Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5548–5556
Marseille, 20-25 June 2022

CLGC: A Corpus for Chinese Literary Grace Evaluation

Yi Li, Dong Yu✉ , Pengyuan Liu
School of Information Science, Beijing Language and Culture University
Beijing, China
liyi_blcu@163.com, yudong_blcu@126.com, liupengyuan@pku.edu.cn

Abstract

In this paper, we construct a Chinese literary grace corpus, CLGC, with 10,000 texts and more than 1.85 million tokens. Multi-level annotations are provided for each text in our corpus, including literary grace level, sentence category, and figure-of-speech type. Based on the corpus, we dig deep into the correlation between fine-grained features (semantic information, part-of-speech and figure-of-speech, etc.) and literary grace level. We also propose a new Literary Grace Evaluation (LGE) task, which aims at making a comprehensive assessment of the literary grace level according to the text. In the end, we build some classification models with machine learning algorithms (such as SVM, TextCNN) to prove the effectiveness of our features and corpus for LGE. The results of our preliminary classification experiments have achieved 79.71% on the weighted average F1-score.

Keywords: corpus, annotation, Chinese literary grace, linguistic features

1. Introduction

Literary grace (文采) reveals the aesthetic function of language and matters of the use by a writer of the language (Wang, 1994; Rastall, 2008). Texts with better literary grace can bring readers a higher aesthetic experience. For instance, when reading the two example sentences in Table 1, we can obviously feel the higher aesthetic values of the first sentence. Literary grace is embodied in phonetics, form, and semantic, including beauties of sound, modification, color, emotion, image, and philosophy. It is regarded as one of the elements constituting the style and quality of articles in China (Wang, 1994; Zhou, 2013; Qi, 2006). The discussion about it has been throughout all the ancient Chinese literary theories. As an indispensable factor of the text, literary grace also has attracted scholars in the natural language processing (NLP) field. Many studies have referred to several aspects of it and been applied in real-world applications, including Automated Essay Evaluation (AEE) (Liu et al., 2016a), quality assessment (Qiao et al., 2021), machine translation, opinion mining, dialogue modeling, and modeling argumentative discourse (Tong et al., 2021).

However, there have been few publicly available corpora annotated with literary grace. The traditional related literature and linguistics study is based on the scattered examples or limited to a specific text, not forming a large-scale and integrated corpus. The existing task-oriented datasets are only focused on the sub aspects of literary grace: figure-of-speech (Birke and Sarkar, 2006; Liu et al., 2018) and rhetoric level (Shi, 2019; Gong, 2016). Both of them are ways to get aesthetic feelings, not equivalent to literary grace. More recently, Fu et al. (2018) constructed a sentence-level dataset that is closely related to literary grace. However, the generation of literary grace is also affected by many factors above the sentence level, such as the article's theme, textual rhetoric, and artistic conception formed by the semantic combination of sentences. Therefore, the linguistic units above the sentence level should be considered during the data collection. Moreover, these datasets are often small-scale and directly oriented to a specific field. It means the existing datasets cannot meet the current research on literary grace. Constructing a corresponding larger-scale, multi-level, high-quality, and adaptable corpus is a matter of concern.

In this work, we constructed a Chinese corpus with literary grace annotation, CLGC, to alleviate the problem of scarce corpora. The texts in CLGC were divided into three classes according to their literary grace level. The language of segments labeled as "1" is plain and less rhetoric. Higher-level labels represent higher literary grace. Table 1 shows sentences selected from level 1 and level 3 texts[1]. This corpus consists of 10,000 texts from novels and prose, including 64% level 1 texts (6,448 texts, 1,192,144 tokens), 19% level 2 texts (1959 texts, 369,946 tokens) and 15% level 3 texts (1593 texts, 294,748 tokens). Compared with the existing corpora mentioned in Sec.2, our corpus has the following features: (1) Focus on the real sense of literary grace, not just limited to the specific features; (2) Provide multi-level annotations that include literary grace level, sentence category, and figure-of-speech type tags; (3) Use a simple majority voting scheme to select the most probably literary grace label as the ground truth; (4) Contain paragraph-level materials varied in the domain. Based on this corpus, we can get rid of the shackles of empiricism and explore literary grace more systematically and scientifically. At the same time, we can also explore relevant linguistic fea-

✉ Corresponding author

[1]Each text contains 200 to 300 tokens, and we provide multi-level annotations for every sentence in the text.

Literary grace level	Sentence	Sentence Category	P	Figure-of-speech Count	Figure-of-speech Type
3	如一个娇媚的女子，面带微笑、静静地守在那里，等待着有缘人穿越万水千山寻梦在这里。(As a charming woman with a smile, stay there silently, waiting for a nice ring, who treks through thousands of rivers and mountains here to fulfill his exquisite dream.)	C	0	1	['BY']
	...				
1	我对汽车喇叭声较敏感，是因为上班的办公楼临近马路，经常被机动车喇叭声打扰，有时刚理好的工作思路，立刻就被喇叭声打断。(I am sensitive to the sound of car horns because my office building is near the road, and I am often disturbed by the sound of car horns. Sometimes, I just got my thoughts together, and then the horn interrupted me.)	C	0	0	[]
	...(The rest of the sentences in the text are omitted here.)				

