A Corpus-Driven Study of the Style Variation in The Grapes of Wrath

Yiyang Hu[1], Qingshun He[1*]

[1] Sun Yat-sen University
[*] Corresponding author's email: heqsh5@mail.sysu.edu.cn

DOI: https://doi.org/10.53482/2022_52_396

ABSTRACT

The novel *The Grapes of Wrath* is distinctive in the arrangement of intercalary chapters and narrative chapters. Existing studies of the narratological distinction of this novel are primarily qualitative. This article conducted a corpus-driven study of the variation of styles in this novel from the perspectives of word cluster, type-token ratio, descriptivity and activity, keyness, and sentiment. The cluster analysis shows that the choice of words in the narrative chapters is more consistent than that in the intercalary chapters. The type-token ratio analysis testifies to the heterogeneity of the intercalary chapters in terms of lexical richness. The descriptivity and activity analysis and the keyness analysis reveal that the narrative chapters are more active than the intercalary chapters. The sentiment analysis finds that the novel is pervaded by negative sentiments and that negative sentiments are more prevalent in the narrative chapters than in the intercalary chapters. The research concludes that the corpus-driven study can provide insights into the narrative structure and the stylistic variation of the novel.

Keywords: corpus-driven; narrative structure; narratological distinction; style variation; The Grapes of Wrath

1 Introduction

1.1 Narrative Structure of The Grapes of Wrath

The Grapes of Wrath written by John Steinbeck is an epic capturing the plights of the Okies whose lives are destroyed by the Dust Bowl and the Great Depression. This novel focuses on the Joad family, who got evicted from the farm in Oklahoma, travelled along Route 66 to California and became migrant farmers in California. By consulting the spatial movement, this novel can be divided into three balanced parts.

In each of the three parts, there are some chapters not directly relevant to the storyline of the Joad family. These chapters are referred to as the intercalary chapters (Swensen, 2015) or the interchapters (Levant, 2007); they provide the narrative chapters with an epic scope. Burcar (2018) regarded this narrative method as a dialectical montage that juxtaposes the seemingly disparate scenes to reveal the structural interconnectedness in a larger context. This setting leads the readers’ attention to the cause of the systematic exploitation. Hamilton (2016) stressed the palimpsest effects of different types of chapters.

However, existing studies mainly focus on the non-verbal aspects of the narrative without considering the linguistic features. Without the description of the language, the interpretation of the narrative structure will be subjective and less convincing. The dichotomy of the intercalary chapters and the narrative chapters can be a framework for the study of style variation.

1.2 Narratology and Stylistics

Stylistic studies can provide narrative analyses with valuable insights. Most narratological studies are characterized by the analysis of text fragments and the exclusion of language, while stylists tend to focus on the linguistic features below the sentence level, such as the rhyme pattern, the word choice and the syntactic structure. It seems to exist a boundary between narratology and stylistics. Rimmon-Kenan (1989) pointed out the paradox that narratologists paid little attention to the linguistic features of the text while narratology theories were greatly influenced by linguistic theories. According to Rimmon-Kenan (1989), the reconstruction of the represented events in literary works should depend on language for its very essence. The language of the text can be described in terms of style. Stylistics is a method of describing linguistic features in the text, which provides evidence for or against the interpretation and evaluation of literature (Short, 1984). Literary criticism without the basis of stylistics lacks a convincing argument. According to Phelan (1981, p. 6), the sense of style can be developed into “those elements of a sentence or passage that would be lost in paraphrase”. After paraphrasing the story, the same narrative technique can be adopted, but the linguistic features such as the dictation and the syntactic structure will be changed.

The loss of style in the paraphrase shows that a narratological study without stylistic analysis cannot fully figure out the art of the author’s presentation. In the analysis of *A Farewell to Arms*, Phelan (1996) consulted the style of paratactic sentences and the use of adverbials to figure out the tension of narrative voices. He derived the hidden narrative voices from the stylistic features, thus building a bridge between narratological techniques and stylistic features. The exploration of this kind of interaction can also be found in the analysis of *Marabou Stork Nightmares* conducted by Short (1999), according to whom the viewpoint shift in the narrative structure and the movement of narrative levels are influenced by the style variation in the text. The choice of language has considerable impacts on the way texts are structured. To get a broader view of how *The Grapes of Wrath* is presented, linguistic features should be considered with organizational techniques.

1.3 Corpus Stylistics

The corpus in stylistics has gained popularity with the development of natural language processing. The relationship between linguistic description and literary appreciation can be defined as the corpus stylistic cycle (Mahlberg, 2014). In this cycle, the researcher can use the corpus method to investigate the linguistic phenomena in quantitative ways, thus giving innovative linguistic descriptions and supporting literary appreciation. The corpus method features the quick collection and calculation of repetitions. This rough indication of the high-frequency elements can be helpful in stylistic studies.

According to Halliday (1971), the foregrounding elements can be found not only in literary deviances, i.e. the use of ungrammatical forms, but also in deflections, i.e. the departure from some expected patterns of frequency. Deviances can be seen as the law-breaking elements while deflections as law-making elements. It should be noted that these repetitions are not always foregrounding elements of stylistic relevance. Halliday (1971) emphasized the positive virtues of counting deflections under the condition that they were relevant to the thematic meaning of the text. Leech and Short (2007) maintained that there existed a dividing line between the foregrounding and the unmotivated prominences. Not all statistical deviances can be regarded as literary foregrounding. It can be seen that the statistical significance is not always relevant to the interpretive significance, but the empirical evidence can make it easier for critics to find clues of foregrounding. To improve the reliability of statistical deviance of linguistic features, researchers of corpus stylistics usually adopt the approach of matching texts against corpus. This approach is influenced by the concept of semantic prosody, referring to "a consistent aura of meaning with which a form is imbued by its collocation" (Louw, 1993, p. 157). The matching process can be achieved by comparing the extract from the individual text with a general corpus to explain the creative use of words. The comparison usually focuses more on the function of concordances rather than on the frequency of words. In the stylistic study of O'Connor's *Eyes*, Hardy (2004) investigated the frequencies of keywords in the context in comparison with the general corpus. This approach can give us evidence of the interference of negative words in the results of sentiment analysis. According to McIntyre and Walker (2019), matching texts against corpora can also be used in the calculation of keyness. The quantitative information of words can help assess whether the choice of the word can be regarded as a departure from the expected choice.

