Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Value Assessment and Linguistic Feature Mining Based on Regression Analysis in Ancient Literary Texts

Yali Li[1,✉]

[1] *College of Humanities and Arts, Xi'an International University, Xi'an, Shaanxi, 710077, China*

ABSTRACT

The value assessment of ancient literary texts and the mining of linguistic features are indispensable parts of academic research and ancient cultural inheritance. This paper uses the multiple regression model as a quantitative analysis tool for value assessment to evaluate the value of ancient literary texts. At the same time, for the linguistic features of ancient literary texts, we put forward the quantitative descriptive definitions of words, phrases, sentences and other multi-layer and multi-latitude, and establish the corresponding calculation formulas. After the assessment of the value of ancient literary texts, it can be learned that, except for the artistic law and the breadth of dissemination, the ancient literary texts are positively correlated with other influencing factors such as the writing method and the rhythm and rhyme, and the gap between the predicted value of the value assessment and the real value is small, with an error of 40% or less in 90% of the cases. In the mining analysis of linguistic features using The Peony Pavilion and The West Wing as research objects, the average word length of the former is slightly higher than that of the latter, while the difference in the distribution of long and short sentences of the latter is relatively large. Meanwhile, the average dependency distance of The Peony Pavilion is 2.42, which is higher than that of The Story of the Western Wing by 0.1, making syntactic analysis more difficult.

Keywords: Multiple regression model, linguistic features, dependency syntax, ancient literary texts

1. Introduction

Ancient Chinese literary classics are the treasures of Chinese civilization for more than 5,000 years, they not only record the historical changes and cultural development of the Chinese nation, but also contain rich ideological wisdom and artistic value [1-2]. With the change of the times, these classics still have unique modern value and have a profound influence on the culture, education, aesthetics and social concepts of contemporary society. Ancient literary classics have two basic characteristics, historical inheritance and artistry [3-4].

✉ Corresponding author.
E-mail address: xianliyali@163.com (Y. Li).
Received 18 February 2024; Revised 10 April 2024; Accepted 25 December 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-230

Ancient literary classics occupy an important position in the history of Chinese literature and have a profound influence on the development of later literature. These classics not only provide later literary scholars with rich creative materials and sources of inspiration, but also provide them with valuable artistic experience and theoretical support [5-6]. At the same time, the ancient literary classics have also shaped the unique cultural character and aesthetic concepts of the Chinese nation and become an important part of Chinese culture. In terms of ideology, the ideas of Confucianism, Taoism and other schools of thought embedded in ancient literary classics had a profound impact on the governance of ancient society, the handling of interpersonal relationships and the cultivation of personal character [7-8]. In terms of art, ancient literary classics have attracted countless readers and viewers with their unique artistic charm. These works not only show the high talent and excellent skills of ancient literati, but also set an example and benchmark for the development of Chinese literature. At the same time, ancient literary classics have also promoted the exchange and integration of Chinese literature with the literature of other countries, and promoted Chinese literature to the world [9-10].

As an important part of ancient Chinese culture, ancient Chinese literature has a pivotal position in the traditional literary system of the Chinese nation [11]. In-depth study and understanding of ancient Chinese literature, we will find that the creators of many ancient literary works have the ultimate pursuit of the creation of linguistic mood, especially in some lyrical poems and songs, the creators are more willing to contain the feelings in the language, the mood in the scene, the scene blending, giving the language and text of the highly contagious mood and flavor, so that readers in the process of savoring the words feel the unique charm of ancient literature, and at the same time get a pleasurable aesthetic experience. This makes the readers feel the unique charm of ancient literature in the process of savoring the words, and at the same time obtain a pleasant aesthetic experience [12-14].

The characteristics of ancient Chinese literature can be summarized as follows: firstly, ancient Chinese literature has fine words, focuses on the precision of words, has rich and varied sentence patterns, is good at using various rhetorical techniques, and shows strong expressive and infectious power of words [15]. Secondly, ancient Chinese literature has profound ideological connotations, and most of the ancient literary works are related to the creator's thinking and understanding of human nature, society and other issues, with deep and intriguing meanings. Third, ancient Chinese literature attaches great importance to rhythm and rhyme, especially in poetry, paying more attention to rhyme and pursuing the beauty and musicality of tone [16]. Fourth, ancient Chinese literature advocates the way of mediocrity and elegance, and does not pursue ornate embellishments, but more simple and elegant temperament. Fifth, ancient Chinese literature takes historical figures and events as the main material, fully reflecting the ideology and life pursuit of people in the social environment at that time [17].

In this paper, the research focus is concentrated on the value assessment and linguistic features of ancient literary texts. In order to realize the quantitative analysis of the value assessment of ancient literary texts, the multiple regression model is introduced as a quantitative analysis tool. In the construction of the multiple regression model, the least squares method is utilized to determine the parameters of the regression model, and the conventional test methods of determining the goodness-of-fit, regression equation test and regression coefficient test are proposed. With the help of the constructed multiple linear regression model, the index system of factors influencing the value of ancient literary texts is established, and regression analysis is carried out on the sample data of ancient literary texts. In the study of linguistic features of ancient literary texts, quantitative and descriptive statistics are carried out from the linguistic features of words, phrases, sentences, paragraphs and grammar, and relevant formulas are proposed to analyze the linguistic features in terms of the crowdedness of the text, richness of vocabulary, rhythmicity of the text, and brokenness of the text. In the mining analysis of linguistic features, two classic ancient literary works, The Peony Pavilion and

The Story of the Western Wing, are selected as research objects, and their linguistic features and syntactic features are studied in depth.

2. Multiple linear regression model

The regression model is represented as follows.

Based on the observed value sample data is recorded as $(y_i, x_{i1}, x_{i1}, \cdots, x_{ik}), i = 1, 2, \cdots, n$ and introduced into the formula, it is derived:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, i = 1, 2, \cdots n \tag{1}$$

Expand and get:

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_n \end{cases} \tag{2}$$

In this equation, β_0 can be regarded as a constant term, $\beta_1, \beta_2, \cdots, \beta_k$ can be regarded as a regression coefficient term, γ is known as the explained variable, and $X_1, X_2, \cdots, X_k$ is the value of the independent variable that affects the explanatory variable, i.e., the explanatory variable. ε is the random disturbance term, $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ is the mutually independent residual series disturbance term and satisfies $\varepsilon_i \sim N(0, \delta^2), i = 1, 2, \cdots, n$. When $k = 1$, this model is univariate linear regression, while $k \geq 2$, the model is multiple linear regression [18].