Table 1: Sentences selected from level 1 text and level 3 text ("C" means declarative sentence; "BY" means bi yu, including metaphor and simile; "P" means paragraph-level rhetoric)

tures and observe their performance at different levels of language, so that find new language rules. The corpus and features can also help machine learning models evaluate the literary grace of texts and be applied in similar NLP tasks, like Graded Reading and Automated Essay Scoring. Our contributions include:

- We provide a relatively large-scale Chinese corpus with literary grace annotation, containing multi-level labels. The corpus can be used to facilitate the study of literary grace in literature, linguistics, and natural language processing.
- We analyze the corpus and explore linguistic features that affect readers' aesthetic judgments from three aspects: word, sentence, and paragraph levels.
- We propose a new task, named Literary Grace Evaluation, which aims to make a comprehensive assessment of the literary grace level. Then we classify the corpus with machine learning models. The result shows that the features we chose are very effective in the literary grace evaluation task.

2. Related Works

Datasets Related to Rhetoric Most of rhetoric-related datasets are just about figure-of-speech, especially metaphor, such as VU Amsterdam corpus (Steen, 2010), MOH-X (Saif et al., 2016), TroFi (Birke and Sarkar, 2006) and Chinese-Simile-Recognition (Liu et al., 2018). Chang et al. (2004) and Wang (2020) did different works. They regarded rhetoric as expression skills, not just the figure of speech. They constructed the rhetoric level corpora collected from students' essays. In these corpora, every essay was scored by 2-3 teachers, and the average score was the essay's final score. The former corpus contains 693 essays, and the later corpus contains 366 essays.

Datasets Related to Elegant Sentences The public essay websites, such as JuKu[2] and LeLeKeTang[3], are important corpus resources. These websites collected essays from different grades, and each essay was commented on and marked by professional teachers. These marked sentences can be considered elegant sentences, the others as normal sentences. Fu et al. (2018), Chen (2021), Gong (2016) all collected elegant sentences from them. Fu et al. (2018) also manually annotated 21,053 sentences with two tags (elegant and non-elegant) by two annotators back-to-back, including 3990 elegant sentences and 17,063 normal sentences. Qiao et al. (2021) annotated a batch of blogs in the same way.

Tasks and Applications Based on the rhetoric datasets, Chang et al. (2004) and Wang (2020) proposed a text rhetoric evaluation task and analyzed the possible linguistic features related to the rhetoric level, including token numbers, adjectives, idioms, non-verbal metaphors, sentence pattern, and figure-of-speech. Fu et al. (2018) proposed a task of elegant sentence recognition in Chinese essays of high school students. They presented a deep neural network combining Convolution Neural Network (CNN) and Bi-directional Long Short-Term Memory (BiLSTM) to recognize grace sentences (up to 89.23% classification accuracy) and applied it to the AEE task. The experiment proved that the elegant sentence feature could improve the performance of AEE. Chen (2021), Gong (2016) did the same work to assist automatic essay scoring. Qiao et al. (2021) combined elegant sentence recognition with the AMR model for blogs quality evaluation and proved this method can improve the evaluation performance (from 80% to 85.85%).

3. The Corpus Construction

3.1. Definition of Literary Grace

Literary grace as a result of aesthetic judgment is both prescriptive and empirical (Rastall, 2008). We can se-

[2] http://www.pigai.org/

[3] http://www.leleketang.com/zuowen/

lect the most probable literary grace label using a majority voting scheme. As a basis of our annotation effort, we begin with Liu Xie's theory (Zhou, 2013), which lay the foundation of the views on Chinese literary grace. In this theory, a text with high literary grace should have the following characteristics: (1) Beauty of color: the picture painted by words is colorful and vivid, never dull; (2) Beauty of sound: the rhythm, tone, and fluctuation of language and characters can bring readers a catchy and tuneful sense; (3) Beauty of modification: the text language is rich and varied, mainly referring to sentence pattern and rhetoric; (4) Beauty of emotion: the text has complex and sincere emotion and resonates with readers; (5) Beauty of image: the scenery, character, and image depicted by language can bring the aesthetic feelings; (6) Beauty of philosophy: the text has a profound theme and contains the truth of life. These features are directly related to the language itself. From the aspect of phonetics, we can emphasize the beauty of sound; From the aspect of form, we can emphasize the beauty of modification; From the aspect of meaning, we can emphasize the beauty of color, emotion, image, and philosophy (Qi, 2006).

These six forms of literary grace have been generally recognized in China, and most of the related research since then has been carried out around these six aspects (Xu, 2013). Based on this theory, the text which stands out in ***just one*** of sound, modification, color, emotion, image, and philosophy six aspects can be considered as having high literary grace.

