linguistic contexts.

Bi-LSTM-CRF's potential superiority over BERT in ancient Chinese text part-of-speech tagging experiments can be attributed to its specialized design for sequence labeling tasks, proficiency in capturing sequential patterns and

linguistic nuances, adaptability to smaller datasets, robustness in handling noisy annotations, and suitability for domain-specific vocabulary. Additionally, its simpler architecture may confer an advantage in tasks with lower complexity. The choice between Bi-LSTM-CRF and BERT depends on factors like dataset size, linguistic traits, and task requirements. Both models have strengths, but Bi-LSTM-CRF's tailored nature and sequential modeling make it a favorable choice for certain ancient language text segmentation tasks.

In an experiment focused on part-of-speech tagging, the large language model demonstrated remarkable performance. It exhibited a high level of accuracy in predicting parts of speech within sentences. When compared to the Bi-LSTM-CRF and BERT models, the large language model showcased competitive results in terms of tagging accuracy and efficiency. While Bi-LSTM-CRF excelled in capturing intricate sequential patterns, the large language model's extensive pre-training on diverse text corpora granted it a strong understanding of language semantics and context, enabling it to make accurate predictions. However, the large language model's reliance on a massive amount of data might lead to overfitting on domain-specific tasks, which is where models like BERT, designed with a narrower focus, could outperform it. The choice between these models should be based on factors such as dataset size, computational resources, and the level of linguistic nuance required for the particular part-of-speech tagging task.

B. Text Linguistic Features

Based on the aforementioned experimental results, we selected the pre-trained language model that exhibited the most favorable overall performance and applied it to perform full-text model annotation within the constructed corpus of canonical texts. Guided by the annotation outcomes, we conducted a comprehensive analysis and comparative study of the language style features at the lexical level across cross-lingual canonical texts.

1) Lexical part of speech:

