分类号： H315.9　　　　　　　　　　单位代码： 10335

密　级：　　　　　　　　　　　　　学　号： 21805060

浙江大学

硕士专业学位论文

中文论文题目：语料库驱动的文学性文本与非文学性文本的机器翻译表现研究

外文论文题目：A Corpus-driven Study of Machine Translation Performance on Literary and Non-literary Text

申请人姓名：　刘美君

指导教师：　梁君英（教授）

专业学位类别：　翻译硕士

专业学位领域：　英语笔译

所在学院：　外国语言文化与国际交流学院

论文提交日期　2020 年 3 月 20 日

语料库驱动的文学性文本与非文学性文本的机器翻译表现研究

论文作者签名：刘美启

指导教师签名：梁君英

论文评阅人 1：

评阅人 2：

评阅人 3：

评阅人 4：

评阅人 5：

答辩委员会主席：刘翼斌

委员 1：王之光

委员 2：徐雪英

委员 3：

委员 4：

委员 5：

答辩日期：2020年5月29日

A Corpus-driven Study of Machine Translation Performance on Literary and Non-literary Texts

Author's signature: 刘美辰

Supervisor's signature: 梁君英

Thesis reviewer 1: __________

Thesis reviewer 2: __________

Thesis reviewer 3: __________

Thesis reviewer 4: __________

Thesis reviewer 5: __________

Chair: 刘翼斌

(Committee of oral defence)

Committeeman 1: 王文光

Committeeman 2: 徐雪英

Committeeman 3: __________

Committeeman 4: __________

Committeeman 5: __________

Date of oral defence: 2020.05.29

Acknowledgements

I would like to extend my sincere gratitude to all the people who have offered me assistance and support on my graduation thesis.

First of all, I want to express thanks to my supervisor professor Liang Junying. She guides me to a scientific and rigorous way of thinking which indeed helps me a lot in the preparation of my graduation thesis. Besides, the enthusiastic and positive attitude towards work she has also encourages me to face every challenge in my life with confidence and ease. I feel so lucky to stay in her research group over the last two years of graduate study.

Secondly, I am very grateful to all the faculty members in the School of International Studies. I have learned a lot from their teaching courses for the past six years in Zhejiang University and all things I have learned lay the foundation of my present study.

My gratitude also goes to my Thai friend Ye Dafu who helps me finish the python program that calculates linguistic indicators and to all the members of professor Liang Junying's research group who offer me advice.

Finally, I would like to acknowledge all my family do for me. Thanks to my parent's support and understanding, I spend a wonderful college time. And thanks to my boyfriend Liao Yuan, he gives me much comfort and encouragement when I am at a loss.

摘要

近几年机器翻译蓬勃发展，极大地便利了人们的生活。人们普遍关注的问题是：机器翻译能否取代人类翻译？前人研究结果显示，以传递信息为主的说明文翻译，一定程度上存在被机器翻译取代的可能性。但是以传递人类情感为主的文学翻译，将面临怎样的机遇与挑战呢？该问题目前缺少系统的量化的研究证据。

为探究机器翻译的发展现状，评估其当下是否有能力翻译文学作品，本研究进行了对比研究，分别测试机器翻译在文学性文本和非文学性文本上的表现，旨在检测机器翻译在处理两类文本时是否存在显著差异。本研究首先采用基于语料库的翻译研究方法，计算出文本的三个语言学指标：类符形符比（TTR）,实词密度和平均句长，之后进行案例分析，验证数据结果。

数据结果表明，非文学性文本的机器翻译与其原始文本高度相关，三个语言学指标间均存在显著正相关，而文学性文本的机器翻译与其原始文本之间只有类符形符比一个语言学指标呈显著正相关。案例分析结果显示，机器翻译在文学性文本上的表现尚不能达到用户期望，输出文本有多处错译、漏译。总体而言，非文学性文本的机器翻译表现优于文学性文本的机器翻译表现。

关键词：机器翻译，文学性文本，非文学性文本，类符形符比，实词密度，平均句长

Abstract

Machine translation, as an emerging technology, has greatly facilitated people's lives, which gives rise to a widely held concern over whether machines have reached the intelligence level of human translators. Previous studies show that, to some extent, it is possible for machine translation to handle expository texts that mainly deliver objective information. But what opportunities and challenges may machine translation encounter when it comes to literary works that mainly convey human emotions? At present, there is not much systematic and quantitative research on this issue.

To explore the current development of machine translation and assess its ability to translate literary works, the present study conducted a comparative study by testing machine translation performance on both literary text and non-literary text to explore if there was a significant gap between two types of machine translation. A corpus-driven analysis was carried out first to calculate three linguistic indicators, Type Token Ratio (TTR), Content Words Density and Average Sentence Length, then followed by case analysis to verify the results.

The results show that the machine translated non-literary text is highly correlated with its raw text with significant positive correlation found in all three linguistic indicators, while between literary text and its machine translation only one linguistic indicator TTR reports significant positive correlation. Further, the case analysis result of many mistranslations and omissions existing in the output text indicates that machine translation performance on literary text is unable to meet users' expectations yet. Overall, the study concludes that the machine translation of non-literary text outperforms that of literary text.

Key Words: Machine Translation, Literary Text, Non-literary Text, Type Token Ratio, Content Words Density, Average Sentence Length

Contents

List of Figures

List of Tables

1. Introduction

The introduction part begins with the background information and significance of present study to state reasons for choosing this research topic. Then it is followed by the research purpose and three major research questions. After that it ends up with a brief introduction to the thesis structure.

1.1 Background Information and Significance of Present Study

In times of the fourth industrial revolution when artificial intelligence and corresponding high technologies prevail in the market, a multitude of jobs will be replaced by machines of high efficiency and accuracy. Different from the former three revolutions that all ended at weeding out workers who did simple and repetitive tasks, the fourth one extends further by imperiling high-tech employments. What's the boundary of AI development? A puzzling question lies ahead most industries.