3.2. Data collection

Data Source Our corpus only focuses on the literary style[4], including prose documents and novels. We selected and crawled 600 Chinese prose documents (200 prose documents for each type) and 300 Chinese novels (Original Medium-Short Novels): sanwenwang.com[5] and readnovel.com[6]. The first website consists of different types of Chinese prose, including lyric, narrative, and philosophical prose. Readnovel.com is a very popular novel reading website in China, available in multiple types of novels. These low-threshold original websites can provide us with diverse and popular texts. To prevent adding excessive personal experience during the annotation, we avoided the works of famous authors when selecting the texts, and all these texts have removed authors' names. All the HTML tags were eliminated during this step, together with style sheets, objects, figures, etc.

Data Processing Traditional work measured the literary grace at the whole essay or sentence level. As is mentioned in Sec.1, there are many linguistic features above the sentence level. Thus the sentence is not an appropriate unit to analyze literary grace. On the other hand, ordinary expression generally takes up a higher proportion than elegant expression in texts. So the whole text is too macro to analyze the literary grace, and it also takes too much effort to read the whole text during annotation. To get an idea of the right length, we first annotated 200 texts (100 for novels, 100 for prose documents), and each text was annotated by two annotators, majored in literature. In this task, annotators can freely extract what they think is the best part of the text based on the principle mentioned in Sec.3.1. At the end of annotation, we selected the segments two annotators choose consistently[7] to calculate the average length of the segments. In these 200 texts, there were 332 segments the two annotators extracted, 204 of them mainly consistent. The average length is 212 tokens. Thus, we first split the raw data into 250-word segments, then proofread them one by one, deleting semantically incomplete sentences. Finally, we collected a total of 10,500 segments, the length of each text between 150 and 250 tokens.

3.3. Annotation Guidelines

The annotation task is three-fold, including literary grace, figure-of-speech, and sentence category annotation.

For literary grace annotation, we made "0" and "1" two tags. Annotators should read all the segments of a task (50 segments per task) firstly, then choose one of the following tags for each segment:

1. **1: expresses this is a literary grace segment.** As is mentioned in Sec.3.1, the segment which stands out in ***just one*** of sound, modification, color, emotion, image, and philosophy these six aspects can be considered as high literary grace.

2. **0: expresses this is an ordinary segment.** The segments which do not meet the above requirements.

For sentence category and figure-of-speech type annotation, we used LTP[8] (Language Technology Platform) to split all the segments into sentences firstly. Annotators should determine the sentence category, figure-of-speech type, and figure-of-speech counts during this annotation. We refer to the theory of Huang and Liao (1997) and Chen (2001) for the definition of sentence category and figure-of-speech and ensure that annotators are familiar with them through annotation training. Each figure-of-speech and sentence category is replaced by the first letter in Chinese pinyin. The following shows the annotation guidelines for this task:

[4]Different linguistic style has different specific manifestations of literary grace. For example, the technical text emphasizes simplicity and rigor, while the literature text emphasizes vividness and figurativeness (Qi, 2006). So the literary study must be carried out in linguistic styles

[5]https://www.sanwenwang.com

[6]https://www.readnovel.com

[7]The segments were chosen freely, so the consistent segment can be slightly different in the length.

[8]https://github.com/HIT-SCIR/ltp

		Tag	Examples
Sentence Category	Declarative	C	马克吐温是一位著名的美国作家。(Mark Twain is a famous American writer.)
	Interrogative	Y	你收到他的来信了吗？(Have you heard from him?)
	Exclamatory	G	那该有多好啊！(How wonderful it would be!)
	Imperative	Q	大家快过来呀！(Come on!)
Figure-of-speech Type	Simile& Metaphor	BY	她笑得像花儿一样灿烂。(She smiled as brightly as flowers.) 这是花的海洋。(This is a sea of flowers.)
	Personification& Skeuomorphism	BN	小鸟在唱歌。(The bird is singing.) 敌人夹着尾巴跑了。(The enemy ran with his tail between his tails.)
	Repetition	FF	沉默呵，沉默呵!(Silence, silence!)
	Parallelism	PB	夜是寂静的，是温和的，是梦幻的。(The night is quiet, gentle, and dreamy.)
	Contrast	DO	天有多高，山有多高。(The high of the sky is the high of the mountain.)
	Transferred Epithet	TG	你笑得很甜。(Your smile is sweet.)
	Allusion	YY	失败乃成功之母，你不要放弃。(Failure is the mother of success. Don't give up.)
	Paragraph-level	P	沉寂！沉寂！几亿年的沉寂！(Silence! Silence! The permanent silence!)

Table 2: The annotation tags and examples

1. Read the sentence and annotate its category. You have four tags to choose: "C" (declarative sentence), "Y" (interrogative sentence), "G" (exclamatory sentence), "Q" (imperative sentence). The examples for each sentence category tag are shown in the Table 2.