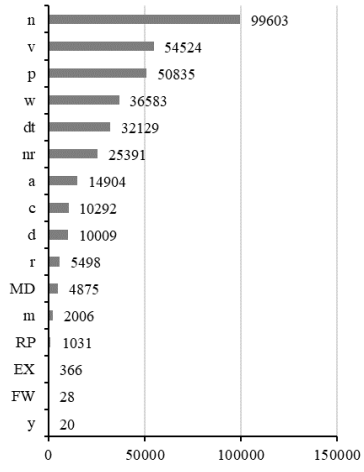


Fig. 2. Frequency of Part-of-Speech Tagging in English Texts

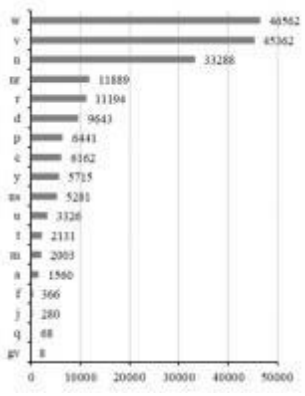


Fig. 3. Frequency of Part-of-Speech Tagging in ancient Chinese Texts

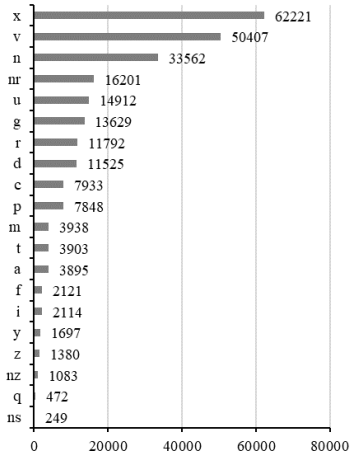


Fig. 4. Frequency of Part-of-Speech Tagging in modern Chinese Texts

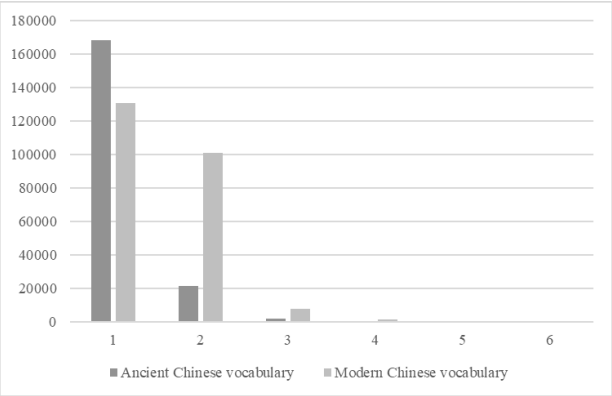
Utilizing the aforementioned statistics, a comprehensive comparative analysis of the parts of speech can be conducted, drawing insights from the provided tabular data. The statistics encompass part-of-speech markers in three diverse linguistic contexts: ancient Chinese, English, and modern Chinese. This analysis aims to illuminate the distinctive linguistic characteristics and tendencies inherent in each language variant, while also highlighting the potential implications for cross-language translation and communication. Examining the data, it is evident that each language exhibits its own distinct distribution of part-of-speech markers. In ancient Chinese, the prominence of certain markers, such as "n" (general noun) and "v" (verb), underscores the foundational elements of its syntactic structure. Meanwhile, in English, markers like "n" (general noun) and "v" (verb) maintain their significance, but variations emerge in the prevalence of other markers such as "a" (adjective) and "p" (preposition), reflecting English's unique grammatical features. Modern Chinese, characterized by its evolving linguistic landscape, reveals distinct patterns with markers such as "n" (general noun) and "v" (verb) attaining a significant presence. The increased occurrence of "x" (a marker denoting anon-specific category) suggests the influence of evolving linguistic trends in modern discourse.

The comparative analysis also offers insights into the role of function words. In modern Chinese, for example, the notable prevalence of "u" (auxiliary) and "d" (adverbial)

markers may underscore the role of these function words in indicating grammatical relationships and nuances.

Furthermore, the comparative analysis of part-of-speech markers has implications for translation. Translators must navigate the challenges posed by the differing syntactic structures and language-specific nuances. The variations observed in the distribution of markers may impact the clarity, accuracy, and stylistic equivalence of translated texts.

2) Lexical word length:



The data showcases a clear inverse relationship between word length and lexical frequency, evident in both ancient and modern Chinese vocabularies. As word length increases, the frequency of occurrence decreases dramatically. This is a well-documented phenomenon in linguistics, commonly referred to as Zipf's Law. Zipf's Law posits that a small number of words are highly frequent, while the majority occur rarely. In the context of this dataset, this law manifests in the low frequency of longer words.

In the realm of ancient Chinese vocabulary, the data reflects a significantly reduced occurrence of longer words. The stark drop in frequency as word length increases aligns with the characteristics of classical Chinese, which was characterized by a concise and contextuallyrich writing style. The prevalence of shorter words highlights the linguistic economy present in classical Chinese texts, where a single character often encapsulated intricate meanings. Conversely, the modern Chinese vocabulary manifests a more balanced distribution across word lengths. While the data still adheres to Zipf's Law, the frequencies of longer words are relatively higher compared to ancient Chinese. This can be attributed to the evolution of the Chinese language over time, influenced by socio-cultural changes and language contact. Modern Chinese accommodates a broader range of vocabulary to express contemporary concepts, leading to a more even distribution acrossword lengths.

From a linguistic perspective, this discrepancy can be attributed to the differences in linguistic structures and writing conventions between ancient and modern Chinese. Classical Chinese often relied on monosyllabic morphemes, resulting in a higher frequency of shorter words. In contrast, modern Chinese, while still rooted in monosyllabic morphemes, has evolved to incorporate polysyllabic and compound words, leading to amore even distribution of word lengths. Moreover, cultural and sociolinguistic factors also come into play. The shift from classical to modern Chinese coincided with significant changes in society, technology, and global interactions. Modernization and internationalization necessitated the adoption of new terminologies, contributing to a richer vocabulary with diverse word lengths.