Back in 1956, the Dartmouth Summer Research Project on Artificial Intelligence (DSRPAI) first proposed key aspects of artificial intelligence problems like how to make machines understand and use human language, marking the beginning of AI research (McCarthy et al., 2006). As the computer doubles its memory and speed every year, AI application has been expanding its reach into various fields. Machine translation, as a hot AI field in recent years, has gone through generations of upgrading from "example-based" to "neural network", which achieves a great rise in translation speed and accuracy and particularly unleashes its talents to the full in dealing with professional documents with the aid of copious stocks of technical terms.

Translators react to the coming wave of AI technology with a mixed feeling of hope and concern. On the one hand, translators across all fields take advantage of many newly developed technological tools like online dictionaries, translational corpus and translation memory software that make their work more agreeable and effective and thus "release their cognitive resources for complex tasks by relieving them of repetitive and boring tasks" (Taivalkoski-Shilov, 2018). On the other hand, the threat of replacement now not only comes from other professional translators as in the past, but also from raw or post-edited machine translations which in clients' view represent cheaper price and shorter delivery time.

To explore how translators feel about machine translation, a research group from Zurich employed automatic sentiment analysis on a large collection of online posts (from Facebook, LinkedIn and Twitter) about translators' opinions on machine translation, and the experimental evidence shows that negative perceptions in social media outnumber positives (Läubli & Orrego-Carmona, 2017). After studying a full range of translation cases in market, Way (2013) proposed a rule of thumb that the degree of human involvement in translation scenario should be associated with the lifespan of the content, and claimed that "those translators who argue that there is only one level of quality, namely 'flawless' human translation, are obviously stuck in the dark ages". In other words, it is moot to question whether machine translation is useful or not, and the focus now is to distinguish which types of source texts are suitable for using machine translation and which still rely on human translators.

When it comes to which type of texts shall still rely on human translators, the first one must be literary works, which generally represent the wisdom collections of human beings passing down through generations. In fact, "no one seriously believes that a computer could understand Shakespeare since the intuitions of most people, both in and outside academia, are that computers are too rigid, formulaic, and cold to pick up the subtle nuances, emotions, and creative brilliance manifested in literature" (Graesser et al., 2011). Although the proposal that machine understands literature would be blasphemous for readers and translators who focus on literary works, it is undeniable that machine translation might one day rise to the human intelligence level. Hence one essential task facing both developers and translators is to always follow up machine translation performance on literary works. Only by knowing how far machines have achieved on this translation type can researchers timely update their research direction and translators better plan for the future.

1.2 Research Purpose

With an aim of assessing the machine translation performance, a quantitative analysis is applied to study linguistic indicators of chosen texts to get a comprehensive and objective result first. In this study, besides preparing a literary text as a study object, a non-literary text is needed as well for later carrying out a contrastive analysis which better helps reveal the current development of machine translation performance on literature. Three time-honored linguistic indicators Type Token Ratio (TTR), Content

Words Density and Average Sentence Length of both literary and non-literary texts as well as their machine translations are calculated and analyzed. Previous researches suggest that Type Token Ratio and Content Words Density are the most commonly used linguistic indicators to assess lexical richness and Average Sentence Length is helpful in revealing the characteristic of writing style (Richards, 1987; Laufer & Nation, 1995; Yule,1939). Pearson correlation analysis is then carried out to examine whether the machine translated texts are positively correlated with their original texts.

To verify the data results, the present study sets a case analysis part before discussions and conclusion. Admittedly, data could provide powerful evidence for general patterns, but sometimes they are limited in convincing readers of this type of general pattern, particularly in translation studies because there are a thousand Hamlets in a thousand people's eyes. A combination of both comprehensive data research and concrete case analysis is supposed to better support the findings.

In an effort to probe into the linguistic differences between machine translation performances on literary and non-literary texts from both quantitative and qualitative perspectives, and to explore whether the literary works are translatable for recently updated machine translation, the present study is designed to answer the following three research questions:

(1) Is there any significant positive correlation in three linguistic indicators between literary text and its machine translation?

(2) Is there any significant positive correlation in three linguistic indicators between non-literary text and its machine translation?

(3) Is machine translation performance on literary text equal to that of non-literary text?

1.3 Thesis Structure

The thesis is composed of seven chapters in total.

Chapter one is a brief introduction to the present study. The first section focuses on the research background and significance. The second one is about the research purpose and major research questions that guide the study.

Chapter two is literature review. The first section mainly introduces three development stages of machine translation and different academic views on its fast development. The second section gives a brief review of various literary translation theories home and abroad and lists a few researches about machine translation attempts on literary works. The third section focuses on the corpus-related translation researches.

The research design is given in Chapter three. In the first section, the reason for choosing *Love in a Fallen City* and *Bain China Shopper Report 2019* as literary and non-literary texts respectively are explained. The second section introduces three linguistic indicators Type Token Ratio (TTR), Content Words Density and Average Sentence Length. After that, research tools used in the present study are listed in the third section.

Chapter four of Data Analysis and Results is generally divided into three sections, each one focusing on one indicator. Every section first presents data of paragraphs chosen from sample texts and then lists the Pearson correlation coefficient between two pairs (pair 1: literary text and its machine translation; pair 2: nonliterary text and its machine translation).

Chapter five chooses five literary paragraphs and five non-literary paragraphs to conduct case analysis. Each case would center on how machine translates specific words and sentences and whether the translation quality is acceptable or not.

Chapter six discusses what has been found in the present study. The first section gives out explanation on the results obtained from corpus analysis. The second section evaluates the machine translation performance on two types texts based on case analysis result.

The final chapter concludes major findings of the present study and proposes suggestions for future studies on this subject.