From this, it can be seen that:

$$\begin{cases} E(\varepsilon) = 0 \\ \text{var}(\varepsilon) = \sigma^2 \end{cases} \tag{3}$$

Similarly, the multiple linear overall regression equation is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{4}$$

The multiple linear sample regression equation is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k \tag{5}$$

The model can be expressed in matrix form:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{6}$$

$$\left.\begin{array}{c} y = A\beta + e \\ e \sim N(0, \delta^2 I) \end{array}\right\} \tag{7}$$

In the matrix, n is the sample size, k is the number of independent variables, y is the vector of dependent variable observations, and X is the design matrix with X the actual variable values that make up the matrix.

2.1. Model parameter estimation

When determining the parameters of the multiple linear regression model, it can be estimated by the sample data of the observed things, and the least squares method is one of the methods commonly used in predictive analysis [19].

The dependent variable Y has the following relationship with the many independent variables $X_1, X_2, \cdots, X_k$:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k + e \tag{8}$$

The least squares method can be used when determining parameter estimates $b_0, b_1, \cdots, b_k$. The basic idea is to make the residuals sum of squares:

$$\sum e^2 = \sum (Y - b_0 - b_1 X_1 - b_2 X_2 - \cdots - b_k X_k)^2 \tag{9}$$

Minimize. The partial derivatives of the part on the right hand side of the equation are found, and the resulting partial derivatives are made zero, and finally this system of equations is solved. In general using such a method can be expressed in terms of matrices, that is, the coefficient matrix solution method [20].

If the linear relationship between the dependent variable Y and the independent variable is expressed as:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k + e \tag{10}$$

Then:

$$Y = XB + e \tag{11}$$

Among them:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}, e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \tag{12}$$

where Y is the $n \times 1$ matrix, consisting of the individual observations of the dependent variable, B is the $(k+1) \times 1$ matrix, which is the specific parameter vector, e is the $n \times 1$ matrix, which is the residual terms obeying a normal distribution, i.e., $e \sim N(0, \sigma^2 I)$, n is the number of sample data, k is the number of independent variables, and X is the $n \times (k+1)$ matrix, which consists of the individual observations of the independent variables and becomes the design matrix.

The resulting parameter estimates:

$$b = (X'X)^{-1} X'Y \tag{13}$$

where X' is the transpose matrix of matrix X and $(X'X)^{-1}$ is the inverse matrix of matrix (XX).

The standard error is:

$$\hat{\sigma} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - m - 1}} = \sqrt{\frac{\hat{e}^T \hat{e}}{n - m - 1}} \tag{14}$$

The residual sum of squares $e^2 = e'e$, e' is the transpose matrix of e.

And so:

$$\varphi = e'e = (Y - XB)'(Y - XB) \tag{15}$$

Expanded:

$$= Y'Y - Y'XB - B'X'Y + B'X'XB \tag{16}$$

$$= Y'Y - 2B'X'Y + B'X'XB \tag{17}$$

By the Fundamental Theorem $(AB)' = B'A'$ for matrices, we get $Y'XB = (B'X'Y)'$, and $B'XY$ is a matrix of 1×1. $Y'XB$ is also a matrix of 1×1, so YXB is the same value as $B'XY$.

The partial derivative of the sum of squares of the residuals and making it 0 is satisfied:

$$\frac{\partial e'e}{\partial B} = -2X'Y + 2X'XB \tag{18}$$

$$Y'XB = B'X'Y \tag{19}$$

$$B = (X'X)^{-1}X'Y \tag{20}$$

where X' is the transpose matrix of matrix X and $(X'X)^{-1}$ is the inverse matrix of matrix XX.

2.2. Routine testing

1) Measurement of goodness of fit. Goodness of fit is an identification value used to compare the validity of regression effects. The value of goodness of fit also depends on the number of independent variables in the model, the increase of independent variables can not change the total sum of squares, but can increase the regression sum of squares, that is, the denominator did not change, the numerator becomes larger, the value of goodness of fit will become larger. Which is defined as:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\hat{y})^2} \tag{21}$$

where SSR is the sum of squares of regression, SSE is the sum of squares of residuals, and SST is the sum of squares of total deviations.

R^2 increase is not related to the goodness of the model, so when there are more independent variables in the model, it is not very appropriate to use the value of goodness of fit for comparative analysis, and in order to eliminate the degree of dependence of the goodness of fit on the number of independent variables in the model, we use the method of correction.

The modified goodness-of-fit formula is:

$$\overline{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{SSE}{SST}\cdot\frac{n-1}{n-k-1} = 1 - (1-R^2)\frac{n-1}{n-k-1} \tag{22}$$

$$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2 \tag{23}$$

It can be concluded that $\overline{R}^2$ takes into account the average of the sum of squares of the residuals rather than the sum of squares of the residuals itself, and thus, the larger $\overline{R}^2$ is, the better, in linear regression modeling for predictive analysis.

2) Test of regression equation. Let the original hypothesis be $H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0$ and the alternative hypothesis be $H_1 : \beta_i, i = 1,2,\cdots,m$, not all null.

Construct the F statistic:

$$F = \frac{SSR/m}{SSE/n-m-1} = \frac{\sum(\hat{y}_i - \overline{y})^2/m}{\sum(y_i - \hat{y})^2/(n-m-1)} \tag{24}$$

Where $\sum(\hat{y}_i - \overline{y})^2$ is the sum of squares of regression (SSR) which has m degrees of freedom and $\sum(y_i - \hat{y})^2$ is the sum of squares of residuals (SSE) which has $n-m-1$ degrees of freedom [21].

After calculating the F values from the above equation, the model was then tested by checking against the F distribution table. For the confidence interval $\alpha = 0.05$, check the values F_α in the F distribution table with degrees of freedom m and $n-m-1$, and if $F \geq F_\alpha$, it can be obtained that there is a linear correlation between y and $x_1, x_2, \cdots, x_m$, and vice versa, the linear correlation between the two is not obvious.

3) Regression coefficient test. Set the original hypothesis as $H_0 : \beta_i = 0$ and the alternative hypothesis as $H_1 : \beta_i \neq 0, i = 1, 2, \cdots, m$.

Construct the statistic:

$$t_{\beta_i} = \frac{\beta_i}{S_{\beta_i}}, i = 1, 2, \cdots, m \tag{25}$$

where S_{β_i} is the standard deviation of the regression coefficient β_i, calculated as:

$$S_{\beta_i}^2 = S^2 \cdot C_{ii}, i = 0, 1, \cdots, m \tag{26}$$

where S^2 is the regression variance, $S^2 = \frac{\sum e_t^2}{n-k-1}$, n is the number of observations, k is the number of independent variables C_i is the elements on the main diagonal of the matrix $(X^{'}X)^{-1}$, thus:

$$S_{\beta_i} = \sqrt{S^2 \cdot C_{ii}} \tag{27}$$

Based on the given significance level α, a two-sided test for degrees of freedom $n-m-1$ in the distribution table of query t yields the critical value $t_\alpha / 2(n-m-1)$.

