International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

IJLLT
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

The Language Difficulty of the Original Version and the Rewritten Version of *The Time Machine*: A Corpus-based Study

Xiaoyi Wang
College of Foreign Studies, Jinan University, Guangzhou, China
Corresponding Author: Xiaoyi Wang, **E-mail**: 644182923@qq.com

| ABSTRACT
This study examines the language difficulty of the original version and rewritten version of *The Time Machine* from the perspective of lexical complexity, syntactic complexity, and semantic complexity. We aim to answer the following two research questions: (1) Is there a significant difference between the language difficulty of the original version and rewritten version of The Time Machine? If yes, in what ways? (2) Is the rewritten version well-targeted at its intended readers? The data consisted of three corpora, the original version of *The Time Machine*, the rewritten version, and passages from English subjects in the National College Entrance Examination (NCEE) of China. The results demonstrate a significant difference between the original and rewritten versions on most measures of language difficulty, but in many respects, the rewritten version still leaves much to work on before becoming suitable reading material for its target readers. With such analysis carried out, we seek to offer some pedagogical implications to better facilitate English learning in China.

| KEYWORDS
Rewritten novels; language difficulty; lexical complexity; syntactic complexity; semantic complexity; corpus linguistics

| ARTICLE INFORMATION
ACCEPTED: 27 January 2023 **PUBLISHED:** 04 February 2023 **DOI:** 10.32996/ijllt.2023.6.2.13

1. Introduction

The language difficulty of written texts has been a long-standing focus in academic research, which refers to the amount of effort one needs to summon in order to understand a text. In other words, language difficulty means the readability or feasibility of a text (McNamara, Graesser, McCarthy & Cai, 2014). With the advent of various corpora and related research tools, a growing body of corpus-based research has generated valuable insights into the analysis of language difficulty of written texts, offering researchers a more accurate and scientific lens to look at this particular question of language difficulty.

In China, English is learned by thousands of students as their foreign language. When considering English study, language difficulty is a main concern which cannot be ignored by students, as well as teachers and textbook designers. It is a critical factor when deciding whether something is suitable for a certain group of learners. Therefore, corpus tools are bound to be a great aid in facilitating English learning.

1.1. Language difficulty of written texts

The lexical features and syntactic features are the two main perspectives in corpus linguistics from which we can measure language difficulty (Liu & Chen, 2013). From the lexical perspective, Read (2000) proposed that lexical richness can be interpreted as composing of lexical density, lexical sophistication, lexical variation and errors in vocabulary use. Lexical density was first put forward by Ure, referring to the ratio of content words to all the words in a text (Ure, 1971). According to Lu (2012), content words include nouns, adjectives, notional verbs and adverbs derived from adjectives. Lexical density serves as a criterion for assessing the degree of formality——the more formal a text is, the higher lexical density will be; otherwise, the more colloquial a text is, the lower lexical density will be (Liu & Chen, 2013). Lexical sophistication is defined to be the coverage of advanced words in a text

(Read, 2000), and lexical variation means the range of vocabulary use (Malvern, Richards & Chipere, 2004). These two criteria, by telling us how advanced and varied the vocabulary of a text is, can reveal the lexical complexity of the target text. Errors in vocabulary use are often related to learners' language.

From the syntactic perspective, an important measure of language difficulty turns out to be syntactic complexity. Syntactic complexity can be decomposed into several sub-measures, including the mean length of sentences, clausal elaboration, coordination, subordination, phrasal complexity, the complexity of certain construction, etc. (Ortega, 2003).

1.2. Corpus-based research on language difficulty concerning English learning

Adopting those language difficulty measures, many studies have explored the language difficulty of some written materials related to English study. One focus of such research has been primarily on English textbooks. For example, Jiang, Li, and Zhao (2016) conducted a comprehensive analysis of the syntactic complexity of three national English textbooks with the help of Lu's L2 Syntactic Complexity Analyzer (Lu, 2010). They reported no significant difference between these three textbooks on fourteen syntactic complexity measures, but when compared to general English and academic English, there exists a significant difference between these two genres and the three textbooks as a whole.