2. Literature Review

The literature review consists of three sections. The first section focuses on the history of machine translation. The second section summarizes different academic theories on literary translation and researches about machine translation attempts on literary texts. The third section mainly elaborates on the significance of corpus study method and its application in translation studies.

2.1 Literature Review on Machine Translation

2.1.1 History of Machine Translation

The term Machine Translation refers to computerized systems producing translations of oral or written texts in different languages. The ideal goal of machine translation is to automatically produce high-quality translations without any human assistance. As Hutchins (1995) concluded at the start of *Machine Translation: A Brief History*, "The translation of natural languages by machine, first dreamt of in the seventeenth century, has become a reality in the late twentieth". Before the landing of functioning translation machines was a longtime exploratory stage. In 1933, George Artsrouni invented a storage device similar to a multilingual mechanical dictionary, through which equivalent words of other languages could be found to facilitate human translation and in the same year Petr Smirnov-Troyanskii put forward a systematical idea of how to achieve easier translation in a mechanical way (Hutchins, 2007). Later in 1951, Yehoshua Bar-Hillel, a linguistic researcher at the Massachusetts Institute of Technology (MIT), started to devote his time on machine translation realization and convened the first machine translation conference on which one important idea that although perfect machine translation was a virtual impossibility, pre-editing or post-editing was expectable took shape (Hutchins, 2007).

Machine translation study gradually entered the right track and went through stages from "Rule-based Machine Translation", "Example-based Machine Translation" into "Neural Machine Translation (NMT)". The "Rule-based Machine Translation" relies on linguistic knowledge like grammars, semantics and transfer rules to produce complete sentences by combining linguistic units (Hutchins & Somers,1992). It outputs translations often more faithful to the original due to terminology consistency while less natural and fluid due to lack of smoothing techniques (Forcada et al., 2011).

Concerning “Example-based Machine Translation”, Nagao (1984) first emphasized the importance of examples in machine translation research and proposed a method called “Machine Translation by Analogy” which was claimed to be more inclined to human thinking: rather than translating a sentence by doing detailed linguistic analysis, people would like to translate input sentences by consulting proper examples in dictionaries. The “Exampled-based Machine Translation” mechanism retrieves similar examples from large-scale databases and adapts them to translate source texts, which is, in the opinion of most people, more effective than rule-based one since adding a new rule into the translation system is time-consuming and hard to anticipate its effect (Sumita et al., 1990). Both machine translation systems have pros and cons, and neither of them is inherently better than the other. With a great rise in computational power the “Example-based Machine Translation” will achieve higher translation accuracy and speed, and with proper post-editing strategies the “Rule-based Machine Translation” will be more humanlike.

“Neural Machine Translation (NMT)” was introduced into market years ago and outperformed machine translations of fore-mentioned two types in terms of quality on most cases (Du & Way, 2017). Different from its predecessor “Example-based Machine Translation” that collects a mass of small sub-components and then separately turns them into other languages for use, NMT intends to train a well-organized neural network by adopting an encoder-decoder approach that reads sentences and directly outputs a correct translation (Bahdanau et al., 2014). As Cho et al. (2014) explained, “the encoder extracts a fixed-length representation from a variable-length input sentence, and the decoder generates a correct translation from this representation”. Since the birth of NMT, promising results have been being reported in comparison with other machine translation paradigms and undoubtedly NMT marks a big step forward in machine translation field, but meanwhile there are still limitations for NMT model that need to be addressed (Castilho et al., 2017). Koehn & Knowles (2017) listed six challenges for NMT: 1. NMT sacrifices adequacy for the sake of fluency in some point; 2.NMT possesses a steep learning curve in regard to the amount of training data; 3.NMT is weak in translating low-frequency words of highly-inflected categories; 4. NMT is weak in dealing with very long sentences (usually above 60 words); 5. NMT attention model does not always take the role of a word alignment model; 6. Beam search decoding is only applicable to narrow beams and is useless while exposed to a larger search space.

2.1.2 Views on Machine Translation

Today many big companies like Google, Tencent and Baidu are working full-speed on developing and advancing their own NMT products, due to which NMT has been applied into various real life scenarios to facilitate people's daily life. It is beyond question that in this new state Google Neural Machine Translation (GNMT) is the brightest star who possesses the most advanced techniques critical to translation accuracy and speed. Wu et al. (2016) conducted a research to test GNMT performance, concluding that GNMT system had achieved the translation quality achieved by average bilingual human translators on some test sets.

The fast-developing machine translation does bring great benefits to users by improving translation speed or reducing translation cost, while at the same time it causes a lot of concerns and thus will not replace human beings for the foreseeable future. Taivalkoski-Shilov (2018) reminded of ethical issues regarding machine translation of literary texts which might lead to low quality and unnecessary noises. The popularity of machine translation may interfere with the actual language acquisition process or even reduce the desire to learn a second language (Groves & Mundt, 2015). Prates et al. (2018) noticed that there was a growing concern about the phenomenon dubbed as machine bias which meant that trained statistical models reflect gender or racial bias, so they conducted a study that "built sentences in constructions like 'He/She is an Engineer' (where 'Engineer' is replaced by the job position of interest) in 12 different gender neutral languages such as Hungarian, Chinese, Yoruba, and several other" and then translated these sentences into English in Google Translate, finding that translation tools such as Google Translate showed gender biases and a strong inclination for male defaults.

To sum up, the development of automated translation will not result in the passivation and exclusion of human beings, since translation machines are artifacts based on human creativity and require lifelong learning from human beings.

2.2 Literature Review on Literary Translation

Literary translation is an art. The key challenge for literary translation is that translators have to preserve both the meaning and reading experience. Sometimes the difficulty of translating a literary work is no less than creating one since translators are required to be extremely skilled in using two languages and considerate by taking many factors like cultural differences into account in order to bring readers the best reading experience.