1.4 Hypothesis

The Grapes of Wrath is exquisitely weaved by the alternation of intercalary chapters and narrative chapters. In this article, we will conduct a corpus-driven study of the style variation of *The Grapes of Wrath* especially the relationship between the two types of chapters to reflect upon the role of language as a medium in the construction of the story. For this purpose, we can work on the hypothesis that the intercalary chapters and the narrative chapters of the novel have different linguistic features in terms of word cluster, type-token ratio, descriptivity and activity, keyness, and sentiment.

2 Methodology

2.1 Corpus

The novel consists of 30 chapters which can be divided into three parts according to the spatial movement. The first ten chapters talking about the eviction of the farmers are set in Oklahoma. The following ten chapters describe the events on Route 66 until the characters arrive at the destination of California. The last ten chapters concern the lives of migrant farmers in California. Of the 30 chapters, 16 are intercalary. The size of each of the intercalary chapters is smaller than 5,000 words, while 13 out of the 14 narrative chapters contain more than 5,000 words each. Most of the intercalary chapters form alternations with the narrative chapters in the narrative structure. The details of the chapters are shown in Table 1.

Table 1. An overview of the chapters in the novel

Part 1			Part 2			Part 3		
Chapter	**Intercalary/ Narrative**	**No. of words**	**Chapter**	**Intercalary/ Narrative**	**No. of words**	**Chapter**	**Intercalary/ Narrative**	**No. of words**
1	I	1,396	11	I	841	21	I	942
2	N	3,752	12	I	1,951	22	N	16,502
3	I	892	13	N	11,762	23	I	2,329
4	N	6,215	14	I	975	24	N	5,981
5	I	3,673	15	I	4,159	25	I	1,462
6	N	9,933	16	N	13,417	26	N	22,334
7	I	2,164	17	I	2,985	27	I	1,039
8	N	8,662	18	N	12,450	28	N	9,112
9	I	1,464	19	I	3,530	29	I	1,197
10	N	11,227	20	N	17,518	30	N	7,696

This research also uses the fiction sub-corpus of the Brown Corpus as the reference corpus to help calculate the keyness of the keywords. The fiction sub-corpus covers a variety of novels, including 126 sample texts totalling 253,997 words. See Table 2.

Table 2. An overview of the fiction sub-corpus of the Brown Corpus in the novel

Genre category	No. of texts	No. of words
General fiction	29	58,338
Mystery and detective Fiction	24	48,231
Science fiction	6	12,043
Adventure and Western	29	58,433
Romance and love story	29	58,675
Humor	9	18,277
Total	126	253,997

2.2 Procedure

This study adopts the top-down approach. An overview of the text can be presented by cluster analysis. Then the MATTR (Moving Average Type-Token Ratio), descriptivity, activity, nominality, keyness and emotion shift are calculated. Finally, we zoom in on the patterns or details of interest and give an evidence-based interpretation.

Cluster analysis is based on the assumption that the works of different authors can be classified according to their distinct stylistic features. These features are the author's unique "signals" or style (Jockers, 2013, p. 63). Cluster analysis concerns the stylistic similarity and difference of texts, and hence can also be used to compare different chapters in one novel. According to Jockers (2014), high-frequency features of different texts can be compared through Euclidean Distance. The closer the distance is, the more common the feature usage habits are. The formula of Euclidean Distance is expressed as follows.

$$(1) \qquad d\,(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_i - q_i)^2 + (p_n - q_n)^2}$$

In Formula 1, d represents the distance, p and q are two texts, $p_1, p_2 \dots p_i, p_n$ are the measures of feature vectors in text p and $q_1, q_2 \dots q_i, q_n$ are the measures of feature vectors in text q.

Type-token ratio (TTR) has been widely used in literary research to analyze lexical richness, but this indicator can be affected by the text length. For the reliability of TTR, Covington and McFall (2010) proposed the MATTR as shown in Formula 2.

$$(2) \qquad \text{MATTR}\ (L_w) = \frac{\sum_{i=1}^{L_t - L_w} V_i}{L_w(L_t - L_w + 1)},$$

In Formula 2, L_w stands for the arbitrarily chosen length of a window, L_t for text length in tokens, and V_i for the number of types in an individual window. This approach is helpful to calculate the TTR for the subset from 1 to L_w, then for the subset from 2 to $L_w + 1$, and ends when $L_w + 1$ reaches the end of the text. In this way, the indicator of lexical richness is not influenced by the segment boundaries or text length. In this study, the MATTR[1] software is used to calculate the MATTR.

Descriptivity, activity and nominality are stylometric indicators that can describe the features of a text. The verb-adjective ratio and the noun-verb ratio have been used to differentiate styles, genres and authorships (e.g., Boder, 1940; Antosch, 1969; Popescu et al., 2013; Chen and Kubát, 2021). Descriptivity is defined as the division of the number of adjectives to the sum of verbs and adjectives and activity is equivalent to 1 minus descriptivity. Nominality is defined as the ratio of nouns to the sum of nouns and verbs. The formulas are as follows.

[1] This software is designed by Covington and McFall (2010) to eliminate the effect of text length on calculating the type-token ratio.