Many English tests have received equal attention, including grade tests and academic examinations (e.g., Tang, 2009; Liu & Chen, 2013; Xu & Yan, 2021). In one such study, Liu and Chen (2013) examined the language difficulty of the reading comprehension section in CET-4 and CET-6 from the perspective of lexical and syntactic complexity. In terms of lexical complexity, which in turn is designed to contain three dimensions——lexical density, lexical sophistication, and mean word length, the language difficulty of CET-6 ranks higher than CET-4 in all three measures. In terms of syntactic complexity, which is designed to contain the mean sentence length, passive voice, and nominalization, the language difficulty of CET-6 ranks above CET-4 in the mean sentence length and nominalization. Notably, this research attaches special attention to passive voice and nominalization, which are categorized as two major ways to realize grammatical metaphors (Lassen, 2003). In fact, the importance of grammatical metaphors in language difficulty analysis has also been acknowledged by other scholars. When other factors are equal, the more grammatical metaphor a text contains, the more difficult the text will be, and learners have to exert greater efforts in order to understand it (e.g., Fan, 1996; Tang, 2009).

Although textbooks and all kinds of exams have long been identified as crucial components in English study, there are still many other indispensable resources. English novels are one of them. Rich with cultural elements, those abundant English novels provide the most natural way to strengthen one's English ability. However, given the limited vocabulary and grammar knowledge of some Chinese learners, those original novels may not be the most suitable learning material. This is how many rewritten versions targeted at certain age groups come into being, for example, the *Book Worm* series published by Oxford University Press and Foreign Language Teaching and Research Press. Evidently, the language difficulty of those rewritten versions must be set at an appropriate level so that targeted learners can achieve the "i+1" effect (Krashen, 1980). But the research on the language difficulty of those rewritten versions of original English novels remains scant.

1.3. The current research

In response to the research gap indicated above, the current study examines the language difficulty of the original version and rewritten version of *The Time Machine*, a world-renowned science fiction written by Herbert George Wells in 1895. With the help of relevant corpora and information retrieval tools, we seek to analyze language difficulty from the perspective of lexical complexity, syntactic complexity, and semantic complexity. We aim to answer the following two research questions: (1) Is there a significant difference between the language difficulty of the original version and rewritten version of *The Time Machine*? If yes, in what ways? (2) Is the rewritten version well-targeted at its intended readers?

The current research is motivated both by the absence of prior research on language difficulty analysis of rewritten English novels for English learners and also by the availability of a well-established system of language difficulty measurement.

2. Material and methods

2.1 Corpora

Our data consist of three corpora——the original version of *The Time Machine*, the rewritten version of *The Time Machine*, and a collection of the reading passages of English subjects in the National College Entrance Examination (NCEE) of China. The original version of *The Time Machine* was drawn from free ebooks by Project Gutenberg. The rewritten version was chosen from the *Chuangtoudeng* 3000-vocabulary series of English novels, published by the Beijing Institute of Technology Press in 2019. Based on original versions, the whole series is written by contemporary American writers in fluent modern English, which guarantees its readability. The most prominent characteristic of this series is its vocabulary range setting. Strictly centered on 3300 core words in

English, this series aims to provide reader-friendly reading material for English learners whose level is at or above high school. The building of these two corpora is for the purpose of answering the first research question.

Besides, the reason for building the corpus of English reading passages in the National College Entrance Examination (NCEE) of China is to add a reference corpus. In order to answer the second research question, we need a standard which can reveal the average level of high school students or above, and the reading passages in National College Entrance Examination turn out to be a fair choice on account of its 3500 vocabulary requirement (*English Curriculum Standards for Senior High Schools*, 2017). The total words of this corpus are set to be as close as the total words of rewritten version corpus for the convenience of comparison. Each of these three corpora is a plain text file in English.

Table 1 Corpus descriptive statistics.

Corpus Name	Total Words
Original version	33,543
Rewritten version	25,213
Passages in NCEE	25,547
Total	84,303

2.2 The measurement of language difficulty

In the current study, using the above listed corpora, we first determine whether there exists a significant difference between the language difficulty of the original version and rewritten version of *The Time Machine* and in what ways. Then compared with passages in NCEE, we analyze whether the language difficulty of the rewritten version is set on an appropriate level.

All written materials are firstly part-of-speech tagged using CLAWS. Then we carry out language difficulty measurement from the perspective of lexical complexity, syntactic complexity, and semantic complexity. On the lexical level, four indices are measured to quantify the language difficulty: mean word length (MWL), lexical density (LD), lexical variation (LV), and lexical sophistication (LS). When measuring lexical density, we use Lu's (2012) categorization of content words. When computing lexical variation, we do not choose the most often used "type**/**token" formula, for the result is likely to be affected by the length of the text. We use the formula "type$^{\mathbf{2}}$**/**token" which is considered to be more scientific and accurate (Wolfe-Quintero, Inagaki & Kim, 1998, as cited in Wen, 2006). As for the measurement of lexical sophistication, the corpus tool Range32 (Heatley, Nation & Coxhead, 2002) is of great help. It compares the vocabulary of your target text to its ready made reference word lists and informs you how advanced the vocabulary of the text is.