2. Reread the sentence and determine what kind of figures-of-speech it contains. You have eight tags to choose: "BY" (bi yu, Simile and Metaphor), "BN" (bi ni, Personification and Skeuomorphism), "FF" (fan fu, Repetition), "PB" (pai bi, Parallelism), "DO" (dui ou, Contrast), "TG" (tong gan, Transferred Epithet), "YY" (yin yong, Allusion), "P" (paragraph-level rhetoric). The examples for each figure-of-speech type tag are shown in Table 2

3. Count the number of the figure-of-speech and link each of them by "_".

4. Some special situations are considered as follows:

   - If there is more than one figure-of-speech type in one sentence, please annotate all the types.
   - Use the tag "P" to annotate figure-of-speech, which crosses the sentence. This tag should be marked between the sentence-type and figure-of-speech number tags. For example, the last sentence in Table 2 should be annotated with "_G_P_1_FF" for each sentence.

5. The whole pattern of this annotation is "sentence _ sentence category _ (P) _figure-of-speech numbers _ the first type of figures-of-speech _ the second type of figures-of-speech _ the third...".

3.4. Annotation Process

For the literary grace annotation, we hired 25 annotators, literary majors and divided them into five groups. Each group annotated 2,000 segments, and five annotators annotated each segment. The segments were divided into 200 questionnaires on wenjuan.com[9]. We also set limits on reading time, at least one minute of every segment. Through this approach, each item was annotated with five judgments.

For the figure-of-speech and sentence category annotation, we hired 15 annotators, linguistic major, and divided them into five groups. Each group annotated 2,000 segments, and each segment was annotated by two annotators, proofread by one. They used the commenting function in the NotePad for annotation firstly. Then they consorted together and agreed on a final version. In the case of further inconsistencies, the third annotator determined its final label.

We also performed annotation training and trial annotations for both tasks to ensure every annotator is familiar with the guidelines. In trial annotations, we sampled 50 segments out of the raw data. Every annotator was asked to annotate them following the guidelines and checked the full results once again when they finished. Meanwhile, we revised the annotation guidelines where they were vague.

3.5. Inter-Annotator Agreement

Table 3 shows the inter-annotators' reliability coefficients results of different annotation tags calculated by Intraclass correlation coefficient (ICC) (Shrout and Fleiss, 1979). ICC can reveal the reliability coefficients of two or more nominal data. It is generally believed that the value below 0.4 indicates poor reliability, and above 0.75 indicates good reliability. The global ICC values among sentence category, figure-of-speech count, and paragraph-level figure-of-speech type tags are over 0.95, which shows that each annotator is in near-perfect agreement concerning these tags. For figure-of-speech type tags, all of them show better consistency except "FF" (fan fu, Repetition) and "TG" (tong gan, Transferred Epithet). The sentence labeled with "FF" must satisfy two requirements: (1) repeatedly uses the same sentence or word; (2) expresses strong feelings. However, the judgment on the later re-

[9] https://www.wenjuan.com/

		Group1	Group2	Group3	Group4	Group5
Sentence Category tags		0.981	0.972	0.986	0.991	0.989
Figure-of-speech count tags		0.999	0.965	0.979	0.983	0.993
Over-sentence figure-of-speech type tags		0.981	0.921	0.945	0.942	0.911
Figure-of-speech type tags	BY (Simile and Metaphor)	0.799	0.762	0.743	0.831	0.764
	BN (Personification and Skeuomorphism)	0.782	0.736	0.729	0.731	0.754
	FF (Repetition)	0.454	0.465	0.449	0.436	0.482
	PB (Parallelism)	0.723	0.784	0.783	0.728	0.751
	DO (Contrast)	0.645	0.661	0.646	0.693	0.704
	TG (Transferred Epithet))	0.434	0.462	0.445	0.493	0.557
	YY (Allusion)	0.786	0.794	0.813	0.821	0.745
	P (Paragraph-level rhetoric)	0.655	0.791	0.749	0.762	0.701
		Group6	**Group7**	**Group8**	**Group9**	**Group10**
Literary grace tags		0.477	0.483	0.459	0.487	0.499
Literary grace tags of filter segments		0.848	0.766	0.735	0.858	0.863

Table 3: The inter-annotators' agreements of different tags

quirement varies from person to person[10]. As for "TG", the boundary of "BY"(bi yu, Simile and Metaphor) and "TG" is vague, so annotators easily confuse them. These are the reasons for their low consistency, and we will discuss this phenomenon further in follow-up studies.