(3) $$activity = \frac{verbs}{verbs + adjectives}$$

(4) $$descriptivity = \frac{adjectives}{verbs + adjectives}$$

(5) $$nominality = \frac{nouns}{verbs + nouns}$$

Before calculating these indicators, we need to tag the part of speech of the tokens using the TreeTagger in the R Studio[2]. Some packages[3] are used for conducting macro analyses such as word frequency study, descriptivity and activity comparison, and sentiment analysis.

The departure from the expected word choice can be assessed by keyness. According to McIntyre and Walker (2019), it is better to calculate keyness by log-likelihood because log-likelihood does not assume a normal distribution. It shows how much evidence there is for the difference between the target corpus and the reference corpus. Rayson and Garside (2000) presented the process of calculating log-likelihood. Firstly, a contingency table is constructed as follows.

Table 3. Contingency table for word frequencies (Rayson and Garside, 2000, p. 3)

	Target corpus	**Reference corpus**	**Total**
Frequency of words	a	b	a + b
Frequency of other words	c – a	d – b	c + d – a – b
Total	c	d	c + d

Then the expected frequencies for the target corpus ($E1$) and the reference corpus ($E2$) are calculated respectively. The LL value can be calculated in Formula 8.

(6) $$E1 = c \times (a + b)/(c + d)$$

(7) $$E2 = d \times (a + b)/(c + d)$$

(8) $$LL = 2 \times \left(a \times log\left(\frac{a}{E1}\right) + b \times log\left(\frac{b}{E2}\right)\right)$$

The LL just indicates the existence of a difference between two corpora without telling the scale of difference. Thus, the log ratio is required to measure the effect size of the difference. According to

[2] A programming language like R is a more versatile tool than most ready-made software applications (Gries, 2017). R Studio can be used to calculate Euclidean Distance and draw the dendrogram by the “dist” function and the “hclust” function.

[3] The packages we used in this study are korpus.lang.en (Michalke, 2020), dplyr (Wickham et al., 2021), tidyr (Wickham, 2021), coreNLP (Arnold and Tilton, 2016), stringr (Wickham, 2019) and ggplot2 (Wickham, 2016).

Hardie (2014), log ratio is defined as the binary log of the ratio of relative frequencies, indicating the size of the difference. The calculation of log ratio (*LR*) is expressed as follows.

$$LR = log_2\left(\frac{\frac{a}{c}}{\frac{b}{d}}\right) \tag{9}$$

The sentiment variation across the plot can be roughly evaluated by the frequency of words denoting sentiments. To identify the sentiment words, researchers usually match texts with the lexicon of sentiments, but there are some problems in the matching. Hunston (2007) recognized the problem of over-simplified exploitation of intertextuality in corpus linguistics and argued that attitudinal meanings were not always transferable from one text to another. It is necessary to check the precise phraseology in the context to avoid the uncareful classification of being positive or negative. It is possible that some negative words such as *no, not* and *without* can appear preceding the words indicating emotion. Their disruption in sentiment analysis can be checked and evaluated through *n*-grams. After gathering the 2-grams consisting of one negative word and one negated word, researchers can roughly measure the degree of their disruption in the sentiment analysis. The choice of lexicon can influence the result of sentiment analysis and hence more than one lexicon is required. This study consults the Bing (Hu and Liu, 2004) and the NRC (Mohammad and Turney, 2013) sentiment lexicons[4].

3 Results and Discussion

3.1 Unsupervised Cluster Analysis

It can be expected that Steinbeck must have controlled his selection of words to write the intercalary chapters. This selection is restricted by the location of the chapter. Each chapter is influenced by its surrounding chapters to a certain degree. However, the intercalary chapters whose contents and styles are generally different from the narrative chapters make the selection of words complicated. It is hard to evaluate the style variation just by close reading. Thus, a cluster analysis is conducted to have an overview of the styles of the chapters. See Figure 1.

[4] The ratio of negative words to positive words in the lexicons can influence the result of sentiment analysis. The Bing contains 3,324 negative words and 2,312 positive words, while the NRC includes 4,782 negative words and 2,006 positive words. The NRC also categorizes the sentiment words into anger, anticipation, disgust, joy, trust and so on.