On the syntactic level, we use the L2 Syntactic Complexity Analyzer developed by Lu (2010) to automate the syntactic complexity analysis. The analyzer takes a written English language sample in plain text format as input and generates 14 indices of syntactic complexity of the sample. The 14 indices are listed in Table 2.

Table 2 Indices of syntactic complexity (Lu, 2010).

Index name		Index description
Length of production unit	MLC	Mean length of clause
	MLS	Mean length of sentence
	MLT	Mean length of T-unit
Sentence complexity	C/S	Sentence complexity ratio
Subordination	C/T	T-unit complexity ratio
	CT/T	Complex T-unit ratio
	DC/C	Dependent clause ratio
	DC/T	Dependent clause per T-unit
Coordination	CP/C	Coordinate phrases per clause
	CP/T	Coordinate phrases per T-unit
	T/S	Sentence coordination ratio
Particular structures	CN/C	Complex nominals per clause
	CN/T	Complex nominals per T-unit
	VP/T	Verb phrases per T-unit

On the semantic level, we explore the grammatical metaphor in a different corpus and mainly focus on nominalization. Given that nominalization expresses abstract concepts and processes (Liu & Chen, 2013), we believe it will contribute to the difficulty of

content understanding; therefore, we put nominalization onto the semantic level. All indices of language difficulty and their relevant corpus tools are presented in Table 3.

Index name		Corpus tools
Lexical level	Mean word length	WordSmith Tools
	Lexical density	AntConc
	Lexical variation	Range32
	Lexical sophistication	Range32
Syntactic level	14 indices of syntactic complexity	L2 Syntactic Complexity Analyzer
Semantic level	Nominalization	AntConc

Table 3 Language difficulty indices used.

3. Results

3.1 Lexical complexity

3.1.1 Mean word length

For mean word length, we use WordSmith Tools 4.0 to automate required values. The result in Table 4 shows that the mean word length of each corpus is 4.30, 4.13, and 4.47, indicating no significant difference. In order to get a more comprehensive picture of the lexical complexity in terms of word length, we also counted the proportion of words above 7 letters in each corpus (see Table 5) based on the statistics of words of different lengths produced by WordSmith Tools. By performing a chi-square test and checking with the crosstab, we confirm that there exists a significant difference between each pair of the three corpora on the proportion of words above 7 letters. The number of above-7-letter words in the rewritten version (7.68%) is significantly lower than that in the original version (11.12%), but passages in NCEE have even more words above 7 letters (11.98%) compared to the original version.

Table 4 Mean word length values.

Corpus name	Mean word length	Standard deviation
Original version	4.30	2.37
Rewritten version	4.13	2.14
Passages in NCEE	4.47	2.40

Table 5 Word length values.

Corpus name	Total	Above 7 letters (%)	x^2	p
Original version	32,833	3,651 (11.12%)	285.222	0.000
Rewritten version	24,839	1,907 (7.68%)		
Passages in NCEE	25,443	3,049 (11.98%)		

3.1.2 Lexical density

For lexical density, AntConc is used to automate relevant measures after all texts have been part-of-speech tagged using CLAWS (the used tagset is C5). Again with the chi-square test performed, there is found to be a significant difference between each pair of the three corpora. The result follows the same trend with word length values in that the lexical density of the rewritten version (43.26%) is significantly lower than that of the original version (44.81%), but passages in NCEE turn out to have the highest lexical density value (48.21%) among the three.

Table 6 Lexical density values.

Corpus name	Noun	Adjective	Notional verb	Adverb	Total	Lexical density	x^2	p
Original version	6,280	2,515	3,936	2,298	15,029	44.81%	132.942	0.000
Rewritten version	4,032	1,199	3,655	2,021	10,907	43.26%		
Passages in NCEE	6,168	1,697	3,196	1,256	12,317	48.21%		

3.1.3 Lexical variation

For lexical variation, based on the type and token values generated by Range32, we use the more scientific and accurate "type**²/**token" formula to compute the lexical variation values. From the result presented in Table 7, it is observed that the same

kind of situation emerges. Passages in NCEE rank top in terms of lexical variation value, and the rewritten version ranks bottom with a remarkable difference.

Table 7 Lexical variation values.