When it comes to literary grace tags, the agreements range from 0.45 - 0.5, which is relatively low. It is normal to have different aesthetic judgments, so this result cannot be used to judge quality. We made **filter segment** to help select corpus. The filter segments are non-corpus segments with correct tags, which were taken to identify whether annotators were paying attention to the task or behaving as outliers. The tag to the filter segments was fully agreed upon by the other 20 people majoring in linguistic who did not become a participant in the following annotation. During the annotation process, each task was presented in blocks of 48 target segments and two filter segments (one literary grace segment and one ordinary segment). Suppose annotators did not annotate these segments correctly. In that case, they should annotate all the tasks again until all the filter segments were annotated correctly. The ICC values (first-time annotation) among literary grace tags of these filter segments range from 0.73-0.87, which means the judgment of these annotators is reliable. Figure 1 shows the distribution of the vote related to the literary grace tag. Based on the simple majority voting scheme, we divided these segments into three levels:

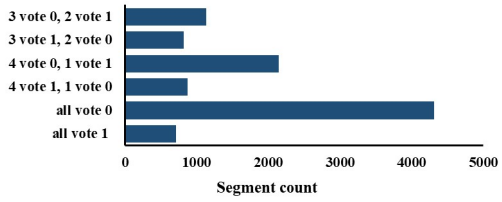

Figure 1: The distribution of the vote related to the literary grace tag

- Level 1: Ordinary segments. Level 1 means that at least four annotators (5 annotators in total) consider the text as having no literary grace;
- Level 2: Transition segments between level1 and level3. Level 2 means that 2 or 3 annotators make a different judgment;
- Level 3: High literary grace segments. Level 3 means that at least four annotators (5 annotators in total) consider the text as having literary grace.

The basic statistics of our corpus are shown in Table 4.

Level	Type	Count	Character	Word	Sent
1	prose	4115	759692	601500	26505
	novel	2333	432452	355869	18772
2	prose	1370	255446	201753	8685
	novel	589	114500	92194	4695
3	prose	1513	279714	220769	9588
	novel	80	15034	11903	558

Table 4: The basic statistics of CLGC

4. Analysis

We further analyze the corpus from three aspects to find out linguistic features that affect Chinese readers' aesthetic judgment. In this section, we only focus on the segments with a majority vote (level 1 and level 3 segments) in the corpus. There are 8041 segments with a majority vote in the corpus. The sentence splitting, tokenization, POS tagging were done with LTP (Che, Wanxiang and Feng, Yunlong and Qin, Libo and Liu, Ting, 2020).

[10]For example, 我的残忍让我痛哭流泪，我的自作自受让我悔恨不已。(My cruelty made me cry, and my self-inflicted made me remorse.)

4.1. Word Level Features

To observe the relationship between words and literary grace more intuitively, we constructed a formula by using word frequency difference as follows:

$$W = (P_{high} - P_{low})/P_{total}$$

$P_{high}, P_{low}, P_{total}$ represent the frequency of the word in level 3, level 1 and whole corpus respectively. The value of W reveals the relationship between word and literary grace. Closer to 1 means this word tends to be in more literary texts. On the contrary, closer to -1 means this word tends to appear in plain texts. With ± 0.5 as the boundary, we selected words with $W > 0.5$ as high literary grace words, words with $W < -0.5$ as low literary grace words. A Spearman's Rank Correlation test using the r statistic was calculated to test the relationship between the frequency of these words in the segments and the literary grace level of segments. There is a high correlation shown in Table 5 ($p < 0.01$), which means the literary grace-related words exactly exist.

	High Literary Grace	Low Literary Grace
Number	14015	32493
r	.644**	.658**
Example	时光 time, 梦 dream, 花 flower, 岁月 years, 风 wind, 生命 life, 树 tree, 美丽 beauty, 心 heart, 阳光 sunshine	他 he, 他们 they, 找 find, 什么 what, 问 ask, 现在 now, 拿 take, 事 thing, 家 home, 可是 but

Table 5: The words related to literary grace (Examples are selected from TOP 20 that are easily translated into English)

Lexical Semantic Shown in Table 5, we can find the semantic differences among them: (1) The words related to low literary grace are mainly about the necessities of life, such as food, clothing, house, and transportation; (2) The words related to high literary grace are more artistic and mainly about the nature and feelings. To verify this deduction, we calculated the Spearman correlation coefficient (r) between the word frequency of different semantic fields and literary grace level by using HIT-CIR Tongyici Cilin (Extended) (Che, Wanxiang and Feng, Yunlong and Qin, Libo and Liu, Ting, 2020). Table 6 shows the semantic fields which $r > 0.2^{**}$. This proves that the words related to machinery, meteorology, natural objects, plant, color, and condition are easier to bring readers high aesthetic experience, e.g.花 (flower), 岁月 (years), 风 (wind). In comparison, the words related to the general term, relatives, society, social contact, address, and interjection are harder to bring this experience, e.g.他们 (they), 警察局 (police station), 妈妈 (mom).