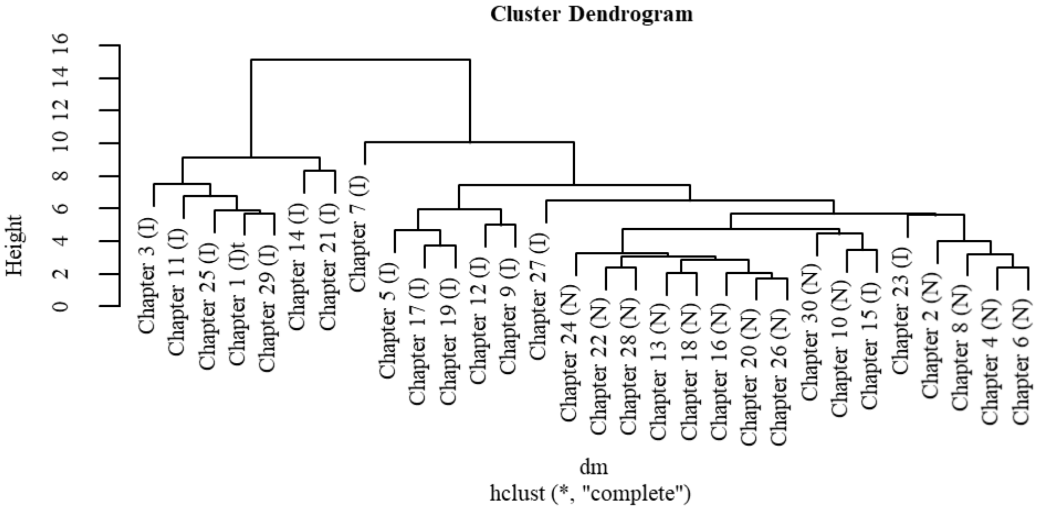

Figure 1. Cluster dendrogram of all chapters in *The Grapes of Wrath*

The dendrogram shows that most intercalary chapters are clustered on the left while the narrative chapters are on the right. The narrative chapters in Section 1 (i.e. Chapters 2, 4, 6 and 8) are merged to form their cluster, indicating that they share a high similarity. Other narrative chapters cluster nearby, the branch heights of which are below 4. The branch heights of all the narrative chapters are below 6, while the average branch height of the intercalary chapters is above 6. Even in the same section of the novel, the intercalary chapters are far away from their surrounding intercalary chapters. This attests to the heterogeneity of the intercalary chapters and the relative homogeneity of the narrative chapters.

One clade of the second-highest branch has only one leaf (i.e. Chapter 7). The isolated chunk can be interpreted as the uniqueness of Chapter 7 in the novel. The words that make Chapter 7 unique can be identified by comparing the frequency of words used in this chapter with that in the rest chapters[5]. According to Silge and Robinson (2017), the most distinct words for each file can be sorted out from the word frequency lists by consulting the log odds ratio for each word. The calculation of the log odds ratio is expressed in Formula 10.

(10) $$log\ odds\ ratio = ln(\frac{(\frac{n+1}{total+1})_{Ch-7}}{(\frac{n+1}{total+1})_{the\ rest}})$$

In formula 10, *n* is the number of times the word in question is used by each file, and *total* refers to the total words for each file. This study takes the top 15 most distinctive words for Chapter 7 and the rest chapters. The plot of these words is shown in Figure 2.

[5] With the help of the R package tidyr (Wickham, 2021), word frequencies are counted for Chapter 7 and the rest chapters.

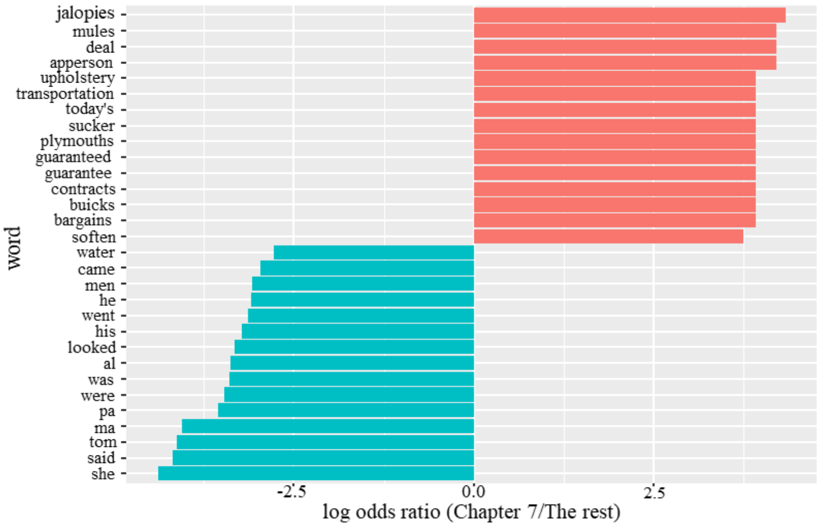

Figure 2. Comparing the odds ratio of words from Chapter 7 and that from the rest chapters

Most distinctive words in Chapter 7 form the semantic field of transportation, such as *jalopies, mules, transportation, Apperson, Buicks, Plymouths, cushions* and *upholstery*. Chapter 7 seems to be relevant to the deals and bargains in the car store. The most distinctive words in the rest chapters are characters and personal pronouns, such as *Tom, ma, pa, Al, he, his* and *she*. The lack of personal pronouns in Chapter 7 can be explained in the special speech presentation. Chapter 7 is composed of the dialogue presented in free direct speech. This special speech presentation can partly affect the lexical distinctiveness of personal pronouns in Chapter 7. In these dialogues, the speakers are mainly the owners of the car store and the salesmen involved in the dealings. As the distance between the speakers and the narrator is relatively short in the free direct speech, personal pronouns like *he* and *she* are less likely to appear in Chapter 7 than in those chapters with an omniscient narrator.

3.2 Lexical Richness

Intercalary chapters usually describe the background which is not directly relevant to the main storyline. They may provide extra information and description and are hereby expected to have a higher value of lexical richness. However, Figure 3 shows no evidence of higher MATTR values in the intercalary chapters.

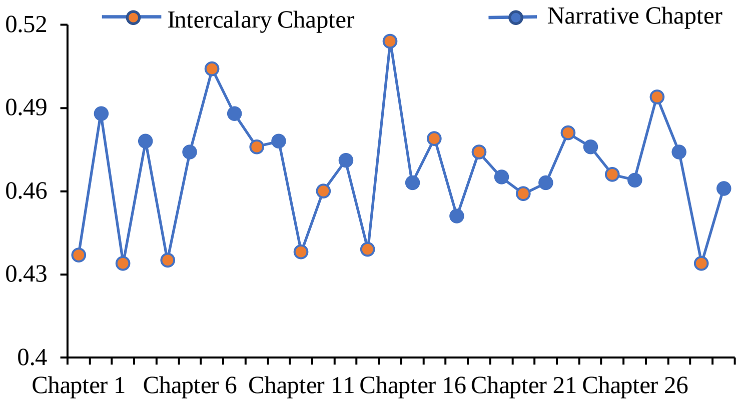

Figure 3. Variation of MATTR in *The Grapes of Wrath* (MATTR window size is 500)

According to the boxplot (see Figure 4), the average MATTR value is higher in the narrative chapters (0.471) than in the intercalary chapters (0.464). The result in the intercalary corpus varies considerably. Both the extreme values are found in the intercalary chapters. For example, the maximum value (0.514) appears in Chapter 15. The great variation testifies to the heterogeneity of the intercalary chapters.