Corpus name	Tokens	Types	Lexical variation
Original version	32,667	4,664	665.90
Rewritten version	24,854	2,502	251.87
Passages in NCEE	24,844	4,224	718.17

3.1.4 Lexical sophistication

There are three ready made base lists available in Range32. The first includes the most frequent 1,000 words of English. The second includes the 2nd 1,000 most frequent words, and the third includes some other words which are frequent in upper secondary school and university texts from a wide range of subjects (Heatley, Nation & Coxhead, 2002). The result is presented in Table 8.

Table 8 Lexical sophistication values.

	Original version	Rewritten version	Passages in NCEE	x^2	*p*
Word list 1	26,761 / 81.92%	22,394 / 90.10%	20,403 / 82.12%	866.156	0.000
Word list 2	1,976 / 6.05%	1,451 / 5.84%	1,608 / 6.47%	9.097	0.011
Word list 3	698 / 2.14%	236 / 0.95%	980 / 3.94%	499.375	0.000
Not in the lists	3,232 / 9.89%	773 / 3.11%	1,853 / 7.46%	989.645	0.000
Total	32,667	24,854	24,844		

By performing the chi-square test and checking up with the crosstab, it is observed that there exists a significant difference between the original version and rewritten version in terms of the proportion of words in word list 1, 3, and *Not in the lists*. The proportion of words in word list 1 rise from 81.92% to 90.1% in the rewritten version, and the proportion of words in word list 3 and *Not in the lists* both decreased, falling from 2.14% to 0.95% and 9.89% to 3.11% respectively. All these figures indicate a sharp decrease in the language difficulty of the rewritten version. When compared with passages in NCEE, the results seem to deviate from what we expected. We find there also exists a significant difference between the rewritten version and passages of NCEE in each of the four levels. Words in word list 1 take up a much higher proportion in the rewritten version (90.1%) than in passages in NCEE (82.12%). Meanwhile, the proportions of words in word list 2, 3, and *Not in the lists* are all notably lower in the rewritten version.

3.2 Syntactic complexity

For syntactic complexity, Lu's L2 Syntactic Complexity Analyzer (2010) automated 14 indices for each corpus, summarized in Table 9. By performing ANOVA and follow-up LSD post-hoc comparisons, we find that except for the two measures, C/S and DC/C, there exists a significant difference between the original version and rewritten version on all other measures (with one or two asterisks).

Table 9 Syntactic complexity values.

Index name	Original version	Rewritten version	Passages in NCEE	F	*p*
MLC	9.1719	8.5899	9.2179	58.019	0.000*
MLS	16.8726	15.9315	17.2131	51.246	0.000*
MLT	13.9040	14.0276	15.4790	57.664	0.000**
C/S	1.8396	1.8547	1.8674	0.872	0.675
C/T	1.5159	1.6330	1.6792	3.841	0.006**
CT/T	0.3893	0.4594	0.4561	2.016	0.017**
DC/C	0.3238	0.3676	0.3725	0.496	0.693
DC/T	0.4909	0.6003	0.6255	3.195	0.003**
CP/C	0.2347	0.1722	0.1903	2.557	0.031**
CP/T	0.3557	0.2813	0.3195	2.705	0.026**
T/S	1.2135	1.1357	1.1120	4.524	0.002**
CN/C	0.9277	0.7487	1.0364	3.606	0.004*
CN/T	1.4063	1.2227	1.7404	2.914	0.021*
VP/T	1.8615	2.2627	2.2583	5.387	0.000**

Given the research questions of the current study, measures with significant differences between the original and the rewritten can then be divided into two categories, according to the comparison with passages in NCEE. If there is no significant difference between the rewritten version and NCEE, we assume the rewriting process to be appropriate (with two asterisks), but if there exists a significant difference, then the rewriting is assumed to be inappropriate (with one asterisk).

For MLC, MLS, CN/C, and CN/T, the rewritten version displays a lower syntactic complexity value compared with the original version, but the syntactic complexity of NCEE turns out to be the highest among the three corpora.

For C/T, CT/T, DC/T, CP/C, CP/T, T/S, and VP/T, it is suggested that there exists a significant difference between the original and rewritten versions, but with no such difference between the rewritten version and passages in NCEE. The original version is syntactically easier than the other two corpora in C/T, CT/T, DC/T, and VP/T, while it is syntactically more complex than the other two in CP/C, CP/T, and T/S.

For MLT, although there still exists a significant difference between the rewritten version and NCEE, the syntactic change that happened to the original version is in the right direction, so we also regard it as appropriate.