3) Lexical richness and lexical density:

The Type/Token Ratio (TTR), commonly employed by linguists, serves as a pivotal metric for gauging the lexical diversity inherent within a given text. The token count, representing the total vocabulary within a corpus, is designated as the token, while the type signifies the non-repeated count of forms, encapsulating the total number of distinct word forms. The ratio between the two components inherently reflects the lexical variance exhibited within the corpus. Generally, a higher Type/Token Ratio (TTR) value indicates a greater degree of vocabulary variation and richness within the text. However, the frequent recurrence of non-content words such as Chinese particles "而 (ér), 何 (hé), 乎 (hū), 乃 (nǎi), 之 (zhī)," and English articles "the, of, a" within substantial textual corpora may lead to distortion in TTR application. Within this context, researchers have proposed the Standardized Type/Token Ratio (STTR), which entails computing the type/token ratio within each unit capacity of the corpus and subsequently averaging the results to derive the standardized type/token ratio. Furthermore, correlated studies indicate that the trajectory of the token ratio curve parallels that of the logarithmic curve. Consequently, an augmented computation involving logarithmic smoothing has been introduced to the original formula, enhancing the reliability of the outcomes.

$$\text{TTR} = \frac{Types}{Tokens} \times 100\%$$

$$\log \text{TTR} = \frac{log\ Types}{log\ Tokens} \times 100\%$$

Lexical density, calculated as the ratio of content words to the total number of words, is a frequently employed linguistic metric for gauging the amount of information in a text. Content words, encompassing nouns, content verbs, adjectives, and adverbs, are indicative of words that convey substantive meaning. The lexical density, derived from the ratio of content words to the total vocabulary size, serves as an intuitive indicator of the richness or sparsity of language within a given text. By statistically analyzing the occurrence of characters, function words, and content words in classical Chinese, modern Chinese, and English translations, and employing the metrics of Type-Token Ratio (TTR), log TTR, and lexical density, the distinctive linguistic characteristics of the canonical texts are juxtaposed. The ensuing results are presented in the following table.

Table Head	**Ancient Chinese**	**Modern Chinese**	***English***
Token	191279	250882	348094
Type	11789	11789	18537
Substance word	90210	87864	179040
TTR (Type / Token Ratio)	0.0616	0.0470	0.0532
log TTR	77.09%	75.41%	77.02%
Lexical density	51.93%	45.02%	31.43%

The numerical values in the table represent the overall corpus size of the text, indicating that the corpus size of ancient Chinese texts is relatively small, followed by modern Chinese and then English translations. The figures of different symbols in the text represent the number of distinct vocabulary items, illustrating the highest diversity in vocabulary usage in modern Chinese. The type-token ratio (TTR) of ancient Chinese is the highest, reaching 0.0616, while the TTR of modern Chinese and English translations are comparable. The smoothed type-token ratio (STTR), another metric, shows a relatively minor difference among the three, with modern Chinese having the highest value, followed by ancient Chinese and then English translations having the lowest STTR. The lexical density of ancient Chinese texts is the highest at 51.93%, whereas modern Chinese has a lexical density of 45.029%. English translations exhibit the lowest lexical density, amounting to only 31.43%. The modern Chinese translation supplements the omitted subject and conjunction in the classical Chinese sentence, while the English translation further adds auxiliary words indicating the tense of the action. Accordingly, this reaffirms that classical Chinese vocabulary is richer than modern Chinese and that Chinese vocabulary is more extensive than English. The lexical density also validates the theory of translation simplification in linguistics. From ancient Chinese to modern Chinese and from ancient Chinese to English, as well as from modern Chinese to English translations, translated texts tend to utilize more refined and concise linguistic vocabulary.

V. Discussion

The intricate architecture of LLMs, while affording them the ability to comprehend and generate complex linguistic patterns, also renders them susceptible to dependencies that may hinder their applicability. One such limitation pertains to instruction optimization, a critical aspect that affects the precision and efficiency of analyses conducted using LLMs. Instruction optimization involves the alignment of input instructions or prompts with desired output. LLMs rely heavily on these prompts to generate contextually relevant and accurate responses. However, the challenge arises when intricate or specialized prompts are required to elicit nuanced responses. LLMs may struggle to accurately interpret such instructions, leading to suboptimal outcomes in terms of analysis results. This dependence limitation is of particular concern in scenarios where the analysis demands a fine-grained or domain-specific approach. When the nuances of a particular linguistic or contextual domain are essential for accurate analysis, the reliance on generic prompts provided to LLMs can hinder their effectiveness. The model's output is contingent on the quality and appropriateness of the instructions it receives, which, in turn, affects the reliability of the analysis results.