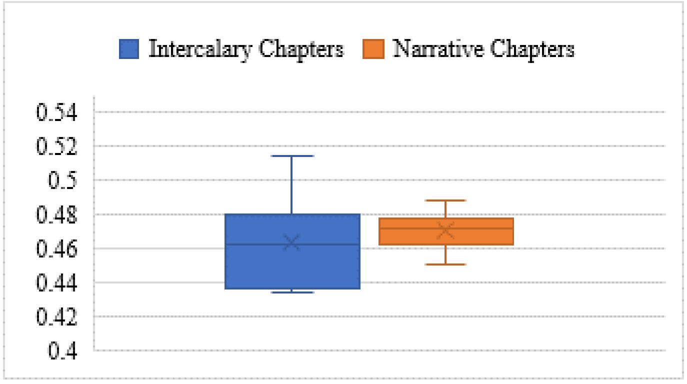

Figure 4. The boxplot of MATTR of intercalary chapters and narrative chapters

The lexical Richness of Chapter 15 can be attributed to Steinbeck’s choice of words. The specific area of lexical richness in Chapter 15 can be roughly located by tracing its 3,638 subsets. As shown in Figure 5, the first half of these subsets is richer in vocabulary than the second half. In the first half, there exist three peaks of value, indicating intensive addition of new information. The intermittent upslope shows a decrease in word repetitions. The downslope following each peak suggests a more closely related extension which possibly repeats words that have appeared previously. These intense fluctuations reveal the degree of fragmentation in the text structure of Chapter 15. Moreover, the average of the second half of subsets shows a lower value of MATTR. These subsets turn out to be a conversation. Generally, people in conversations discuss an issue at length by using simple language without introducing many new words.

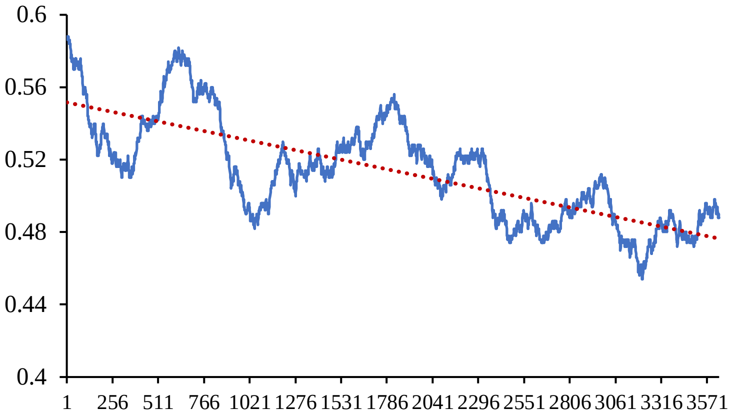

Figure 5. Variation of MATTR of subsets in Chapter 15 (MATTR window size is 500)

The second peak value (0.582) of MATTR comes from the 623rd subset. A brief extraction of this subset is shown below.

> Cars whisking by on 66. License plates. Mass., Tenn., R.I., N.Y., Vt., Ohio. Going west. Fine cars, cruising at sixty-five.
>
> There goes one of them Cords. Looks like a coffin on wheels.
>
> But, Jesus, how they travel!
>
> See that La Salle? Me for that. I ain't a hog. I go for a La Salle.
>
> 'F ya goin' big, what's a matter with a Cad'? Jus' a little bigger, little faster.
>
> I'd take a Zephyr myself. You ain't ridin' no fortune, but you got class an' speed. Give me a Zephyr.
>
> Well, sir, you may get a laugh outa this — I'll take a Buick-Puick. That's good enough.

(Steinbeck, 2006, p. 154)

The number of proper nouns and the presentation of speech can expatiate lexical richness in this subset. These sentences are characterized by short and incomplete sentences that are common in spoken language. Proper nouns of cars and places account for a large proportion of words in these sentences. Moreover, free direct speech makes the talk discursive and flexible, thus reducing the possibility of repetition. These talks seem like a semi-transcription of the hubbub and impose a bunch of information on readers without introducing the identity of the speakers. If speakers can be easily identified, the content of the talk will be limited by the specific relationship between the speakers. It is the vague presentation of the scene and speakers that paves the way for a broader scope and a higher information density. Therefore, the relevance between MATTR and conversation deserves rethinking. The conversation shown in the subset has a higher value of MATTR, whereas the ordinary conversation shows a lower value of MATTR. The way of presenting conversations should be considered in the assessment of the text structure because it may impact the information density.

3.3 Descriptivity, Activity and Nominality

The descriptivity, activity and nominality of all the chapters of the novel are shown in Figure 6 and the Boxplot of descriptivity and nominality of the two types of chapters is shown in Figure 7.

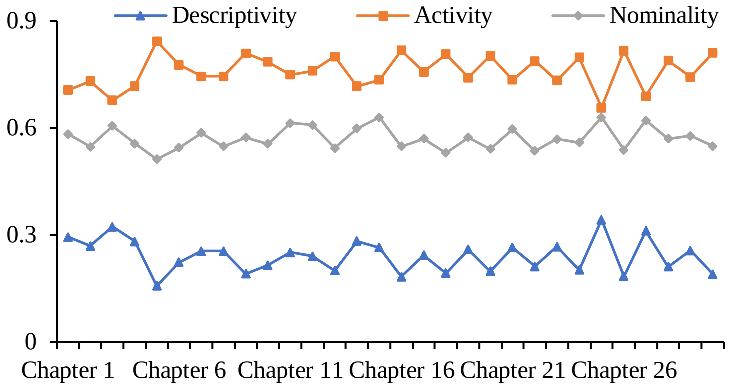

Figure 6. Descriptivity, activity and nominality of all chapters

It can be seen from Figure 6 that all the intercalary chapters have higher values of descriptivity in the last 10 chapters, i.e. Section 3. This difference is not obvious in the first two sections. This partially corroborates Levant's (2007, p. 22) criticism that "the third part…. marks an artistic decline". In the first two sections, the narrative structure does not obviously affect the style. In Section 3, the author pays much attention to the narrative structure. According to Levant (2007), style is a concomitant of structure. The high degree of manipulation seems artificial because it may lead to the overlooking of the relationship between the stories in the two types of chapters.