3.3 Semantic complexity

Taking the space limitation into consideration, together with the time-consuming work of counting up all the nominalizations in three corpora, in this research, we only focus on four kinds of frequently-appearing nominalizations, which are realized by suffixation. With the help of AntConc, we can automate the result presented in Table 10.

Table 10 Nominalization values.

Corpus name	-tion(s)	-ment(s)	-ity/ities	-ness(es)	Total	Ratio	x^2	*p*
Original version	270	86	88	112	556	1.70%	94.414	0.000
Rewritten version	78	60	38	46	222	0.89%		
Passages in NCEE	238	70	131	28	467	1.88%		

By performing the Chi-square test and checking with the crosstab, it is suggested that from the perspective of nominalization, the language difficulty of the rewritten version is significantly lower than that of the original version and also than that of the passages in NCEE, with no notable difference being reported between the later two.

4. Discussion

4.1 Lexical complexity

For word length, the similar mean word length of these three corpora is not enough to indicate similar lexical complexity. Another key factor in determining lexical complexity lies in the number of long words. The most-frequently used words in English often keep within 7 letters, so words with more than 7 letters belong to long words, which are more difficult for learners to memorize as well as understand (Liu & Chen, 2013). From the original version to the rewritten version, the above-7-letter words become significantly less, which contributes to the lessening of language difficulty.

Ex. 1. The fire burnt *brightly*, and the soft *radiance* of the *incandescent* lights in the lilies of silver caught the bubbles that flashed and passed in our glasses. Our chairs, being his patents, embraced and caressed us rather than submitted to be sat upon, and there was that *luxurious* after-dinner *atmosphere*. (original).

Ex. 2. Sitting around the dinner table, a teacher, a doctor, a psychologist, and myself —— all of us pleasantly full from the meal we had just finished and pleasantly relaxed from the wine we had been drinking. (rewritten).

Examples 1 and 2 are the corresponding parts in the two versions. The rewritten version rewrote this part in a way that many details are missed out, as well as some long words. In example 1, there are several words longer than 7 letters (in italics) which all disappear in example 2, condensed into the word “pleasantly”.

But when compared to passages in NCEE, the rewritten version seems to miss its target. The truth is that passages in NCEE have the biggest number of long words among the three corpora, which indicates that the rewritten version may be too easy for high

school students, let alone upper-grade readers. Although the passages of NCEE often contain lots of proper nouns —— names of people, places, etc. —— which are usually long words, such as " Hemningway", "Massachusetts Institute of Technology", etc., the significant difference between the rewritten version and NCEE is still too large to ignore. Therefore, in the rewriting process, it is not a must-do to get rid of all the long words in order to reduce the language difficulty.

For lexical density, content words are counted. It is known that content words express information while grammatical words serve to link information together. If a text is filled with too many content words, it is natural to infer that this text conveys too much information. Besides, a larger proportion of content words is a hallmark of a more formal style, as a colloquial style is likely to employ more words which are not crucial for the conveying of information. These two characteristics of high lexical density may both double the amount of time and cognitive pain it takes for learners to understand a text, and this is the very reason why the rewritten version chose to decrease the ratio of content words.

Ex. 3. "No," he said suddenly. "Lend me your hand." And turning to the Psychologist, he took that individual's hand in his own and told him to put out his forefinger. (original).

Ex. 4. "Now, doctor, *would you be* so kind as *to* give me your hand?" "*What do you* plan to do?" asked the doctor, a little frightened..... The scientist again addressed the doctor, "Now *please*, *sir*, your hand." (rewritten).

The lowered language difficulty of the rewritten version is demonstrated by its more conversational style. The rewritten version elaborates on the conversation part in the original version, expanding and adding details to it so as to make it easier for readers to follow the story line. Additionally, this conversational style can vividly present the described situation, as if figures in the book are standing in front of you, which makes the reading more enjoyable and less difficult. As a result, there are more non-content words or phrases (in italics) such as "would you be... to...", "what do you", and "please, sir", which are characteristic of conversations.

Although the language difficulty does decrease, whether the rewritten version is well-targeted at its intended readers remains to be checked. The comparison to the reference corpus once again reveals an inappropriate language processing, for the lexical density of passages in NCEE, is actually very high. The main theme of passages in NCEE often centers on science, social phenomena, health problems, some narrative stories and so on, which all can be marked as information-heavy. Furthermore, the genre difference —— novels are likely to have more conversations and be more informal compared to description and argumentation —— can already bring in the effect of lexical density decrease; in the rewriting process, we may not need to spend extra efforts to create a version with even lower lexical density.
For lexical variation, the rewritten version, as expected, uses less varied words, greatly easing the burden of reading comprehension.