2.2.1 Academic Theories on Literary Translation

Many scholars put their focus on literary translation studies and propose many literary translation theories, trying to find the best way to keep the original beauty of literary works. Many golden rules proposed by famous translation theorists like Eugene A. Nida and Peter Newmark are applied into practice in literary translation, and translation theories exclusively belonging to literature works also emerge in endlessly. Professor Lefevere (2016) whose major research field was in translation and comparative literature study once proposed “Rewriting Theory” to guide literary translation. Kazakova (2015) suggested that in dealing with literary translation which was a specific type with sophisticated structures of information, effective strategies of bilingual information processing including observer-strategy, helper-strategy and enlightener-strategy were needed as attempts to manage the complications.

It is known to all that how to translate literary works has always been a difficulty and the crux lies in the cultural differences, starting from which scholars of every country pull out all the stops to study how to keep the original cultural features of literary works to the maximum extent while ensuring that foreign readers could fully understand and accept the content. Taking Chinese literature as an example, it always stands out as a shining pearl in world literature and is well known for its long history and rich types. Writings in classical Chinese like *The Analects of Confucius*, *The Art of War* and *A Dream in Red Mansions* are all concise in format but profound in significance. To better translate these literary works and promote Chinese culture, great efforts have been made at the academic level. While translating *Evolution and Ethics*, Yan Fu proposed the translation principle of “Faithfulness, Expressiveness and Elegance” which was regarded as the standard of translation (Yan & Zhu, 2017). Chinese renowned scholar Qian Zhongshu emphasized the consciousness of innovation and came up with “Sublimation Theory” in dealing with literary translation (Zheng, 2001). Xu Yuanchong, the celebrated Chinese literature translator, founded the translation theory of “Three Principles of Beauty” (beauty in sense, sound and form) and once stated that Chinese literature translation should be performed by Chinese translators (Xu, 2003 & 2016). Professor Xu Jun proposed “Three Levels (mind, semantics and aesthetics)” to guide Chinese literature translation (Xu, 1989). Although many literature translation theories have been proposed to help facilitating translators' work, literature is still considered to be the most difficult text type to translate or even to be untranslatable. In the process of translating cultural classics, there are always elements that cannot be copied, which is called untranslatability of literary works (Fu, 2019).The main threat to the existence of world literature is untranslatability or rather cultural untranslatability (Sun, 2019). Thus, literature translation, an art of language in need of lots of wisdom and creativity, is seen as the toughest business in translation field and even the most eminent translators cannot guarantee perfect translation.

2.2.2 Machine Translation Attempts on Literary Texts

In recent years, machine translation is undertaking fast development and demonstrating its unique advantages, and many written texts have been put into test getting promising results. Millions of people now use free web-tools like Google Translate every day to satisfy translation needs since machine translation offers strong baseline performance to deal with any input text (Way, 2013).

However, machine translation of literary texts is unacceptable in most people's mind, and most case studies on machine translated literary texts do show negative results. Is there any room for improvement on machine translation of literary texts in the future? A question confuses many researchers. To unveil new insights on literature through a computational approach, Graesser et al. (2011) analyzed literary texts and political speeches with tools Coh-Metrix and Linguistic Inquiry Word Count to calculate linguistic indicators like word concreteness and referential cohesion. Voigt and Jurafsky (2012) examined the expression of referential cohesion in literary and non-literary texts, pointing out that one important direction for future machine translation research is to integrate discourse features above the sentence level. To address what makes literary translation hard and figure out what role machine translation could be, Jones and Irvine (2013) used existing machine translation system to translate literature samples from French into English and analyzed the results from a qualitative perspective, discovering that though machine translation can ensure fluency it cannot handle domestic and foreign ideas. To study the machine translation applicability to literary texts, Toral and Way (2015) once conducted a case study on machine-assisted translation of literary text by training statistical machine translation models on novels and their professional translations, suggesting that machine translation performance had progressed in recent years and translation between related languages would be easier for posting-editing. They continued exploring the machine translation performance on a novel between related languages (Spanish and Catalan) and affirmed the usefulness of machine translation on novels between closely-related languages in accordance with the study results that almost 20% cases translations produced by machine were equal in quality to the ones produced by professional human translator and only 10% cases were in need of at most 5 character edits (2015). To further explore whether machine translation is helpful or not on literary texts, Toral et al. (2018) carried out an experiment in which *Warbreaker* (a novel originally written in English and in need of translating into Catalan) was separately translated by phrase-based machine translation (PBMT) and neural machine translation (NMT) first and then six professional translators translated consecutive fragments of sentences in three alternatives: from scratch, post-editing PBMT and post-editing NMT, finally arriving at the conclusion that both PBMT and NMT contributed to substantial improvements in translation productivity of 18% and 36% respectively (word per hour).

To sum up, scholars have conducted various experiments to assess the machine translation performance on literary texts to figure out whether machine translation is applicable to literature, and if the answer is yes, what should be done next to better facilitate the machine translation work. Although high-quality machine translation cannot be ensured yet, the value and usefulness of machine translation are acknowledged by many researchers since many experiments show that machine translation does increase the translation productivity through post-editing. Hence one essential task facing both developers and translators, as mentioned before, is to always follow up machine translation performance on literary translation to know how far machines have achieved on this translation type and timely update research directions as well as plan for the future.

2.3 Literature Review on Corpus-related Research

Unlike blooming translation theories, translation research methods are usually composed of only experiences summary and case studies. In recent years, some scholars start to employ corpus-related approaches in translation research and more efforts are needed to enrich this research method.