Part-of-speech We also made a Spearman's Rank Correlation test to find out the relationship between the part-of-speech and literary grace level. Overall, the coefficient correlation is low and not significant. However, there is a certain correlation between descriptive words and literary grace. Descriptive words are used to describe colors, states, and modalities of people or objects, e.g.雪白 (white), 滚烫 (hot), 毛茸茸 (fluffy). They are related to the appearance and shape of people and objects, which can give readers a vivid feeling. For example, 这有一朵白花&这有一朵雪白雪白的花 (the two sentences share the same meaning "Here is a white flower."), the only difference is that the latter uses the descriptive word, making it more vivid. The average count of descriptive words in the level 3 texts is 1.300, more than twice the level 2 texts (0.630) and more than four times the level 3 texts (0.326). The correlation coefficient between them and literary grace level is 0.329^{**} ($p < 0.01$). It proves that descriptive words are more commonly used in high literary grace texts.

Semantic field	r	Semantic field	r
Nature	.225**	Meteorology	.373**
Plant	.202**	Machinery	.230**
Color	.209**	Condition	.203**
Relatives	-.251**	General Term	-.210**
Society	-.224**	Social Contact	-.301**
Status	-.262**	Vocative Expression	-.220**

Table 6: The Spearman correlation coefficient ($r > 0.2^{**}$) between different semantic field and literary grace level

Type-Token Ratio Type-token ratio (TTR), also known as vocabulary size divided by text length (V/N), is a simple measure of lexical diversity[11]. The closer the TTR ratio is to 1, the greater the lexical richness of the segment is. The average TTR value based on character and word of level 3 text is 0.660 and 0.735, which is slightly higher than level 1 text (0.622 and 0.712). The Spearman rank correlation coefficient between the TTR (character), TTR(word) and literary grace level are 0.200^{**} ($p < 0.01$) and 0.113^{**} ($p < 0.01$) respectively. It shows that the lexical diversity of high literary grace text is richer than low literary grace text.

4.2. Sentence Level Features

Sentence Category Table 7 shows the basic statistics of sentence category. Declarative sentences (76.93% of total) account for the most significant proportion. Next are interrogative (15.17%) and exclamatory sentences (7.04%). The last is imperative sentences (0.84%). The distribution of these is in line with our daily language usage. Just like normal sentences, declarative and interrogative sentences can fulfill most of our language usage needs in daily life and construct the framework of our discourse world. In comparison, the specific context limits the usage of exclamatory and imperative sentences. We also calculated the Spearman rank correlation coefficient. The result shows a very weak and

[11] It is affected by the length of the text sample, but the text length is mainly same in our corpus.

Figure-of-speech Type									
	BY	BN	FF	PB	DO	TG	YY	P	Overall
Total 1	1570	747	389	923	418	131	889	537	5683
Total 2	1219	688	205	632	516	108	518	436	4123
Total 3	2014	2457	185	907	1143	306	552	538	7687
Overall	4803	3892	779	2462	2077	545	1959	1511	4903
AVG 1	0.24	0.11	0.06	0.14	0.06	0.02	0.13	0.08	0.88
AVG 2	0.62	0.35	0.10	0.32	0.26	0.05	0.26	0.22	2.10
AVG 3	1.36	1.54	0.11	0.56	0.71	0.19	0.34	0.33	4.82
r	.379**	.440**	.059**	.238**	.312**	.175**	.127**	.129**	.535**

Table 7: The basic statistics of figure-of-speech tags, the meanings of each tag in the first line are: bi yu (Simile and Metaphor); bi ni (Personification and Skeuomorphism); fan fu (Repetition); pai bi (Parallelism); dui ou, 对偶 (Contrast); tong gan (Transferred Epithet); yin yong (yin yong, Allusion); paragraph-level rhetoric

negligible correlation between the declarative, imperative sentences and literary grace. However, when it comes to exclamatory ($r = -0.110^{**}, p < 0.01$) and interrogative sentences ($r = -0.149^{**}, p < 0.01$), the coefficient shows a negative correlation, which means that the use of these two categories can lead to lower literary grace.

	C	Y	Q	G	All
Total 1	35074	7680	3768	424	46946
Total 2	11011	1867	859	139	13876
Total 3	8865	1289	404	41	10599
AVG 1	5.43	1.19	0.58	0.06	7.28
AVG 2	5.62	0.95	0.43	0.07	7.08
AVG 3	5.56	0.80	0.25	0.02	6.65
Overall	54950	10836	5031	604	71421
r	.038**	-.149**	-.110**	-.035**	-.059**

Table 8: Sentence category count and correlation coefficient between them and literary grace level ("C": declarative; "Y": interrogative; "Q": exclamatory; "G": imperative)

Figure-of-Speech Type Table 7 shows the basic statistics of the figure-of-speech types. BY (Simile and Metaphor) and BN (Personification and Skeuomorphism) are the most common types of figure-of-speech, 27.45% and 22.24% of the total respectively. TG (Transferred Epithet) and FF (Repetition) are the two least used types, 4.45% and 3.11% of the total respectively. As expected, the high literary grace (level 3) text contains more count of figure-of-speech, about four times more than the low literary grace text (level 1). It mainly reflected in the use of BY, BN, PB (Parallelism) and DO (Contrast). The Spearman rank correlation coefficient between the average count of them and literary grace level is shown in Table 7. The figure-of-speech number is highly correlated with literary grace level ($r = 0.535^{**}, p < 0.01$). As for the different types of figure-of-speech, BN is the most correlated with the literary grace level ($r = 0.440^{**}, p < 0.01$), then is BY ($r = 0.379^{**}, p < 0.01$), PB ($r = 0.238^{**}, p < 0.01$) and DO ($r = 0.312^{**}, p < 0.01$). The last is TG ($r = 0.175^{**}, p < 0.01$).