Ex. 5. The *thing* the Time Traveller held in his hand was a glittering metallic *framework*, scarcely larger than a small clock and very delicately made. On this table, he placed the *mechanism*. The only other object on the table was a small shaded lamp, the bright light of which fell upon the *model*. (original).

Ex. 6. In his hands, he held a small metal *object*, no larger than an alarm clock. *It* looked very fragile because of all the small instruments inside. He set *it* carefully on the table and moved a lamp closer so that we could see *it* more clearly. (rewritten).

Examples 5 and 6 form a clear comparison. In the original version, four different words (in italics) are used to describe the Time Machine —— "thing", "framework", "mechanism", and "model", while in the rewritten version, only one word ", object" is used, with three "it" (in italics) representing the object later. Although those four different words in the original version all refer to the same thing, it still requires extra cognitive effort to make clear this fact. More basically, greater lexical variation equals to larger vocabulary size, which provides another challenge for English learners.

However, the rewritten version once more missed its target here. The lexical variation of passages in NCEE turns out to be the highest among the three. This has something to do with its information-heavy characteristic, realized by constantly presenting new information, including some proper names. Therefore, in order to make better use of rewritten English novels so that they can facilitate the English learning of high school students (and upper-grade learners), the vocabulary should be as varied as needed. Being so limited only to guarantee readability is something we must avoid.

For lexical sophistication, a sharp decrease is seen in the rewritten version compared to the original version, achieved by maximizing the number of easy words and minimizing the number of difficult words.

Ex. 7. The Time Traveller was *expounding* a *recondite* matter to us. His pale grey eyes shone and twinkled, and his usually pale face was flushed and *animated*. And he put it to us in this way as we sat and lazily admired his *earnestness* over this new *paradox* and his *fecundity*. (original).

Ex. 8. We listened to our host, the scientist, as he prepared to *explain*, *with great excitement*, a *new idea* he had come upon. (rewritten).

In example 7, there are several relatively uncommon words (in italics), which all belong to *Not in the lists* category of Range32. In order to reduce the burden on readers, the rewritten version replaced all those words with easier and more frequently used ones (in italics), which all belong to the word list 1. For instance, "expound" is converted into "explain", "animated", and "earnestness" are altogether converted into "with great excitement", "sleight-of-hand trick" is converted into "joke", etc.

The comparison with the reference corpus consistently produces the same result. The lexical sophistication of the rewritten version proved to be over-reduced. One thing that needs to make clear is that passages of NCEE usually contain some technical terms or some obscure words, but in most cases, all these words are provided with Chinese explanations. During the corpus building process, we erased all Chinese explanations from the texts so as to ensure all the texts were in pure English. As a result, the actual level of lexical sophistication of NCEE is lower than that automated by Range 32. Even so, the gap between the rewritten version and passages in NCEE is too large to put all the blame on those words with erased explanations, which only take up a tiny proportion in number. By employing words with different difficulty and popularity, we can adjust the lexical sophistication of a text according to our needs, but in this case, the adjustment fails to keep within the required scope. This is not conducive for English learners to really get improved with these reading materials.

4.2 Syntactic complexity

For syntactic complexity, one distinct feature is easily spotted. On the subordination level, the rewritten version has a higher value on three of the four measures (C/T, CT/T, and DC/T); on the coordination level, the original version displays a higher value on all the three measures (CP/C, CP/T, and T/S). Therefore, a different preference on grammatical structures can be detected: the original version prefers to expand sentences through coordination, while the rewritten version and passages in NCEE prefer to do this through subordination.

Ex. 9. For instance, here is a portrait of a man at eight years old, *another at fifteen*, *another at seventeen*, *another at twenty-three*, ***and*** *so on*. (original).

Ex. 10. Just look at these pictures. Here, the fellow is eight**.** In the next, he is fifteen**.** And in the last, he is twenty-one**.** (rewritten).

Examples 9 and 10 are a pair of corresponding parts in two versions, constructed in different grammatical structures. There is only one sentence in the original version, and different parts are connected with each other through coordination (in italics). However, instead of relying on coordination, the rewritten version constructed the same part with segmentation, creating several more sentences. This accords with the grammatical pattern of passages in NCEE, which proved to use fewer coordinate structures. Thus this grammatical re-structuring is assumed to be appropriate.