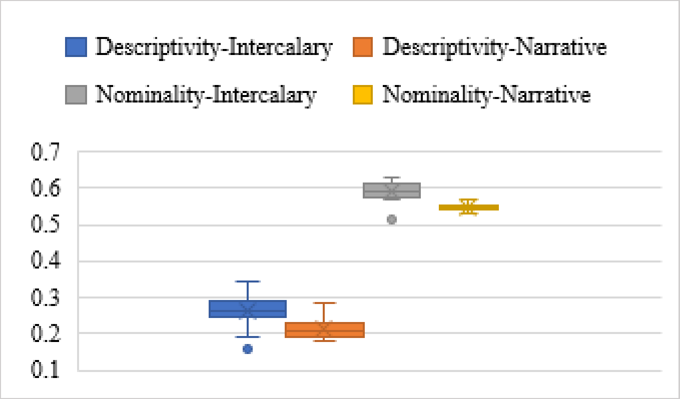

Figure 7. The boxplot of descriptivity and nominality of the intercalary and narrative chapters

The average values in Figure 7 indicate that the narrative chapters are more dynamic than most intercalary chapters. However, there exist two deviant points in descriptivity and nominality of the intercalary chapters respectively, both in Chapter 5. It is reasonable to infer that Chapter 5 is more dynamic than the other chapters. The deviant abundance of verbs signals the departure from the usual diction in the novel. Therefore, the verbs in Chapter 5 deserve in-depth exploration.

3.4 Keyness

In this section, we count, sort and lemmatize[6] the verbs in Chapter 5. It is interesting that the lemma *be* accounts for nearly 65% and *do* for almost 24.8% of all the verbs in this chapter. These prominent statistics spur the comparison of the frequencies of these two verbs in Chapter 5 and the fiction sub-corpus of the Brown Corpus. The results of the comparison will be used in the log-likelihood test. See Table 4.

Table 4. Contingency table for the keyness of *be* and *do* using log-likelihood

	Chapter 5		Brown Corpus	
	be	***do***	***be***	***do***
Frequency of words	134	51	268	154
Frequency of other words	3,390	3,473	253,729	253,843
Total	3,524	3,524	253,997	253,997

The *LL* and *LR* of *be* and *do* can be calculated respectively. The *LL* for the lemma *be* reaches 280.45, a value well above the log-likelihood critical value of 15.13. As the degree of freedom for the data is 1, the value 15.13 is associated with the *p*-value of 0.0001. Likewise, the *LL* for the lemma *do* is about 102.30. Therefore, there is a 99.99 per cent probability that both the results of the lemma *be* and *do* are significant and are not due to chance. The *LR* for the lemma *be* is about 5.17, indicating that the lemma *be* is 16 times more frequent in Chapter 5 than in the Brown Corpus. The *LR* of the lemma *do* is about 4.58, indicating that the lemma *do* is 12 times more frequent in Chapter 5 than in the Brown Corpus.

Lemma *be* will be taken as an example in the following stylistic analysis. The dispersion of the lemma *be* is shown in Figure 8.

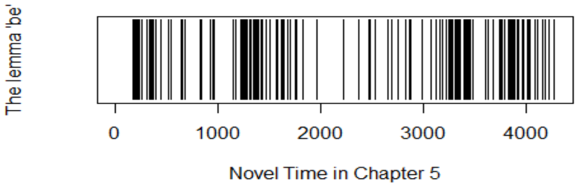

Figure 8. The dispersion plot of the lemma *be* in Chapter 5

It can be seen that there are three clusters of *be* in the plot. The extraction of the second cluster (1250-1450 in the novel time of Chapter 5) is shown as follows.

[6] This process consults the package coreNLP (Arnold and Tilton, 2016).

> We know that – all that. **It's** not us, **it's** the bank. A bank **isn't** like a man. Or an owner with fifty thousand acres, he **isn't** like a man either. **That's** the monster.
>
> Sure, cried the tenant men, but **it's** our land. We measured it…. Even if **it's** no good, **it's** still ours. **That's** what makes it ours – being born on it, working it, dying on it….
>
> **We're** sorry. **It's** not us. **It's** the monster. The bank **isn't** like a man.
>
> Yes, but the bank **is** only made of men.
>
> No, **you're** wrong there – quite wrong there. The bank **is** something else than men…. The bank **is** something more than men, I tell you. **It's** the monster. Men made it, but they can't control it.

(Steinbeck, 2006, p. 33)

This subset is the talk between one tenant and the spokesman for the farm owner. Different forms of the lemma *be* are scattered throughout the sentences. Most instances of *be* are found in the collocation *it's*. The linking verbs are the backbones of these direct statements. It is through the lemma *be* that the different meanings of *it* directly convey the conflict between the spokesman and the tenant. The pronoun *it* in the first paragraph refers to the perpetrator who drives the farmers away. In the second paragraph, *it* refers to the land in the tenants' minds. The last *it* means the bank. It can be seen that the spokesman wants to inform the tenant of the one who is to blame for the eviction, but the tenant is only concerned about the land before his eyes.