LDA Topics	
High	(Life): love, dream, time, heart, life... (Travel): go, love, water, lonely, world... (Memory): life, memory, time, think, whether... (Nature): wind, dream, rain, flower, sky... (Emotion): love, heart, happiness, like, miss...
Low	(Everyday Life): think, know, mom, go, eat... (Society): community, do, lantern, time, sound... (Common Verb 1): know, do, think, eat, say... (Common Verb 2): feel, know, thing, watch, laugh... (Social): grandmother, robot, police, find, may...

Table 9: Five topics retrieved from the corpus of high literary grace and low literary grace

4.3. Paragraph Level Features

Figure-of-Speech Type As is shown in Table 7, the paragraph-level rhetoric has little impact on the literary grace level ($r = 0.129^{**}, p < 0.01$). However, in terms of the average number, the changing trend of paragraph-level figure-of-speech in texts with different literary grace levels is the same as that of other figures-of-speech, which proves that figure-of-speech is an essential feature reflecting the changes in the literary grace of the text.

Text Topic We use LDA (Blei et al., 2003) (Latent Dirichlet Allocation) to analyze the topic difference between texts of different literary grace levels. The output for a run of the program for ten topics of LDA is presented in Table 9. Each topic consists of 10 different words. These topics overlap to some degree. So we merge the similar topics into two main topics as follows:

High Literary Grace: *Romantic topic, includes life, travel, memory, nature, emotion, etc.*

Low Literary Grace:[12] *Daily life topic, includes everyday life, society, common verb, social, etc.*

This result is similar to the features of lexical-semantic.

[12]Note that literary grace is not equal to quality. The meaning here is that the topics are often referred to in a low literary grace text, not the nature of the topic itself.

Method		P	R	F1
Majority		47.10	63.85	51.80
TextRNN		60.07	75.05	66.72
TextCNN		77.07	79.38	75.58
SVM	word-level(semantic, part-of-speech and ttr)	77.71	80.75	77.81
	sentence-level(figure-of-speech)	65.13	72.45	64.46
	paragraph-level (paragraph-level rhetoric, topic words)	76.66	71.35	62.86
	word-level + sentence-level	79.56	81.85	**79.71**
	word-level + paragraph-level	79.24	81.70	79.26
	sentence-level + paragraph-level	67.81	72.55	64.51
	word-level + sentence-level + paragraph-level feature	79.08	81.55	79.17

Table 10: The weighted average precision, recall, f1-score of Literary Grace Evaluation with SVM, TextCNN and TextRNN

It is worth noting that topics or words related to high literary grace are often abstract or distant from our daily lives. This phenomenon is in line with the defamiliarization theory in literature (Crawford, 1984; Shklovsky and Berlina, 2017), which holds that the language of poetry is fundamentally different from the everyday language. Because the former is surprising, unusual, strange, and far away from everyday life. These give a unique atmosphere to ordinary things, and it just so happens that people like to be moved by the unusual.

4.4. Experiments

This section proposes a new task, Literary Grace Evaluation (LGE). Unlike tasks aiming to recognize literary grace from the perspective of metaphor or other rhetoric means, we attempt to construct a comprehensive assessment task of literary grace. LGE can be regarded as a task of text classification. We trained classification models (SVM, TextCNN, TextRNN) on the corpus to prove the effectiveness of the features we chose and the applicability of our corpus.

To test the effect of the features mentioned above on the LGE task, we use SVMs (Joachims, 1998), which can efficiently perform classification on a small corpus, as the classifier to classify the corpus automatically. Our experiment is based on 8,000 training data, 1,000 test data, and 1,000 validation data. The results are presented in Table 10. We added features at different levels separately for literary grace level prediction with the same training, test, and validation data. Compared with the performance based on the majority, the performance can be improved no matter which level of features is added. As for the single feature, using word-level features can improve weighted average F1 by 26.01%, then are sentence-level features (12.66%) and paragraph-level features (11.06%). The combination of features at different levels can further improve the model performance. Based on the word and sentence level features, the weighted average F1 can achieve 79.71%, even better than the performance based on all the features (79.17%). The results prove that all the features are effective for the LGE task. They can improve the model performance to varying degrees. Word-level features have the most significant impact among them. Sentence and paragraph-level features have relatively little influence. However, they also play an essential role in literary grace assessment, further illustrating the necessity of evaluating literary grace in language units above the sentence level.