2.3.1 Corpus Research Method in Translation Studies

The corpus-related translation study refers to the way that we employ real bilingual parallel corpora or translational corpora as research objects, take data statistics and theoretical analysis as research methods, and analyze the nature, process and phenomenon of translation. The corpus-based translation study has achieved rapid development in corpus building, language features, translator styles, translation standard and translation teaching with a series of corpuses, publications and seminars successfully launched (Hu & Mao, 2012). In 1993, the British scholar Mona Baker published *Corpus Linguistics and Translation Studies: Implication and Application*, first introducing the corpus-based approach into translation research (Baker, 1993). Mona Baker, then with her team, built the first Translational English Corpus (TEC) in the world and used it to carry out various corpus-based translation studies (Baker, 1995&1996). Scholars from all over the world, represented by Mona Baker, have applied the corpus-based translation research method to conduct diversified empirical studies and theoretical interpretations, which progressively forms a coherent and comprehensive paradigm for translation studies (Laviosa, 1998).

Olohan and Baker (2000) investigated the application of "say / tell that" in translated English corpus and British National corpus, finding that the former is higher than the latter in the usage frequency of "that". Baker (2000) once studied the translation style differences between Peter Bush and Peter Clark (both British translators) from three perspectives of type token ratio, average sentence length and narrative structure respectively, and the study results showed that the type token ratio and average sentence length data results in Clark's translation were lower than that of Bush and in addition Clark tended to use the past tenses and direct speeches more than Bush, and hence it was concluded that Clarks' translation was more concise and clear. Bosseaux (2006) calculated linguistic items like type token ratio and average sentence length in two French translations of Virginia Woolf's *The Waves* to analyze how these two versions dealt with cultural words and proper nouns, and based on the corpus study results she found that one mainly adopted foreignization strategy trying to reproduce the foreign cultural elements as much as possible and the other was prone to apply domestication strategy trying to eliminate the foreign cultural elements in target language text. Besides providing support to the studying of translation feature and translator style, corpus offers help to guide translation teaching as well. Corpora gives full play to the autonomy and flexibility of students or translators since making use of corpora in translation teaching ensures that students are actively involved in the learning process by collecting texts, evaluating corpus, extracting terms and establishing cross-language equivalence of bilingual texts, and the greatest educational value of corpora is to arouse translator's thinking rather than provide ready-made answers (Zanettin et al., 2014).

In terms of corpus-related translation study, domestic scholars even began to concern for it earlier than foreign scholars and increasing attention from domestic scholars was paid in last two decades. As Yang and Bai (2010) summarized, the corpus translation study in China had gone through three stages of Infancy Stage (1993-1998), Slow Development Stage (1999-2005) and Rapid Growth Stage (2006-2009): In the first stage, only four papers concerning the application of corpus in machine translation were published within six years and not a single one referred to translation theory research; From 1999 to 2005, the number of publications slowly increased to 32 and at this stage domestic scholars began to introduce corpus-based translation study methods and applied corpus to conduct empirical research; Since 2006, domestic corpus translation research has entered a rapid growth stage with a large increase in the number of published papers reaching 90 by 2009. Yang (1993) referred to corpus, a collection of authentic language data based on random sampling, as a powerful tool in machine translation. He (2007) carried out a translation corpus design project based on Chinese and English parallel corpora, in which she proposed five specific implementation steps: preprocessing of parallel corpora, identification of multi-word units in the aligned texts, extraction of bilingual translation units, manual validation of bilingual translation units, and system trial and analyses of results. Hu (2009) explored the motivations for the use of BA construction in Liang Shiqiu and Zhu Shenghao's translations of *Hamlet* by Shakespeare following a corpus-based quantitative analysis of the use and distribution of the distinctive Chinese construction". Wang and Hu (2008) studied the distinguishing characteristics of the use of words in the translated Chinese through comparison of the translated Chinese corpus and original Chinese corpus, finding that compared with original Chinese, translated Chinese had the following characteristics: low lexical variability, low lexical density, explicit function words and increased frequency of common words, and based on these discovered characteristics they explored the problem of “Translation Universals”. Scholars the same time were equally concerned for the corpus assistance in translation teaching. Wang et al. (2007), by conducting corpus presentation experiment, explored the application effect of bilingual corpus translation teaching platform and found that students can gain meaningful translation skills after observing the corpus and accordingly evaluate their own translations, thus reaching a conclusion that the use of corpus in translation classroom teaching is conducive to autonomous learning and formation of stable translation skills. The corpus-based research method, a useful and efficient way to explore language features, undoubtedly became the default choice for many scholars and the corpus-based translation study continued to prosper in last ten years. For instance, Both Chen (2009) and Li (2017) applied corpus in their studies of self-translation features. Ran (2019), after reviewing the book *New Directions in Corpus-based Translation Studies*, highlighted three new development trends in corpus-based translation: 1) combination of process-oriented translation studies with product-oriented translation studies; 2) expansion from bilingual study to trilingual or multi-lingual study; 3) diversified research perspectives.

2.3.2 Corpus Use in Linguistic Researches

Such quantitative research method is certainly not only restricted to translation study field. As Graesser et al. (2011) summarized, "during the last two decades there have been revolutionary advances in computational linguistics, corpus linguistics, discourse processes…". The advances greatly facilitate researchers in analyzing mass text data and exploring the characteristics of language. In the report of *What Linguistics Could Do in the Era of Big Data* published by Chinese Social Sciences Weekly, Liu (2017) highly valued the application of big data in linguistic studies. The same year, Huang and Liang (2017), in the report of *Scientization and Internationalization of Linguistic Research* published by Guangming Daily, claimed that with the development of modern linguistics, the linguistic phenomenon should be studied by applying general methods used in natural sciences to achieve the scientificalization of linguistic studies. Many breakthroughs were made by scholars who embraced quantitative linguistics. Professor Liu (2008) chose twenty corpora from different languages with dependency syntactic annotation and proposed that average dependency distance had a tendency to be minimized in human language. A group led by professor Liang conducted quantitative analysis in interpreting studies to "examine the difference in dependency distance among three interpreting types, namely simultaneous interpreting, consecutive interpreting and read-out translated speech, finding that consecutive interpreting texts entail the smallest dependency distance and suggesting that consecutive interpreting bears heavier cognitive demands than the other two" (Liang et al., 2017).