The talk is not just about the question of who is to blame; but rather it presents different ways of perceiving the world. The perception is embodied clearly in their use of the lemma *be*. The argument of the real attribution forms a tension between different perceptions in the conversation. The tenant is not informed of the bank system. What he cares about are concrete objects, such as the land and people, but the spokesman is conscious of his location in the system. To convey the abstract social system to the tenant, the spokesman directly defines the bank as a monster. The tenant, however, cannot understand why the monster is made of people. The spokesman can only say, "Men made it, but they cannot control it". The tenant faces the fate that he has to abandon his perception of concrete lands and people and accept the destiny of being evicted by the abstract bank.

3.5 Sentiment

We matched each chapter of the novel with the Bing and the NRC lexicons respectively[7]. The value of sentiment was calculated by the value of positive words minus the value of negative words. The result is presented in Figure 9.

[7] The matching process is conducted in R Studio with the help of the dplyr package (Wickham et al., 2021) and tidyr package (Wickham, 2021).

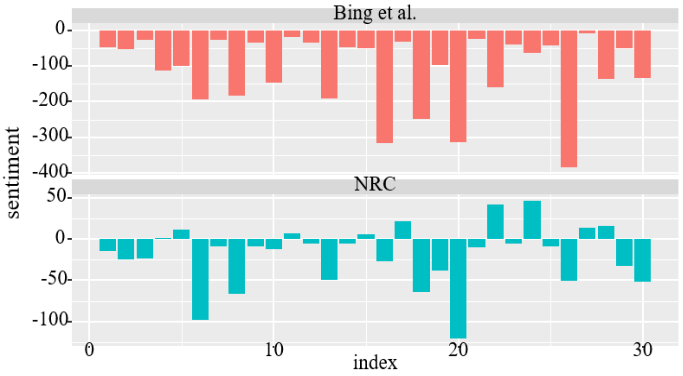

Figure 9. Sentiment through the narrative of *The Grapes of Wrath*

If the novel is matched with the Bing, all chapters contain the value indicating negative sentiments. If the novel is matched with the NRC, only a small number of chapters have positive values and other chapters still show negative sentiments. The existence of a block of negative values shows that negative sentiments permeate the novel. Moreover, the absolute value of the intercalary chapters tends to be lower than that of the narrative chapters. The narrative chapters are more abundant in negative sentiments, whereas the intercalary chapters seem to serve as the subsidiary texts to build up momentum for displaying more sentiments in the narrative chapters.

The result of sentiment analysis in Figure 9 can be roughly checked by calculating the negated words in 2-grams extracted from the novel. Figure 10 shows the numbers of negated words from the Bing lexicon.

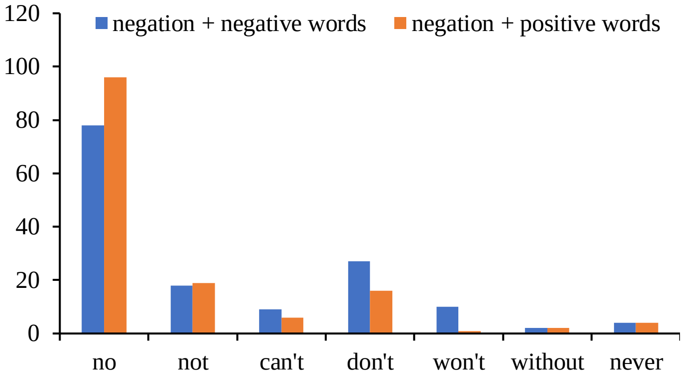

Figure 10. The number of negated words from the Bing lexicon in the sentiment analysis

In Figure 10, most negated words follow the negation *no*. Moreover, the number of negated negative words (144) is nearly as large as that of negated positive words (148). The distribution of negated words supports the result of previous sentiment analysis that negative sentiments generally permeate the novel.

4 Conclusion

The Grapes of Wrath is characterized by intercalary chapters and narrative chapters. This study adopted the narratological distinction of chapters as a framework and conducted a corpus-driven stylistic analysis in terms of word cluster, lexical richness, descriptivity, activity, nominality, keyness and sentiment. Starting from these indicators, the analysis then drills down to the specific value of statistical significance and interprets the linguistic features.

The cluster analysis shows that the narrative chapters are much more homogenous than the intercalary chapters. Of the intercalary chapters, Chapter 7 is unique in the word choice. This uniqueness can be interpreted from the semantic field of the words and the presentation of speech. The MATTR analysis shows that the intercalary chapters vary more than the narrative chapters in terms of lexical richness. Chapter 15 possesses the highest MATTR value, and its lexical richness is relevant to the presentation of speech. The study of descriptivity, activity and nominality finds that the author manipulates the style in Section 3 of the novel. Although the narrative chapters are more active than most intercalary chapters, the intercalary Chapter 5 is the most active in the novel. The keyness analysis shows that the activity of Chapter 5 features the frequent use of the lemma *be*, which helps create the tension between different perceptions of characters. The sentiment analysis finds that negative sentiments generally permeate the whole novel, and the narrative chapters are more abundant in negative sentiments than the intercalary chapters.

Acknowledgements

This research is supported by the National Philosophy and Social Sciences Fund of China (21BYY043).

References

Antosch, F. (1969). The diagnosis of literary style with the verb-adjective ratio. In: Doležel, L., Bailey, R. W. (Eds.). *Statistics and Style*, pp. 57-65. New York: American Elsevier.