It is hard to find such semantically and syntactically comparable pairs in the original version and rewritten version to elucidate this grammatically distinct feature, but many single examples (without comparable match in another corpus) in each version can be offered to add more details.

Ex. 11. Scientific people," proceeded the Time Traveller, ***after*** *the pause required for the proper assimilation of this*, "know very well that Time is only a kind of Space. Here is a popular scientific diagram, *a weather record*. This line I trace with my finger shows the movement *of the barometer*. Yesterday it was so high, *yesterday night it fell*, ***then*** *this morning it rose again*, ***and*** *so gently upward to here*. Surely the mercury did not trace this line *in any of the dimensions of Space generally recognised*? But certainly, it traced such a line, ***and*** *that line*, therefore, we must conclude, *was along the Time-Dimension*." (original).

Ex. 12. Now I want you clearly to understand that this lever, being pressed over, sends the machine gliding into the future, ***and*** *this other reverses the motion*. This saddle represents the seat *of a time traveller*. Presently I am going to press the lever, ***and*** *off the machine will go*. It will vanish, *pass into future Time*, ***and*** *disappear*. Have a good look at the thing. Look at the table too, ***and*** *satisfy yourselves; there is no trickery*. I don't want to waste this model ***and then*** *be told I'm a quack*. (original).

Examples 11 and 12 are two fairly long paragraphs in the original version, in which many coordinate structures densely appear (in italics), including some coordinate phrases such as "of the barometer", "of a time traveller", etc. The use of conjunctions (in bold) is another accompanying feature in these coordinate structures.

Ex. 13. "I can't say ***that*** *I'm too impressed*," said the doctor. "I've seen similar tricks performed before. It's not exactly clear *how they do it*, but I'm pretty sure *it has something to do with the lighting of the room*." "Well," I said, "It seems *I'm going to have an opportunity to ask him more about it*." (rewritten).

Ex. 14. The three new additions to our party were Blank, a famous newspaper editor, a journalist ***whose*** *name I never learned*, and a quiet, bearded gentleman ***who*** *was quite forgettable* because he hardly spoke the entire evening. Strangely enough, the scientist was not present. ***When*** *I asked where he might be*, the doctor said ***that*** *he had left a note* indicating ***that*** *he was going to be late for dinner and that* ***if*** *he had not arrived by seven o'clock, we should begin our meal without him*. (rewritten).

Examples 13 and 14 are representative of the subordinate structures (in italics) in the rewritten version. For passages in NCEE, their main themes make them become information-intensive, which inevitably will produce many subordinate structures. When some information needs to be condensed into a single sentence, or when we want to weave the information into a more connected and compact whole, subordination is bound to be the choice. Therefore, in order to cater to this grammatical pattern of NCEE, the rewritten version prefers to expand sentences through subordination, like examples 13 and 14, while the original version favors coordination more, like examples 11 and 12. This selection of grammatical patterns in the rewritten version is assumed to be appropriate here.

4.3 Semantic complexity

According to Halliday (2000), grammatical metaphor is mainly embodied in nominalization, which is the most powerful way to realize the grammatical metaphor. The use of metaphors in texts poses a challenge to readers' cognitive processing ability, as metaphors do not reflect the actual world directly.

Ex. 15. The Time Traveller proceeded, "Any real body must have *extension* in four *directions*: it must have Length, Breadth, *Thickness*, and —— *Duration*. But through a natural *infirmity* of the flesh, which I will explain to you in a moment, we incline to overlook this fact. There is, however, a *tendency* to draw an unreal *distinction* between the former three dimensions and the latter because it happens that our *consciousness* moves intermittently in one *direction* along the latter from the beginning to the end of our lives." (original).

Ex. 16. It seems our math teachers all believe that space *is measured according to* just those three basic dimensions: height, width, and *depth*. However, they are mistaken here; for there is a fourth dimension." "And what dimension would that be?" asked Filby, now seeming a little uncomfortable. "Why *time*, of course." (rewritten).

In example 15, there are several nominalizations (in italics), which are formed by adding suffixes "-tion", "-ness", and "-ity" to the original word classes. After nominalization, all these words get to express more abstract ideas, making them more demanding for readers to understand. Thus, in order to relieve the burden of comprehension, the rewritten version replaced some of those abstract words with more direct and reader-friendly ones (in italics) —— "extension in four directions" becomes "be measured according to", "thickness" becomes "depth", and "duration" becomes "time". In the meantime, some nominalizations in the original version may be totally erased during the rewriting process, such as "infirmity", "tendency", "distinction", and "consciousness" in example 15.