If $|t_{\beta_i}| \geq t_{\alpha/2}(n-m-1)$, the original hypothesis H_0 is rejected, the regression coefficient β_i is not significantly different from 0, and the t test of the parameter is passed, indicating that it is feasible to choose the independent variable to explain the dependent variable. On the contrary, the original hypothesis cannot be rejected, indicating that the sample observations do not prove that the regression coefficient is significantly different from 0. It means that the t test of the parameter has not been passed, which indicates that it is not feasible to use the dependent variable to explain the independent variable.

2.3. Multicollinearity and its effects

The so-called multicollinearity means that the observations and the influences are very closely related to each other and are not completely unrelated, there is no independent relationship whether it is a single docking or -to-many linkage there is a linkage, which means that this relationship does exist. It is also valid to express the constant relationship as $b_0, b_1, \ldots, b_k$:

$$b_0 = b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \tag{28}$$

Here for the time being it is determined that a presence x_p can also be shown by other variables as:

$$x_p = \frac{a_0 - \sum_{i \neq p} a_i x_i}{a_p} \tag{29}$$

From the correlation phenomenon, we can conclude that when calculating the correlation coefficient, if the result is 1, it means that there is a complete correlation between the two variables, and when calculating the correlation coefficient, the result is 0, it means that the correlation is not

completely correlated with the existence of a certain amount of change, and if it is in the range between 0 and 1, it can be shown that there is a mutual relationship.

If this property exists for the entire prediction process, then the coefficient $a_0, a_1, \cdots, a_k$ at partially 0 yields that $a_0 = a_1 x_1 + a_2 x_2 + \cdots + a_k x_k$ is valid, and from this we utilize the regression coefficients least squares method of estimation $\hat{\beta} = (X^{'} X)^{-1} X^{'} Y$ cannot exist because $(X^{'} X)^{-1}$ does not exist.

If the whole prediction process exists only approximation of this nature, the nagging existence is part of the coefficient of 0 $a_0, a_1, \cdots, a_k$, get $a_0 = a_1 x_1 + a_2 x_2 + \cdots + a_k x_k$ is also valid, but then the relative value of $| XX | \approx 0$, $(X^{'} X)^{-1}$ becomes very large, $\hat{\beta}$ the relative elements of the variance matrix $D(\hat{\beta}) = \sigma^2 (XX)^{-1}$ also becomes very large, that is to say, the precision of the estimated value may be very low.

In order to deal with the problem of multiple covariance, this paper mainly applies the stepwise elimination method to analyze the independent variables. The basic idea of stepwise regression analysis is to make the complex influence environment simple and relevant, mainly by introducing the independent variables into the establishment of the equation, and analyze the degree of influence of the independent variables by testing the index value, find out the most obvious factors affecting the observation value, and observe the rest of the variables to test the gradual elimination of the independent variables that have no obvious influence on the dependent variable. Finally, the best-fit model must eliminate the insignificant influencing factors and keep the most significant influencing factors, so that the multiple regression model is the most suitable predictive analysis for the observation.

3. Assessment of the Value of Ancient Literature Texts Based on Multiple Regression Analysis

In this section, the value of ancient literary texts will be assessed and analyzed based on the theoretical basis of the multiple linear regression model proposed above.

3.1. Indicator System of Factors Influencing the Value of Ancient Literature Texts

On the basis of the existing research on ancient literary texts, we analyze the influencing factors of the value of ancient literary texts, and screen them according to the purpose of this research and the assessment stage. This paper finally constructs the index system of the factors influencing the value of ancient literary texts from three aspects: text value, theoretical value and social value, and the selection of the indexes is shown in Table 1.

Table 1. The index system of the value of ancient literary text value

Primary indicator	Symbol	Secondary indicator	Symbol	Tertiary index	Symbol
Text value	Z	Subject thought	Z_1	Thought content	Z_{11}
		Text feature	Z_2	Language style	Z_{21}
				Rhetorical device	Z_{22}
				Plot structure	Z_{23}
				Character shaping	Z_{24}
		Affective connotation	Z_3	Emotional experience	Z_{31}
Theoretical value	S	Aesthetic orientation	S_1	Aesthetic guidance	S_{11}
		Creative theory	S_2	Creative method	S_{21}
		Art rule	S_3	Rhythm	S_{22}
				Structural method	S_{23}

				Symbolic intention	S_{24}
Social value	M	Social impact	M_1	Historical background	M_{11}
				Cultural background	M_{12}
				Education action	M_{13}
				Revolution enlightenment	M_{14}
		Propagation value	M_2	Propagation span	M_{21}

3.2. Multiple regression modeling

Combined with the index system of influencing factors on the value of ancient literary texts constructed in the previous section, each influencing factor is introduced into the regression model, and the model is established as follows:

$$\begin{aligned}\ln Y = C &+ \beta_1 \cdot Z_{11} + \beta_2 \cdot Z_{21} + \beta_3 \cdot Z_{22} + \beta_4 \cdot Z_{23} \\ &+ \beta_5 \cdot Z_{24} + \beta_6 \cdot S_{11} + \beta_7 \cdot S_{21} + \beta_8 \cdot S_{22} \\ &+ \beta_9 \cdot S_{23} + \beta_{10} \cdot S_{24} + \beta_{11} \cdot M_{11} + \beta_{12} \cdot M_{12} \\ &+ \beta_{13} \cdot M_{13} + \beta_{14} \cdot M_{14} + \beta_{15} \cdot S_3 + \beta_{16} \cdot M_{21} + \mu\end{aligned} \tag{30}$$

where C is a constant.

Y is the value of ancient literary texts, and $\ln Y$ denotes the natural logarithm of the value; the natural logarithm is taken for Y in order to attenuate heteroskedasticity.

Z_{11}-Ideological content, Z_{21}-Language style, Z_{22}-Rhetorical devices, Z_{23}-Plot structure, Z_{24}-Characterization, Z_{31}-Emotional experience.

S_{11}-Aesthetic Guidance, S_{21}-Creative Methods, S_{22}-Rhythm and Rhyme, S_{23}-Structural Chapters, S_{24}-Symbolic Intentions, S3-Laws of Art.