Arnold, T., Tilton, L. (2016). coreNLP: Wrappers around Stanford coreNLP tools. R package version 0.4-2. https://CRAN.R-project.org/package=coreNLP

Boder, D. P. (1940). The adjective-verb quotient: A contribution to the psychology of language. *Psychological Record*, 3, pp. 310-343. https://doi.org/10.1007/BF03393230

Burcar, L. (2018). The (Forgotten) significance of interchapters in John Steinbeck's The Grapes of Wrath: From tenancy to seasonal migrant farm labor. *Arcadia*, 53(2), pp. 360-378. https://doi.org/10.1515/arcadia-2018-0027

Chen, X., Kubát, M. (2021). Rural versus urban fiction in contemporary Chinese literature: Quantitative approach case study. *Digital Scholarship in the Humanities*. (Online) https://doi.org/10.1093/llc/fqab094

Covington, M. A., McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), pp. 94-100. https://doi.org/10.1080/09296171003643098

Gries, S. T. (2017). *Quantitative Corpus Linguistics with R: A practical introduction (2nd ed.)*. New York: Routledge.

Halliday, M. A. K. (1971). Linguistic function and literary style: An inquiry into the language of William Golding's The Inheritors. In: Chatman, S. (Ed.). *Literary Style: A symposium*, pp. 330-368. Oxford: Oxford University Press.

Hamilton, S. (2016). The legacy of Steinbeck's interchapters: The effects of palimpsest on group consciousness and universality. *Steinbeck Review*, 13(2), pp. 169-178. https://doi.org/10.5325/steinbeckreview.13.2.0169

Hardie, A. (2014) 'Log ratio: An informal introduction', Blog post. ESRC Centre for Corpus Approaches to Social Science (CASS). <http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/> (last accessed 29 December 2021).

Hardy, D. E. (2004). Collocational analysis as a stylistic discovery procedure: The case of Flannery O'Connor's Eyes. *Style*, 38(4), pp. 410-427. https://www.jstor.org/stable/10.5325/style.38.4.410

Hu, M., Liu, B. (2004). Mining and summarizing customer reviews. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168-177. New York: Association for Computing Machinery.

Hunston, S. (2007). Semantic prosody revisited. *International Journal of Corpus Linguistics,* 12, pp. 249-268. https://doi.org/10.1075/ijcl.12.2.09hun

Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. Champaign, IL: University of Illinois Press.

Jockers, M. L., Thalken, R. (2014). *Text Analysis with R for Students of Literature*. New York: Springer.

Leech, G. N., Short, M. (2007). *Style in Fiction: A linguistic introduction to English fictional prose (2nd ed.)*. Harlow: Pearson Education.

Levant, H. (2007). The fully matured art: The Grapes of Wrath. In: Bloom, H. (Ed.). *Bloom's Modern Critical Interpretations: The Grapes of Wrath*, pp. 7-37. New York: Infobase Publishing.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In: Baker, M., Francis G., Tognini-Bonelli, E. (Eds.). *Text and Technology: In honour of John Sinclair*, pp. 157-176). Amsterdam: John Benjamins. https://doi.org/10.1075/z.64.11lou

Mahlberg, M. (2014). Corpus stylistics. In: M. Burke (Ed.). *The Routledge Handbook of Stylistics*, pp. 378-393). New York: Routledge.

McIntyre, D., Walker, B. (2019). *Corpus Stylistics: Theory and practice*. Edinburgh: Edinburgh University Press.

Michalke, M. (2020). koRpus.lang.en: Language support for 'koRpus' package: English (Version 0.1-4). Available from https://reaktanz.de/?c=hacking&s=koRpus.

Mohammad, S. M., Turney, P. D. (2013). *NRC Emotion Lexicon.* National Research Council, Canada, 2.

Phelan, J. (1981). *Worlds from Words: A theory of language in fiction*. Chicago: University of Chicago Press.

Phelan, J. (1996). Voices, distance, temporal perspective, and the dynamics of A Farewell to Arms. In: *Narrative as Rhetoric: Techniques, audiences, ethics and ideology*, pp. 59-84). Columbus, OH: Ohio State University Press.

Popescu, I. I., Čech, R.,, Altmann, G. (2013). Descriptivity in Slovak lyrics. *Glottotheory*, 4(1), pp. 92-104. https://doi.org/10.1524/glot.2013.0007

Rayson, P., Garside, R. (2000). Comparing corpora using frequency profiling. In: *The Workshop on Comparing Corpora* (Volume 9), pp. 1-6. Stroudsburg: Association for Computational Linguistics. http://dx.doi.org/10.3115/1117729.1117730

Rimmon-Kenan, S. (1989). How the model neglects the medium: Linguistics, language, and the crisis of narratology. *The Journal of Narrative Technique*, 19(1), pp. 157-166. https://www.jstor.org/stable/30225242

Short, M. H. (1984). Who is Stylistics. *Journal of Foreign Languages*, 5, pp. 14-21.

Short, M. (1999). Graphological deviation, style variation and point of view in Marabou Stork Nightmares by Irvine Welsh. *Journal of Literary Studies*, 15(3-4), pp. 305-323. https://doi.org/10.1080/02564719908530234

Silge, J., Robinson, D. (2017). *Text Mining with R: A tidy approach*. Sebastopol, CA: O'Reilly Media, Inc.

Steinbeck, J. (2006). *The Grapes of Wrath*. New York: Penguin Books.

Swensen, J. R. (2015). *Picturing migrants: The Grapes of Wrath and new deal documentary photography (Vol. 18)*. Oklahoma: University of Oklahoma Press.

Wickham, H. (2016). *ggplot2 - Elegant graphics for data analysis*. Cham: Springer International Publishing.

Wickham, H. (2019). stringr: Simple, consistent wrappers for common string operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr

Wickham, H. (2021). tidyr: Tidy messy data. R package version 1.1.4. https://CRAN.R-project.org/package=tidyr

Wickham, H., François, R., Henry, L., Müller, K. (2021). dplyr: A grammar of data manipulation. R package version 1.0.7. https://CRAN.R-project.org/package=dplyr