We also experiment with an end-to-end approach, using the basic models, TextCNN (Kim, 2014) and TextRNN (Liu et al., 2016b), to classify the corpus. We did not use any features in the experiments and only based on the rough methods. The results are shown in Table 10. Compared with the results of SVMs based on the majority, the deep learning models show a better performance. TextRNN model can achieve 66.72% on the weighted average F1, and the TextCNN model improves to 75.58%, proving that our corpus has good applicability. The scores are lower than SVM models based on lexical-related features. It shows that some features require manual selection, and there is still much room for the performance of deep learning models.

5. Conclusion and Outlooks

In this paper, we construct a Chinese corpus with the literary grace level tag, sentence category tag, and figure-of-speech type tag. From our annotation experiments, we observed a correlation between word semantic information, part-of-speech, TTR, figure-of-speech, texts topic, and literary grace level. Besides, we propose a Literary Grace Evaluation task and trained classification models on the corpus to prove the availability of these features and the applicability of our corpus. Our corpus can be available at `https://github.com/blcunlp/CLGC/`. In future research, we will extend the corpus to different linguistic styles and further explore linguistic features related to the LGE task.

6. Acknowledgements

This work is funded by the Humanity and Social Science Youth foundation of Ministry of Education (19YJCZH230) and the Fundamental Research Funds for the Central Universities in BLCU (No. 21PT04).

7. Bibliographical References

Birke, J. and Sarkar, A. (2006). A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Chang, Y.-M., Lee, C.-H., et al. (2004). *Automated Chinese Essay Scoring System Rhetoric Aspect*. Ph.D. thesis, Natinal Chiao Tung University.

Chen, W. (2001). *Xiu ci xue fa fan* 修辞学发凡 *Rhetoric*. Xiu ci xue fa fan (xin san ban) Rhetoric (New version3).

Chen, Z. (2021). Design and implementation of chinese composition intelligent eveluation system based on deep learning. Master's thesis, University of Chinese Academy of Sciences.

Crawford, L. (1984). Viktor shklovskij: Différance in defamiliarization. *Comparative Literature*, 36(3):209–219.

Fu, R., WANG, D., Wang, S., Hu, G., and Liu, T. (2018). Elegant sentence recognition for automated essay scoring. *Journal of Chinese Information Processing*, 32(6):88–97.

Gong, J. (2016). The design and implement of the rhetoric recogntion system oriented chinese essay review. Master's thesis, Harbin Instityte of Technology.

Huang, B. and Liao, X. (1997). *Xian dai han yu* 现代汉语 *Modern Chinese*. Xian dai han yu Modern Chinese.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

Liu, M., Qin, B., and Liu, T. (2016a). Automated chinese composition scoring based on the literary feature. *Intelligent Computer and Applications*, 1.

Liu, P., Qiu, X., and Huang, X. (2016b). Recurrent neural network for text classification with multi-task learning. *AAAI Press*.

Liu, L., Hu, X., Song, W., Fu, R., Liu, T., and Hu, G. (2018). Neural multitask learning for simile recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1543–1553.

Qi, Z. (2006). 文章 · 情性 · 文采——《文心雕龙》的文学概念及情感美、形式美思想 article · emotion · literary grace - the literary concept of wen xin diao long and the thoughts of emotional beauty and formal beauty. *A Multidimensional Study of Orientalism*, (1):185–202.

Qiao, Y., Gao, Y., and Ma, N. (2021). Research on blog quality evaluation based on amr and graceful sentence recognition. *Scientific and Technological Innovation*, pages 72–73.

Rastall, P. (2008). Aesthetic responses and the 'cloudiness' of language: is there an aesthetic function of language? *La linguistique*, 44(1):103–132.

Saif, M., Ekaterina, S., and Peter, T. (2016). Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33.

Shi, Y. (2019). Research on the automated classification evaluation method of composition of primary school based on rhetoric use. Master's thesis, Central China Normal University.

Shklovsky, V. and Berlina, A. (2017). *Viktor Shklovsky: A Reader*. Bloomsbury Academic.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Steen, G. (2010). *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

Tong, X., Shutova, E., and Lewis, M. (2021). Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Wang, B. (1994). 文采生成浅论 a brief discussion on literary grace. *Seeking Truth*, 4(22):102–105.

Wang, Q. (2020). Research on rhetoric automatic judgment of primary school essays. Master's thesis, Nanjing Normal University.

Xu, X. (2013). "文采"研究述略 a brief introduction to the study of "wen cai". *Culture Journal*, (5):12.

Zhou, Z. (2013). *Wen xin diao long jin yi* 文心雕龙今译. Wen Xin Diao Long Jin Yi 文心雕龙今译.

8. Language Resource References

Che, Wanxiang and Feng, Yunlong and Qin, Libo and Liu, Ting. (2020). *N-LTP: A Open-source Neural Chinese Language Technology Platform with Pretrained Models*. `https://github.com/HIT-SCIR/ltp`.