M_{11}-Historical context, M_{12}-Cultural context, M_{13}-Educational role, M_{14}-Enlightenment to change, M_{21}-Breadth of dissemination.

β_1, β_2, β_3, ..., β_{16} are coefficients and μ is a random error term.

3.3. Multiple regression model coefficient determination

After analyzing the data of 124 ancient literature text samples by multiple linear regression through SPSS software, the following results were obtained as shown in Table 2. Based on the data in the table, the regression results can be obtained as follows.

1) Multiple covariance test. As can be seen from the table, the tolerances are all greater than 0.1 and the VIFs are all less than 10, indicating that there is no multicollinearity between the independent variables.

2) Goodness-of-fit test. The coefficient of determination R^2 of the regression equation is 0.883, and the adjusted R^2 is 0.863, which indicates that the model fits the sample well, and the explanatory variables can explain 86.3% of the changes in the value of the explanatory variables.

3) Test of significance of the equation - F test. The statistical value of F of the overall test of the equation 45.578>F0.01, corresponding to a sig value of 0.000, indicates that the linear relationship between the explanatory variables and the explanatory variables is significant in general, i.e., the overall linear relationship of this model is valid at the 1% level of significance.

4) Parameter significance test. The 12 explanatory variables of creative method, rhythm, structure, symbolic intention, ideological content, linguistic style, plot structure, emotional experience, cultural background, educational role, enlightenment and aesthetic orientation passed the 1% significance level, the 3 explanatory variables of rhetorical techniques, characterization and historical background passed the 5% significance level, and the 2 explanatory variables of artistic rules and breadth of dissemination failed to pass the significance level, i.e. the overall linear relationship between this

model is valid at the 1% significance level. The three explanatory variables of rhetorical techniques, characterization and historical background passed the 5% significance level, while the two variables of artistic rules and communication breadth failed the significance level test.

To summarize, the regression results show that the value of ancient literary texts is positively related to the factors of writing method, rhythm, structure, symbolism, ideological content, linguistic style, plot structure, emotional experience, cultural background, educational function, enlightenment, aesthetic orientation, rhetorical techniques, characterization, and historical background, while there is no obvious linear relationship with the indicators of artistic rules and dissemination.

Table 2. Regression

Interpretation variable	Explained variable $\ln Y$			T	Significance	Common linear statistics	
	Nonnormalized coefficient		Normalization factor			tolerance	VIF
	B	Standard error	Beta				
Constant	-4.365***	0.427	-	-10.242	0	-	-
Creative method	1.646***	0.285	0.281	5.765	0	0.462	2.164
Rhythm	1.838***	0.275	0.293	6.49	0	0.587	1.713
Structural method	2.075***	0.531	0.137	3.928	0	0.849	1.182
Symbolic intention	1.315***	0.213	0.24	6.293	0.002	0.788	1.249
Thought content	0.243***	0.069	0.117	3.255	0.001	0.771	1.319
Language style	0.538***	0.153	0.26	3.436	0.046	0.206	4.803
Rhetorical device	0.295**	0.156	0.144	2.018	0.001	0.192	5.006
Plot structure	0.517***	0.152	0.233	3.488	0.019	0.282	3.575
Character shaping	0.392**	0.168	0.152	2.377	0	0.29	3.489
Emotional experience	0.377***	0.091	0.192	4.19	0.014	0.561	1.76
Historical background	0.233**	0.102	0.108	2.508	0.004	0.574	1.72
Cultural background	0.314***	0.112	0.124	2.961	0	0.651	1.556
Education action	0.506***	0.115	0.171	3.65	0	0.557	1.774
Revolution enlightenment	0.414***	0.112	0.188	4.604	0.546	0.66	1.547
Art rule	0.019*	0.022	0.033	0.609	0.001	0.614	1.646
Aesthetic orientation	0.266***	0.072	0.127	3.428	0.523	0.784	1.26
Propagation span	0.054*	0.089	0.03	0.651		0.806	1.226
R-squared		0.883		F-statistic		45.578***	
Adjusted R-squared		0.863		Prob(F-statistic)		0.000	
* indicates a significant level of 10% * indicates a significant level of 5% * indicates a significant level of of 1%							

3.4. Multiple regression model robustness test

Based on the above regression results, this paper utilizes 10 ancient literary works, quantifies the indicators and brings them into this regression equation for value assessment, and compares the calculated results with the actual value. The data of ancient literary works tested are specifically shown in Table 3.

Table 3. Robustness test

Ancient literature	Thought content	Text feature	Affective connotation	Aesthetic orientation	Creative method	rhythm	Structural method	Symbolic intention	Social impact
The Book Bf Songs	0	1	0	0	0.88	0.88	0.83	0.41	1
Elegies Of Chu	0	1	0	1	0.78	0.36	0.75	0.45	1
The Analects	0	1	0	0	0.72	0.36	0.62	0.52	0
The Lament	0	0	0	0	0.53	0.52	0.68	0.34	0
Record Of the Grand Historian	0	1	0	1	0.81	0.72	0.78	0.81	1
Strange tale from a Chinese Studio	0	1	0	1	0.88	0.88	0.86	0.74	1
The Romance Of the Western Chamber	0	1	0	0	0.42	0.42	0.72	0.82	0
The Peony Pavilion	1	1	1	1	0.68	0.81	0.74	0.81	1
The Scholars	0	1	1	1	0.72	0.72	0.86	0.74	1
Gods And Heros	0	1	0	1	0.67	0.68	0.77	0.75	1

The results of the value assessment were compared with the actual value, and the errors are specifically shown in Table 4. Due to the limited data of ancient literary texts that can be obtained, not all the indicators that may have an impact on the value are covered, so the accuracy of the parameters may be affected to a certain extent. Although the predicted values of value assessment calculated by the multiple regression model are different from the real values, most of the value assessment values have a smaller gap with the real values, and 90% of the errors are within 40%, which shows that the value assessment model is more reliable.

Table 4. Error contrast

Error	Quantity	Proportion
Within 10%	4	40%
10%~20%	3	30%
20%~40%	2	20%
40%~50%	1	10%

4. Linguistic Features of Ancient Literary Texts

In this chapter, we will conduct a multi-latitude, quantitative and descriptive statistical analysis of the multilevel linguistic features of ancient literary texts, such as words, phrases, sentences, paragraphs and syntax. Descriptive statistics refers to the statistical analysis of the overall features of each text, each type of text or the whole corpus in the corpus. On this basis, it can further reveal the linguistic features of weird literary texts in the latitude of text readability, linguistic popularity, linguistic subordination and the sense of rhythm of the work.