However, the significantly higher proportion of nominalization in passages of NCEE again indicates an inappropriate rewriting. What we observed here is the same language difficulty level in terms of nominalization when compared to the original version with passages in NCEE. The reason, as before, can be boiled down to the main themes of passages in NCEE, for social phenomena, health-related problems, science, etc., are positively rich in nominalizations. Therefore, we may not need to make much change in the rewriting process in this respect.

5. Conclusion

This study investigated the language difficulty of the original version and rewritten version of *The Time Machine* from the perspective of lexical complexity, syntactic complexity, and semantic complexity. For the first research question, our analysis revealed a significant difference in the language difficulty between the original version and rewritten version in terms of 4 measures for lexical complexity (word length, lexical density, lexical variation, and lexical sophistication), 12 measures for syntactic complexity (measures concerning sentence length, subordination, coordination, complex nominals, etc.), and nominalization for semantic complexity. For the second research question, on the lexical level, the language difficulty of the rewritten version is over-reduced;

on the syntactic level, the rewriting is appropriately done in terms of subordination, coordination, and verb phrases ratio, but in other respects, such as sentence length, the language difficulty of the rewritten version turns out to be too easy for its intended readers; on the semantic level, the rewritten version is unduly simplified, failing to be useful and suitable reading material for its target readers.

Through these analyses, we seek to offer some pedagogical enlightenment. It's easy to find plenty of rewritten English novels in China, which all claim to be excellent reading material for students at a certain age or on a certain level. Nevertheless, how to control language difficulty is by no means an easy task; as the case in the current study has shown, many rewritten novels may fail to achieve the goal they set. The key is to have a good grasp of the actual language competence of your target readers, which can be realized by collecting some representative reference materials. It is acceptable and can be beneficial to set the language difficulty level a little above the current level of your target readers, but neither being too easy nor too demanding can readers harvest any gains.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors and the reviewers.

Reference:

[1] Fan, W. F. (1996). The influence of grammatical metaphors on the readability of texts. *Journal of Peking University (Philosophy and Social Sciences),* (03), 66-70.
[2] Halliday, M. A. K. (2000). *An Introduction to Functional Grammar*. Beijing: Foreign Language Teaching and Research Press.
[3] Heatley, A., Nation, I. S. P. & Coxhead, A. (2002). RANGE and FREQUENCY programs. http://www.vuw.ac.nz/lals/staff/Paul_Nation.
[4] Jiang, T. C., Li, S. Q., & Zhao, L. M. (2016). A corpus-based study of syntactic complexity in college English textbooks. *Journal of Qiqihar University (Phi & Soc Sci)*, (06), 180-182+185.
[5] Krashen, S. (1980). *The Input Hypothesis*. Washington D. C.: Georgetown University Press.
[6] Lassen, I. (2003). *Accessibility and acceptability in Technical Manuals*. Philadelphia: John Benjaminis B. V.
[7] Liu, B., & Chen, J. S. (2013). Language difficulty of CET-4 and CET-6 reading comprehension —— A corpus-based study. *Journal of Chongqing Jiaotong University (Social Sciences Edition), 13*(5), 141-144.
[8] Lu, X. F. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474-496.
[9] Lu, X. F. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2), 190-208.
[10] Malvern, D., Richards, B., & Chipere, N. (2004). *Lexical density and language development: quantification and assessment*. Houndmills, England: Palgrave MacMillan.
[11] McNamara, D.S., Graesser, A.C., McCarthy, P.M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. New York: Cambridge University Press.
[12] Ministry of Education of the People's Republic of China. (2017). *English Curriculum Standards for Senior High Schools.* Beijing: Beijing Normal University Publishing House.
[13] Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college level of L2 writing. *Applied Linguistics, 24,* 429-518.
[14] Read, J. (2000). *Assessing Vocabulary*. Oxford: Oxford University Press.
[15] Tang, M. H. (2009). A contrastive analysis of the reading comprehension in English major level tests. *Journal of Educational Institute of Jilin Province, 25*(01), 75-77.
[16] Ure, J. (1971). *Lexical Density and Register Differentiation*. London: Cambridge University Press.
[17] Wen, Q. F. (2006). A longitudinal study on the changes in speaking vocabulary by English majors in China. *Foreign Language Teaching and Research, 38*(03), 189-195+240-241.
[18] Xu, P., & Yan, Z. K. (2021). A corpus-based contrastive study of the linguistic complexity of CET-6 reading. *Journal of Anshun University, 23*(1), 49-54.