The corpus model, greatly different from the conventional methods, does provide a new way that enables scholars stand from a more comprehensive and concise perspective to conduct translation studies. However, sometimes accurate statistics results cannot reveal the details of original or translated texts. Methods like case analysis are still needed to better evaluate the translation quality. The combination of multiple research methods may bring about unexpected beneficial effects. In consideration of the coexistence of two research approaches, the present study cannot simply be called a corpus-based research. Given that the study first used corpus-based approach to draw preliminary conclusions based on the data results and then carried out case analysis to verify the conclusions, it is more appropriate to refer to the present study as a corpus-driven research.

2.4 Summary

This part presents threes lines of literature review to support the present study.

The first line enumerates a few landmark events in the beginning stage of machine translation and then introduces three development stages of machine translation (“Rule-based Machine Translation”, “Example-based Machine Translation” and “Neural Machine Translation”) in detail by listing some professionals’ studies on each one type. Besides, it particularly emphasizes the fast development of NMT and its star product Google Translate which is to be used in the present study as the machine translation tool. After that are positive and negative comments on machine translation, both of which lead to a unified conclusion that although machines will not replace human translators in the near future due to various concerns, it is undeniable that with the technical development machine translation will move full speed ahead and any study on machine translation is valuable.

The second line starts with academic researches on literary translation home and abroad. Since literary translation is always viewed as a real challenge and harder to be translated than non-literary texts, scholars propose various theories to guide literary translation while some argue that literary texts are untranslatable for human beings, let alone for machines. Then the second section selects a few machine translation attempts on literary texts. As these attempt results show, although high-quality machine translation of literary texts cannot be achieved yet, the machine performance does contribute to literature translation in some degree and is expected to be better.

The third is the literature review on corpus-related translation researches. This section takes a brief look at the origin of corpus research method and then lists a number of researches that use corpora to study translation activities. Besides, other linguistic studies build language models by using corpora of various sizes as well. The corpus research method is certainly an effective tool to assist translation studies.

Based on the above-mentioned three lines, the present study attempts to investigate the machine performance on literary texts by building a small-scale corpus.

3. Research Design

The primary goal of present study is to assess the machine translation performance on literary texts. To better reveal the current development of machine translation on literary works, an assessment on the quality of machine translated non-literary text is conducted for a comparative purpose. Therefore, the first section is about raw materials, namely the original literary and non-literary texts chosen to be studied in both quantitative and qualitative ways. Selected linguistic indicators and research tools are introduced in the second and third sections respectively.

3.1 Raw Materials

The origin texts were extracted from Eileen Chang's novel *Love in A Fallen City* and *Bain China Shopper Report 2019* respectively representing literary and non-literary texts. Both were written in Chinese and prepared to be translated into English in machine translation.

Eileen Chang, known as one of the most talented and influential female writers in the period of Republic of China, created many classical novels and short stories in her lifetime, all of which feature plenty of detail description and rhetoric expression. Her works like *Red Rose White Rose, Half a Lifetime Romance and Love in A Fallen City* have been translated into many languages and introduced into different countries, a clear demonstration that her literary achievements and works are widely recognized by both inside and outside the industry. The literary fiction *Love in A Fallen City*, one of Eileen Chang's celebrated works, tells a love story set in 1940s Shanghai and wartime Hong Kong. Besides many embedded rhetorical devices like metaphors, personifications, repetitions, double meanings and antitheses, this novel contains a great deal of psychological monologues and Chinese idioms. Although translating such a literary work is an awfully tricky business, all these efforts would pay off in the end since the novel *Love in A Fallen City* is one of the most outstanding Chinese short novels and worthy of being introduced to readers in different countries. As an academic scholar Nicole Huang suggested, "Eileen Chang's most important legacy from the 1940s is her construction of an alternative narrative of war, one that contradicted the grand narratives of national salvation and revolution that dominated the wartime literary scene". How to better translate this type of literary work would be a research problem of great value. The novel *Love in a Fallen City* was used to provide 45 randomly-chosen paragraphs as original literary texts for further study.

Bain & Company is a leading global management consulting firm committed to deliver advises on issues such as private equity investments, mergers & acquisitions, corporate strategy, finance, operations and market analysis. Bain & Company and Kantar Worldpanel have cooperated for eight years on tracking the shopping behaviors of Chinese consumers in order to set a valuable long-term view across 106 fast-moving consumer goods categories purchased for home consumption in China and the report *Bain China Shopper Report 2019* is the latest published survey result. The report's writing is clearly structured by applying induction and deduction methods and the content focuses on data presentation and facts statement. 45 paragraphs were selected randomly from the newly issued *Bain China Shopper Report 2019* as non-literary texts and then put into Google Translate.

The total number of paragraphs used for analysis is 180, with 45 paragraphs of original literary text, 45 paragraphs of non-literary text, 45 paragraphs of machine translated literary text, and the last 45 paragraphs of machine translated non-literary text. An overview of raw materials is shown in Table 3.1:

Table 3.1 Overview of Original Texts and Machine Translated Texts

Language	Type	Content	Number	No.
Chinese	Literary	Love in A Fallen City	45	1.1 - 1.45
English	Literary	Love in A Fallen City	45	2.1 - 2.45
Chinese	Non-literary	Bain China Shopper Report 2019	45	3.1 - 3.45
English	Non-literary	Bain China Shopper Report 2019	45	4.1 - 4.45

3.2 Linguistic Indicators

To obtain an objective and quantified view on machine translation performance on both literary and non-literary texts, it is important to first define representative indicators so that formal empirical studies can follow on. Three indicators Type Token Ratio (TTR), Content Words Density and Average Sentence Length were calculated in the present study. They were chosen to reveal the linguistic features of research texts.