4.1. Text Subordination

Before examining the subordination of ancient literary texts, two concepts need to be introduced, text word formation rate and degree of clustering.

1) Text word formation rate. The text word formation rate (WFR) is the ratio of the sum of the lengths of all words (i.e., the total number of words) to the sum of the lengths of all characters (i.e., the total number of characters) in a text. The formula for calculating the word formation rate is as follows:

$$WFR = \frac{\sum length\ word}{\sum length\ character} \tag{31}$$

2) Degree of Clustering and Text Crowdiness. This study adopts the quantitative index of "degree of clustering" to measure the followness of text. The so-called degree of clustering refers to the product of two orthogonal W elements, namely, the word formation rate and the average word length of a text (the word formation rate reflects the breadth of text followness, while the average word length reflects the intensity of text followness), i.e.:

$$C = WFR \times ALW \tag{32}$$

4.2. Vocabulary richness

Lexical richness, which refers to the richness of a language user's vocabulary use in the process of verbal output, reflects the diversity and maturity of the language M. For the examination of lexical richness, this paper proposes the following proposed measures.

1) Lexical Uniqueness. Lexical uniqueness refers to the proportion of words that appear only once in a text to the total vocabulary of the text. The formula is as follows:

$$LI = \frac{\sum number(hapax)}{\sum number(word)} \times 100\% \tag{33}$$

Since lexical uniqueness is more affected by text length, thus, in order to reduce the impact of text length on the uniqueness rate, this paper adopts the quantitative index of logarithmic uniqueness to represent lexical uniqueness, i.e., the numerator and denominator of the above formula are taken as common logarithms respectively, and then the ratio is calculated:

$$LI = \frac{log \sum number(hapax)}{log \sum number(word)} \times 100\% \tag{34}$$

2) Lexical Diversity. Lexical diversity refers to whether the vocabulary used by the language user is rich and varied with less repetition. The higher the value of lexical diversity, the wider the range of words used. On the contrary, the lower the value of lexical diversity, the narrower the scope of words used in the text.

At present, for the quantitative analysis of lexical diversity, the earlier proposed and most typical measure is the type-example ratio mentioned in the previous section, whose formula is:

$$TTR = \frac{Types}{Tokens} \times 100\% \tag{35}$$

Some researchers found that the curve trend of TTR is very similar to the logarithmic curve, so the basic formula of TTR was improved accordingly using logarithmic operations as a way to weaken the effect of text length on TTR. Logarithmic TTR is proposed, i.e., the common logarithms of word types and word examples are calculated separately, and then the ratio of the two is calculated:

$$Herdan' \sin dexC = \frac{\log Types}{\log Tokens} \tag{36}$$

3) Lexical Density. The so-called lexical density is the ratio between lexical words (i.e., real words) and all the words in the text, which is calculated by the following formula:

$$L, D = \frac{\sum number(contentwords)}{\sum number(words)} \times 100\% \tag{37}$$

4.3. Rhythmicity of the text

Text rhythm is an important component of text form and one of the ways of expressing text meaning.

In Chinese text, sentence length dispersion can well reflect the degree of rhythmic changes in the text. Sentence length dispersion refers to the extent to which the length of each sentence in a text deviates from the average sentence length, and the formula is as follows:

$$D_s = \sqrt{\frac{1}{n} \sum (L_i - \overline{L}_s)^2} \quad (i = 1, 2, 3, ..., N) \tag{38}$$

4.4. Text fragmentation

The so-called brokenness refers to the number of pauses in a sentence, which can reflect the discrete degree of the sentence. Generally speaking, the stronger the written text, the fewer the pauses in the sentence, the lower the degree of fragmentation, and the statement will be more fluent, while the stronger the spoken text, the more insertions and other components in the sentence, the more pauses, the higher the degree of fragmentation. The formula for calculating the degree of fragmentation is given below:

$$\text{Sentence fragmentation} = \frac{\text{Sentence pauses}}{\text{Total sentences}} \tag{39}$$

4.5. Syntactic Dependencies

The study of dependent syntactic features is an important school of research on linguistic features of ancient literary texts. In terms of syntactic features of ancient literary texts, this paper mainly analyzes them statistically in terms of dependency distance.

In this paper, the absolute value of dependency distance is used to calculate the average dependency distance of a sentence or a dependency treebank, for example, the formula used to calculate the average dependency distance of a sentence with more than n words is:

$$\frac{1}{n-1}\sum_{n-1}^{i=1}|dd_i| \tag{40}$$

This formula can be used to calculate the average dependency distance of the whole dependency treebank, if a dependency treebank has n sentences and m dependencies, the average dependency distance of this dependency treebank is:

$$\frac{1}{m-n}\sum_{m-n}^{i=1}|dd_i| \tag{41}$$

Using the above formulas, the average dependency distance of this dependency treebank can be obtained. The dd_i in these i formulas refers to the second dependency in a sentence or a dependency treebank.

5. Analysis of linguistic features mining of ancient literary texts

In this chapter, the two major ancient literary works of The Story of the Western Wing and The Peony Pavilion will be selected as the research objects, combined with the linguistic features of the ancient literary texts proposed above, and the linguistic features mining analysis will be carried out around the texts of the two to explore the linguistic features characteristics of the ancient literary texts.

5.1. Statistical analysis of word and phrase feature descriptions

1) Average word length analysis. Generally speaking, a sentence can be regarded as a sequence of words. The longer the words used in a sentence, the less readable the sentence will be. The distribution of the average word length of the two ancient literary works “The West Wing” and “The Peony Pavilion” is specifically shown in Fig. 1. From the figure, it can be concluded that the distributions of the average word lengths of the two ancient literary works are normally distributed. However, the average word length of The Peony Pavilion is slightly higher than that of The Story of the Western Wing, and the average word length of The Peony Pavilion is more concentrated than that of The Story of the Western Wing, so the shape of the “peak” is more pointed. In addition, the average word length of The Peony Pavilion is also higher than that of The West Wing, which to a certain extent shows that The West Wing has a high degree of colloquialization and is easier to read than The Peony Pavilion.