In light of these considerations, it is noteworthy that our study refrained from utilizing the LLM model for the analysis results. This strategic decision was made to ensure the highest degree of precision and contextual relevance in our analyses. By acknowledging the limitations of LLMs, specifically in terms of instruction optimization, we strive to adopt a comprehensive approach that leverages the strengths of different methodologies while mitigating potential pitfalls. It is important to underscore that the intention is not to disregard the capabilities of LLMs but rather to harness them judiciously in alignment with the unique demands of the analysis at hand. As the field of computational analysis continues to evolve, such considerations regarding the limitations and advantages of different models contribute to a holistic understanding and effective application of advanced technologies.

In the subsequent sections of this paper, we delve into the methodology employed, highlighting the rationale behind our approach and underscoring the importance of a nuanced assessment that takes into account both the potential and constraints of LLMs.

VI. Conclusion

In conclusion, this study elucidates the multifaceted nature of cross-language translation styles within the Chinese context. By harnessing the power of a pre-training model and a corpus-driven methodology, the research underscores the intricate interplay of linguistic elements and contextual dynamics. The selection of translation style profoundly impacts the reception and interpretation of these timeless texts.

The findings hold far-reaching implications for both translation theory and practice. The nuanced analysis of cross-language translation styles contributes to an enriched theoretical framework, enhancing our comprehension of language transfer. Practically, the insights garnered from this investigation offer guidance to translators, enabling them to navigate the intricate terrain of Chinese translation with heightened sensitivity.A judicious amalgamation of fidelity to the original and contextual adaptation is imperative in upholding the enduring relevance of ancient Chinese wisdom in a globalized world. The unveiled translation styles enrich translation theory and provide practical insights for translation practitioners. This study not only advances scholarly discourse but also paves the way for a more profound understanding of the complexities inherent in Chinese translation.

Acknowledgment

This research is supported by the Jiangsu Graduate Practice Innovation Program (SJCX23_0008) and the National Social Science Fund: Research on information science education and development for national strategy(20&ZD332). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

References

[1] Fitzsimmons-Doolan, Shannon. "Language ideologies of institutional language policy: Exploring variability by language policy register." Language policy 18 (2019): 169-189.

[2] Kutlu, Ethan, and Ruth Kircher. "A corpus-assisted discourse study of attitudes toward spanish as a heritage language in Florida." Languages 6.1 (2021): 38.

[3] Zheng, Rong, et al. "A framework for authorship identification of online messages: Writing − style features and classification techniques." Journal of the American society for information science and technology 57.3 (2006): 378-393.

[4] Ji, Li-Jun, Zhiyong Zhang, and Richard E. Nisbett. "Is it culture or is it language? Examination of language effects in cross-cultural research on categorization." Journal of personality and social psychology 87.1 (2004): 57.

[5] Chen, Sylvia Xiaohua, and Michael Harris Bond. "Two languages, two personalities? Examining language effects on the expression of personality in a bilingual context." Personality and Social Psychology Bulletin 36.11 (2010): 1514-1528.

[6] Ong, Kenneth Keng Wee, and Lawrence Jun Zhang. "Metalinguistic filters within the bilingual language faculty: A study of young English-

Chinese bilinguals." Journal of Psycholinguistic Research 39 (2010): 243-272.

[7] Li, Defeng, Chunling Zhang, and Kanglong Liu. "Translation style and ideology: A corpus-assisted analysis of two English translations of Hongloumeng." Literary and linguistic computing 26.2 (2011): 153-166.

[8] Su, Ke. "Translation of metaphorical idioms: A case study of two English versions of Hongloumeng." Babel 67.3 (2021): 332-354.

[9] Duan, Siyu, et al. "Disentangling the cultural evolution of ancient China: a digital humanities perspective." Humanities and Social Sciences Communications 10.1 (2023): 1-15.