3.2.1 Type Token Ratio (TTR)

Type Token Ratio is a time-honored linguistic indicator that measures the richness and diversity of words used in texts. A token is any instance of a particular word form while type refers to a particular unique word form. In other words, tokens refer to units (word occurrences) which form the sample text and types refer to lexical units which form the vocabulary in the sample (Holmes, 1985). At the most basic level, if we count the total number of words in a text as 100, the number of words is often referred to as the number of tokens, and if in the text exists words repetition like “the” repeating ten times and “of” five times, there are 85 so-called types in this text. The result is often multiplied by 100 to express the TTR as a percentage.

In theory, TTR weights range of vocabulary for text sample size, and the larger the resulting TTR value is, the less repetitive the vocabulary usage is (Richards, 1987). Many researches that focus on language development and text analysis apply Type-Token Ratio as an important indicator of words richness. To conduct a corpus-based probe into translational language features of literary self-translation, Li and Li (2017) built a Chinese English Self-Translation Parallel Corpus and calculated various linguistic indicators including Type-Token Ratio. In a study that analyzes how communication changes when people communicate with a chatbot as opposed to another human, one of the conclusions is that human-chatbot communication lacks much of the vocabulary richness found in conversations among people in accordance with the calculation result that human-chatbot messages have a significantly smaller type-token ratio than human-human messages (Hill et al., 2015). To study the difference between Hillary Clinton’s and Donald Trump’s linguistic styles, researchers adopt a corpus-based approach including the calculation of Type Token Ratio to compare their campaign speeches during 2016 U.S presidential election and based on one result that both Type-Token Ratio and Standard Type-Token Ratio of the Clinton corpus outweigh that of the Trump corpus, it is suggested that Clinton has a more diverse vocabulary than Trump (Chen et al., 2019).

However, there is one problem that TTR is affected by the sample text length since the longer the text is, the more likely that the same words will occur again (Covington & McFall, 2010). Once a sample is large enough to have included all active vocabulary of one field, any further sampling of tokens would only lead to a decline in TTR, which is a tough problem that has distorted many research findings (McKee et al., 2000). In view of the adverse effect caused by different corpora sizes, only similar-sized corpora can be compared in using TTR indicator. Thus, the size of each selected Chinese paragraph was controlled between 200 and 250 words.

Calculation Formula:

$$Type\ Token\ Ratio(TTR) = \frac{total\ types}{total\ tokens} * 100$$

3.2.2 Content Words Density

In translation studies, Content Words Density is defined as the percentage of content words in the text and serves as an important indicator for analyzing lexical density. For English, content words usually include nouns, adjectives, adverbs and main verbs while function words include auxiliary verbs, modal verbs, pronouns, propositions, determiners and conjunctions (Stubbs, 2001). For Chinese, content words include nouns, verbs and adjectives while function words include adverbs, pronouns, prepositions, conjunctions, auxiliary words, interjections and the like (Wang, 2012). In linguistics, content words are words that name objects of reality and their qualities, contrasting to function words which are words that have very little substantive meaning and primarily denote grammatical relationships between content words. The distinction is clearly explained by Michael Stubbs in the book *Words and Phrases: Corpus Studies of Lexical Semantics:*

"In English and in many other languages, there is a distinction which divides the whole vocabulary into two major categories: content words tell us what a text is about, and function words relate content words to each other…Content words are also referred to as major, full and lexical words. They carry most of the lexical content, in the sense of being able to make reference outside language. Function words are also referred to as minor, empty, form, structural and grammatical words. They are essential to the grammatical structure of sentences". (Stubbs, 2001, p39-40).

As Laufer and Nation (1995) mentioned in their paper, "since lexical words are the words which primarily convey information, a text is considered 'dense' if it contains many lexical words relative to the total number of words". Concrete and meaningful contents words evoke mental images as opposed to abstract words (Grasser et al., 2011). In general, more content words deliver more information and emotion. Thus, the incidence score of content words is an important linguistic index because it catches a glimpse of how much substantive content there is in the sample text.

Calculation Formula:

$$Content\ Words\ Density = \frac{total\ content\ words}{total\ words}$$

3.2.3 Average Sentence Length

Average Sentence Length is a characteristic of author style and it is computed as the number of tokens in corpus divided by the number of sentences in corpus. Yule (1939) proposed sentence length as a particular method to define authorship and did motivate many people to study the definition and application of “sentence length” for statistic work. In the present study, English sentences are defined to be separated from each other by punctuation marks [.], […], [?] and [!], while Chinese ones are separated by [。], [……], [?] and [!]. Smith (1983) conducted sentence length analysis with known facts of authorship in three cases, finding that the statistic outcomes of sentence length in each case confirmed the results derived by other methods. Kelih et al. (2006) applied a quantitative approach to the research topic of texts classification by considering sentence length as a decisive discriminating factor. The sentence length itself obviously plays an important role in identifying specific text characteristics and meantime it is usually combined with other quantitative measures for further research. When studying the stylistic features of sample texts from a quantitative perspective, sentence length is certainly the one to be taken into account (Pastor et al., 2008). Toral and Way (2018), to assess the quality attainable for novels by NMT, analyzed three features of novels: lexical richness, average sentence length and degree of novelty in regard to the training data. In Coh-Metrix, a computer tool that analyzes texts on cohesion, language and readability, word length and sentence length are viewed as two simple but crucial indicators because standard text readability formulas assess texts on difficulty by relying on them (Graesser et al., 2004). This study calculated average sentence length of every paragraph in chosen texts and its according translated paragraph to see the stylistic features of original ones and whether there is any change in terms of author style after being processed by machine translation.

Calculation Formula:

$$Average\ Sentence\ Length = \frac{total\ words}{total\ sentences}$$

3.3 Research Tools

The research tools in the present study mainly refer to machine translation tool Google Translate as well as linguistic indicators calculation tool. This section gives detailed introdcution to these research tools.