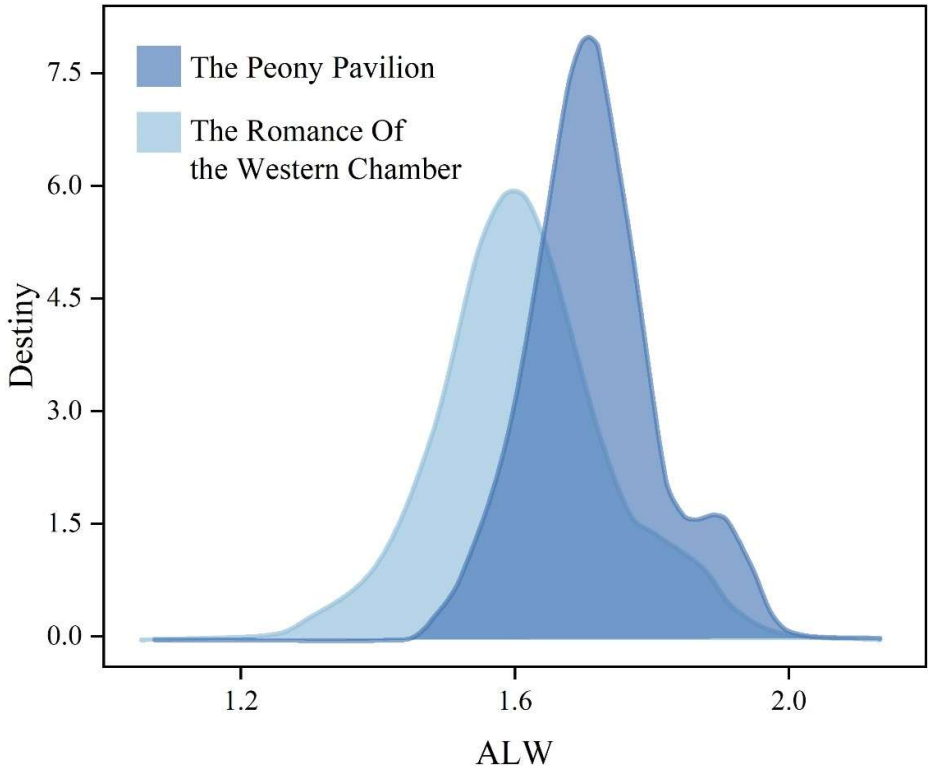


Fig. 1. The average term distribution

2) Analysis of sentence discrete degree. The distribution of sentence discrete degree of the two ancient literary works "The West Chamber" and "The Peony Pavilion" is shown in Figure 2. As can be seen from the figure, the distribution of sentence discrete degree in The Peony Pavilion is more symmetrical, and most of them are concentrated around 20. On the other hand, "The West Wing" is right-skewed, while the distribution of values is not centralized and more scattered, and its average dispersion is on the large side. This shows that compared with The Peony Pavilion, the distribution of long and short sentences in the body of The West Wing is more different.

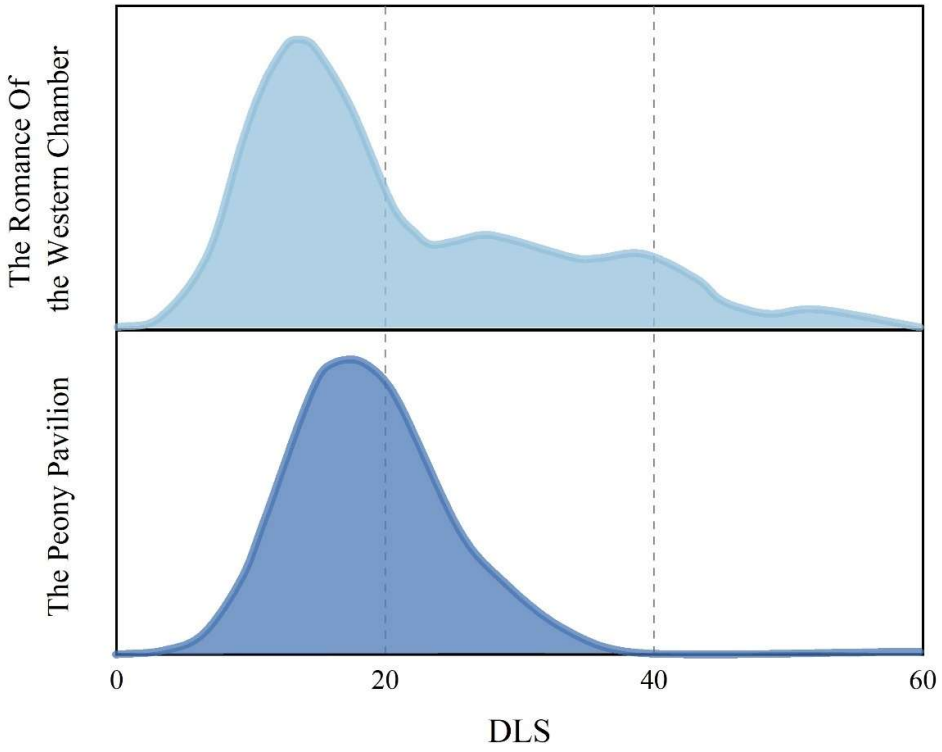


Fig. 2. The discrete distribution of the sentence

5.2. Statistical analysis of syntactic features

In this section, based on the Dependency Treebanks of two ancient literary works, The Peony Pavilion and The Story of the Western Wing, features such as dependency distance, dependency direction and type of dependency relationship are comparatively studied.

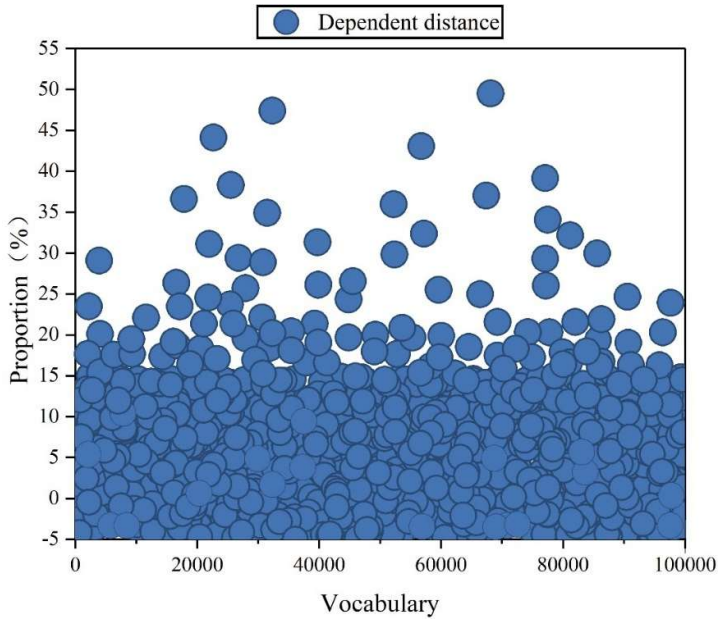
The above syntactic dependency feature formula is used to calculate the average dependency distance of the dependency treebanks of The Peony Pavilion and The Story of the Western Wing, as shown in Table 5. From the table, it can be seen that the average dependency distance of the dependency treebank of The Peony Pavilion is 2.42, and that of The Story of the Western Wing is 2.32. In comparison, the average dependency distance of the dependency treebank of The Story of the Western Wing is shorter. As far as the dependent syntactic analysis is concerned, the dependency distance can reflect the difficulty of sentence comprehension to a certain extent, i.e., the larger the dependency distance is, the more difficult the syntactic analysis is. Thus, it can be seen that the sentence comprehension difficulty of The Peony Pavilion is slightly more difficult than the sentence comprehension difficulty of The West Wing novel.