[10] Lin, Y. Fictionality as a rhetorical resource in Zuozhuan. Neohelicon 45, 213–228 (2018). https://doi.org/10.1007/s11059-018-0423-3

[11] Ye, Xiao, and Min-hua Dong. "A review on different English versions of an ancient classic of Chinese medicine: Huang Di Nei Jing." Journal of Integrative Medicine 15.1 (2017): 11-18.

[12] Wu, Chunlei, et al. "Generate classical Chinese poems with theme-style from images." Pattern Recognition Letters 149 (2021): 75-82.

[13] Li, Haiying, Arthur C. Graesser, and Zhiqiang Cai. "Comparison of Google translation with human translation." the twenty-seventh international flairs conference. 2014.

[14] Chen, Heng, and Junying Liang. "Chinese word length motif and its evolution." Motifs in Language and Text (2017): 37-64.

[15] Li, Qiang, Ruixue Wu, and Young Ng. "Developing culturally effective strategies for Chinese to English geotourism translation by corpus-based interdisciplinary translation analysis." Geoheritage 14.1 (2022): 6.

[16] Wang, Qing, and Defeng Li. "Looking for translator's fingerprints: a corpus-based study on Chinese translations of Ulysses." Literary and Linguistic Computing 27.1 (2011): 81-93.

[17] Hajmohammadi, Mohammad Sadegh, et al. "Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples." Information sciences 317 (2015): 67-77.K. Elissa, “Title of paper if known,” unpublished.

[18] Lee, Joseph J., Tetyana Bychkovska, and James D. Maxwell. "Breaking the rules? A corpus-based comparison of informal features in L1 and L2 undergraduate student writing." System 80 (2019): 143-153.

[19] Li, Qiang, Ruixue Wu, and Young Ng. "Developing culturally effective strategies for Chinese to English geotourism translation by corpus-based interdisciplinary translation analysis." Geoheritage 14.1 (2022): 6.

[20] Levshina, Natalia. "Corpus-based typology: Applications, challenges and some solutions." Linguistic Typology 26.1 (2022): 129-160.

[21] Cohn-Gordon, Reuben, and Noah Goodman. "Lost in machine translation: A method to reduce meaning loss." arXiv preprint arXiv:1902.09514 (2019).

[22] Liu, Mingtong, et al. "Exploring bilingual parallel corpora for syntactically controllable paraphrase generation." Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. 2021.

[23] Coley, Connor W., William H. Green, and Klavs F. Jensen. "Machine learning in computer-aided synthesis planning." Accounts of chemical research 51.5 (2018): 1281-1289.

[24] Luo, Lin. "A Corpus-based Study of Translation Approaches and Strategies on Culture-loaded Words in Literatures: A Case Study of Jiutu and Its English Version." (2023).

[25] Hu, Kaibao, and Kyung Hye Kim, eds. Corpus-based Translation and Interpreting Studies in Chinese Contexts: Present and Future. Springer Nature, 2019.

[26] Chen, Yaru, and Wei Chen. "English translation of long Traditional Chinese Medicine terms: A corpus-based study." Terminology 24.2 (2018): 181-209.

[27] Alfuraih, Reem F. "The undergraduate learner translator corpus: a new resource for translation studies and computational linguistics." Language Resources and Evaluation 54.3 (2020): 801-830.

[28] Weisser, Martin. Practical corpus linguistics: An introduction to corpus-based language analysis. Vol. 43. John Wiley & Sons, 2016.

[29] Geng, LB, et al. "Research Status and Prospects of Computational Linguistics" Linguistic Sciences 20.5 (2021): 491.

[30] Wang, Yizhong, Sujian Li, and Jingfeng Yang. "Toward fast and accurate neural discourse segmentation." arXiv preprint arXiv:1808.09147 (2018).

[31] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[32] Wang, Huiping, and Luyao Wang. "A Research on the Evolution of Chinese Semantics Based on Distributed Representation." Chinese Lexical Semantics: 21st Workshop, CLSW 2020, Hong Kong, China, May 28–30, 2020, Revised Selected Papers 21. Springer International Publishing, 2021.

[33] Alom, Md Zahangir, et al. "A state-of-the-art survey on deep learning theory and architectures." electronics 8.3 (2019): 292.