Table 5. Average dependent distance

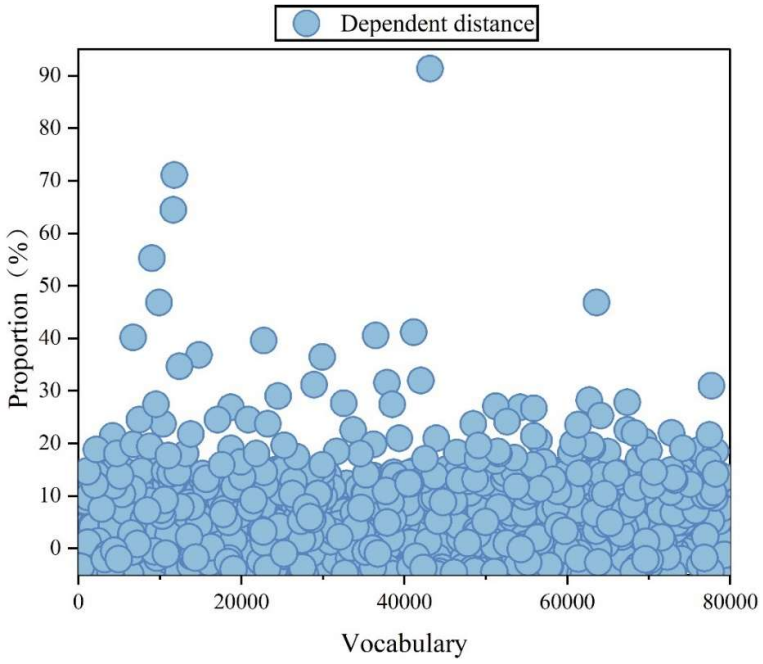
-	The Romance Of the Western Chamber	The Peony Pavilion
Average dependent distance	2.32	2.42

Through the time series graph of all the dependencies contained in the treebank, we can see more clearly the dense dependencies between neighboring words and the minimized distribution of dependency distances characteristic of natural language. The X-axis of the dependency time series graph indicates the number of words in the whole dependency treebank, and the Y-axis indicates the dependency distance of the dependency treebank. The dependency time-series diagrams of The Peony Pavilion and The West Wing are specifically shown in Fig. 3, with diagrams (a) and (b) corresponding to The Peony and The West Wing in that order. As can be seen from the figure, the dependency distances of the two dependency treebanks show the same distribution characteristics. The similarity

lies in the fact that there are both positive and negative dependency distances in the two dependency treebanks, and the distribution of positive dependency distances is more than that of negative dependency distances. The dependency distances of the two dependency treebanks are densely distributed in the period from -5 to 15 on the vertical axis, while the distribution of dependencies with a dependency distance of more than 15 is relatively sparse. It can be seen that the average dependency distances in both dependency treebanks have a tendency to minimize the dependency distances.



(a)The Peony Pavilion



(b)The Romance Of the Western Chamber

Fig. 3. The sequence of dependencies

In addition, on the basis of the two dependency treebanks, the distribution of absolute dependency distances was counted, as shown in Figure 4. The absolute dependency distances of the two

dependency treebanks show the overall distribution characteristic that the larger the dependency distance is, the more the coverage rate is. In the two dependency treebanks, the dependency relationship formed between neighboring words, i.e., the dependency relationship with the dependency distance of 1 accounts for more than 55% of the whole dependency relationship, when the absolute dependency distance of the two treebanks reaches 3, the two dependency treebanks cover more than 80% of the dependency relationship; when the absolute dependency distance reaches 5, the two dependency treebanks cover more than 90% of the dependency relationship. When the absolute dependency distance reaches $\geqslant$10, the dependencies of the two dependency treebanks are completely covered.

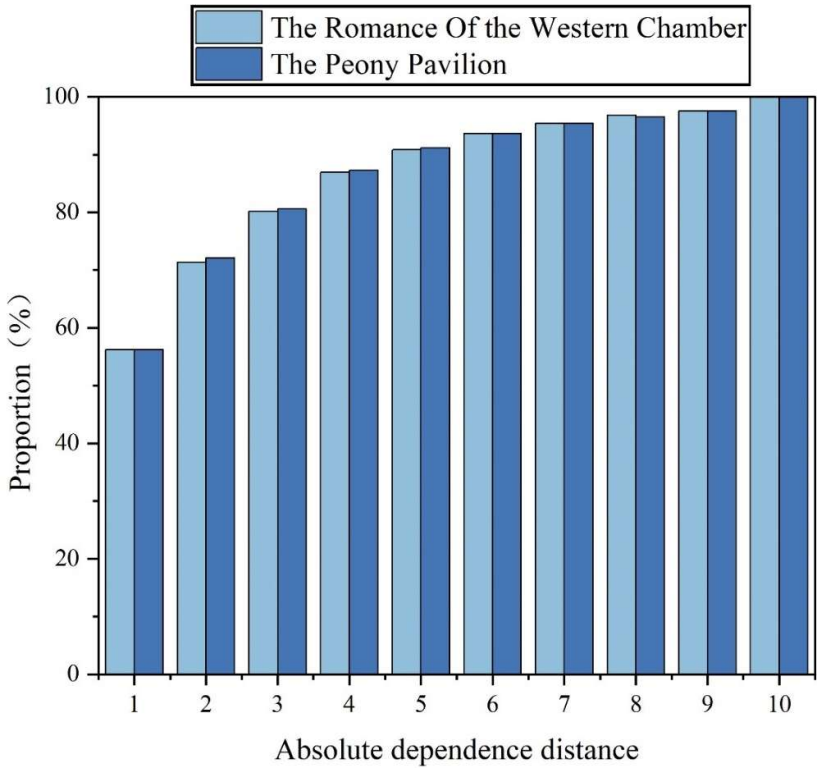


Fig. 4. Absolute dependence distance

6. Conclusion

The research subject of this paper is ancient literary texts, and the research content mainly focuses on the two aspects of value assessment and linguistic features.

For the research of value assessment of ancient literary texts, the multiple regression linear regression model is introduced as a quantitative analysis tool, and the index system of factors influencing the value of ancient literary texts is built. According to the regression results, except for the value of ancient literary texts, which has no obvious linear relationship with the two indexes of artistic rules and dissemination breadth, there is a significant positive correlation between writing method, rhythm and rhyme, structure and chapter, symbolic intention, ideological content, linguistic style, plot structure, emotional experience, cultural background, educational role, enlightenment, aesthetic orientation, rhetorical techniques, characterization, historical background and other factors. . Taking 10 ancient literary works as samples to carry out value assessment, the resulting value assessment predictions differ from the true value, but the difference gap is small, and the percentage of literary works with error within 40% is 90%, which verifies the reliability of the value assessment method proposed in this paper.

Facing the study of linguistic features of ancient literary texts, a multi-latitude quantitative

descriptive statistical analysis is carried out in text readability, linguistic popularity, linguistic subordination and the sense of rhythm of the work, syntactic dependence, and so on. Two classic ancient literary works, The Peony Pavilion and The Story of the Western Wing, are chosen as the research objects, and their lexical and syntactic features are studied and analyzed in depth. The average word length of The Peony Pavilion is slightly higher than that of The Story of the Western Wing, and the average word length is more concentrated, compared with The Story of the Western Wing, which has a high degree of colloquialization and is more readable. The distribution of sentence discrete degree in The Peony Pavilion is more symmetrical, mainly concentrated around 20, while the distribution of value in The Story of the Western Wing is not concentrated, and the average discrete degree is on the large side, and there is a big difference in the degree of discrete degree between the two. In syntactic features, the average dependency distances of the dependency treebanks of The Peony Pavilion and The Story of the Western Wing are 2.42 and 2.32 respectively, and the sentence comprehension difficulty of The Peony Pavilion is slightly greater than that of the novel The Story of the Western Wing. The time-series plots of the two dependencies are all densely distributed between -5 and 15 on the vertical axis, while the distribution of dependencies with a dependency distance of 15 or more is relatively sparse, and both have a tendency to minimize the average dependency distance. When The Peony Pavilion and The Story of the Western Wing have absolute dependency distances up to $\geqslant$10, the dependencies of the dependency treebanks are both fully covered.

References

[1] Wang, X. Y. (2023). The Legacy of Ancient Cultures: Rational Concepts in Ancient Chinese and Ancient Greek Mythology and Their Significance in Modern Literature. Eurasian Journal of Applied Linguistics, 9(1), 110-122.

[2] Sheng, X. (2017). Cultural and value differences of goddess in Ancient Greece and China. Eur Sci J, 13(10), 102-114.

[3] Pollard, D. E. (2021). A Chinese look at literature: the literary values of Chou Tso-jen in relation to the tradition. University of California Press.

[4] Crone, T. (2023). The Scribal Witness: Narrative Authority in Ancient Chinese Literature. Early China, 46, 265-285.

[5] Chen, L., Chen, L., & Li. (2017). The core values of Chinese civilization. New York, NY: Springer.

[6] Yan, L. (2024, September). Research and Reflection on the utilization of ancient Chinese medicine literature resources. In 2024 10th International Conference on Humanities and Social Science Research (ICHSSR 2024) (pp. 75-81). Atlantis Press.

[7] Chen, L., & Yuan, M. (2022). Analysis on the cultivation of college students'cognitive psychology of chinese excellent traditional culture by chinese ancient literature. Psychiatria Danubina, 34, S1170-S1175.

[8] Yu, P. (2018). Poems in Their Place: Collections and Canons in Early Chinese Literature. In Critical Readings on Tang China (pp. 936-966). Brill.

[9] Qiao, Z. (2022). On the practice of inheriting excellent culture in ancient chinese literature from the perspective of educational psychology. Psychiatria Danubina, 34(suppl 1), 723-725.

[10] Teng, C., & Guo, W. (2018, December). Research on Ancient Chinese Literature History from Time Dimension to Spatial Dimension. In 2nd International Conference on Art Studies: Science, Experience, Education (ICASSEE 2018) (pp. 265-267). Atlantis Press.

[11] Zhai, L. (2021). Ancient Literature and Art Based on Big Data. In 2020 International Conference on Data Processing Techniques and Applications for Cyber-Physical Systems: DPTA 2020 (pp. 357-365). Springer Singapore.

[12] Guliayimu, Y. (2022). ANALYSIS ON THE GOAL CONNOTATION AND LEVEL ORIENTATION OF ANCIENT LITERATURE TEACHING FOR ANXIETY COLLEGE STUDENTS. Psychiatria Danubina, 34(suppl 1), 278-280.

[13] Du, Z. (2023). Research on the integration of traditional culture into the teaching of ancient literature courses in colleges and universities in the context of big data. Applied Mathematics and Nonlinear Sciences.

[14] XiaoYu, W. THE HUMANISTIC SPIRIT CONTAINED IN ANCIENT CHINESE AND ANCIENT GREEK MYTHOLOGY. Turkic World Between East and West.

[15] Wu, Y. (2023). Research on Text Value and Linguistic Characteristics in Ancient Literature Based on Text Mining Technology. Applied Mathematics and Nonlinear Sciences, 9(1).

[16] Cao, W. (2020, November). An Ethical Analysis of Imagery Translation in Ancient Chinese Books and Records. In 2020 Conference on Education, Language and Inter-cultural Communication (ELIC 2020) (pp. 437-440). Atlantis Press.

[17] Ying, L. H. (2021). Historical dictionary of modern Chinese literature. Rowman & Littlefield.

[18] Yongcun Shao & Cong Qin. (2024). Strategies for constructing mathematical models of nonlinear systems based on multiple linear regression models. Applied Mathematics and Nonlinear Sciences(1).

[19] Chillarón Pérez Mónica,Vidal Vicente E.,Verdú Gumersindo J. & Quintana Ortí Gregorio. (2024). Few-View CT Image Reconstruction via Least-Squares Methods: Assessment and Optimization. Nuclear Science and Engineering(2),193-206.

[20] Nandakumaran A. K.,Sufian Abu & Thazhathethil Renjith. (2024). Homogenization of Semi-linear Optimal Control Problems on Oscillating Domains with Matrix Coefficients. Applied Mathematics & Optimization(2).

[21] Korkmaz Mehmet. (2020). A study over the general formula of regression sum of squares in multiple linear regression. Numerical Methods for Partial Differential Equations(1),406-421.