3.3.1 Google Translate

In the present study, Google Translate was used as the machine translation tool to process both literary and non-literary texts. Google Translate is a free multilingual machine translation service developed by Google company, which uses deep learning techniques and develops a system called "Neural Machine Translation (NMT)" to ensure greater accuracy of translated texts. The statistics-based translation system Google Translate would calculate probabilities of various translations of a phrase being correct instead of word-for-word translation (Groves & Mundt, 2015). As Wu et al. (2016) explained, "the strength of NMT lies in its ability to learn directly, which is an end-to-end fashion and a mapping from input text to associated output text".

Since the advent of Google Translate, scholars have conducted theoretical and empirical researches using Google Translate as study subject or supplementary research tool in various research fields. Si et al. (2016) reviewed the technical background of Google Translate and discussed the future work in machine translation field in which they concluded that the success of Google Translate lay in its ingenious transition from practical problems into statistics ones as well as its powerful computing capability, but Google Translate only used a small amount of information in Big Data which was a problem to be addressed in the future. To address the issue that researchers who are interested in cross-country comparisons by conducting automated text analysis may get lost in texts of different languages, some analysts suggest to translate all texts into English in Google Translate before starting the analysis (Lucas et al., 2015) and to test the reliability of this method, De Vries et al. (2018) used the *europarl* dataset and then compared term-document matrices with topic model results from high quality translated text and machine-translated text to assess the usefulness of machine translation for bag-of-words models and the results showed that Google Translate was an effective tool for comparative questions. In a famous analysis conducted by Columbia University that built three types of models to classify "tweets" into positive, negative and neutral sentiment, the research group used Google Translate to convert tweet data in foreign languages into English before the manually annotation process, which helped save a lot of time and energy (Agarwal et al., 2011).

As listed above, more and more studies turn to focus on Google Translate performance or its practical application, and many show positive results. Thus, in consideration of the superiority of Google Translate itself and a great deal of available thoughts relating to Google Translate accumulated by forerunners, the present study was designed to continue using Google Translate to output machine translated texts for later analysis.

3.3.2 Calculation Tools

To conduct the Pearson Correlation analysis between original and translated versions, indicators of both Chinese and English texts are needed. However, most linguistics softwares in the market only serve for calculating English text indicators and till today there is no well-developed Chinese linguistics software available to satisfy the research requirement. Such phenomenon is common for languages other than English. Hence, this study applied online natural processing language (NLP) products Natural Language Toolkit (NLTK) and BosonNLP to do word segmentation and POS tagging.

NLTK is a suite of open source libraries and programs for symbolic and statistical natural language processing for English based on an object-oriented scripting language that supports rapid prototyping and literate programming, which provides a uniform and extensible framework for projects, class demonstrations and assignments (Loper & Bird, 2002). The purpose of creating such a platform, as its developers explain, is to allow computational linguists to build NLP components and systems in an easier way, and currently NLTK has been widely accepted by many US universities as a research tool. BosonNLP is a Chinese NLP platform capable of handling sentiment analysis, classification, named entity recognition, keyword extraction and the like with 97.7% accuracy rate and 98.1% recall rate. Word segmentation and POS tagging lay down the foundation of Chinese natural language processing (Wong et al., 2009). To deal with difficulties facing Chinese NLP like high rate of out-of-vocabulary and sentences containing informal words, Min et al., (2015) proposed "an ensemble approach for word segmentation and POS tagging, which combines both discriminative and generative methods to get the advantage of both sides" and the solution has been released and renewed at www.bosonnlp.com for developers and researchers to use for free.

Upon the basis of word segmentation and POS tagging, the numbers of tokens, types, sentences and content words were computed in a program in the python language to calculate indicators of Type Token Ratio (TTR), Content Words Density and Average Sentence Length. Python is one of the fastest growing languages and always preferred by scientists and developers in data analytics since its ever-evolving open source libraries meets the need of Data analytics (Nagpal & Gabrani, 2019).

Based on calculated indicators data, a Pearson Correlation analysis was carried out for later contrastive analysis. The statistics software SPSS 20.0 was applied to run bivariate Pearson correlation. Pearson correlation coefficient is the covariance of the two variables divided by the product of their standard deviations, which is a well-established measurement of correlation ranging from + 1 to -1 with 0 referring to the absence of any relationship. The value of the correlation coefficient is + 1 if the two variables have a complete positive linear correlation, whereas - 1 means a complete negative linear correlation (Chung & Lee, 2001), and the greater the absolute value of correlation coefficient is, the stronger the correlation is. Many scholars claim that they find a strong relationship between what they test, but on most cases, the significance is improperly reported instead of the correlation strength (Akoglu, 2018). The SPSS output of Pearson correlation coefficients includes parameters "r" and "p", in which p-value shows the probability that the result may occur by chance and r value refers to the correlation strength. The correlation strength is usually divided into five levels as following:

0.0 – 0.2	0.2 – 0.4	0.4 – 0.6	0.6 – 0.8	0.8 -1.0
Very Week	Weak	Moderate	Strong	Very Strong

Figure 3.1 Pearson Correlation Strength Classification

And the equation to compute the Pearson's correlation coefficient r for variables x and y is as follows:

$$r = \frac{N\sum xy - \sum x\sum y}{\sqrt{(N\sum X^2 - (\sum x)^2)(N\sum y^2 - (\sum y)^2)}}$$

3.4 Summary

In order to make comparison between machine translation performance on literary and non-literary texts, the present study selected 45 paragraphs (each from 200 to 250 words) of the original literary text *Love in A Fallen City* and 45 paragraphs (each from 200 to 250 words) of the original non-literary text *Bain China Shopper Report 2019,* and then put these chosen paragraphs into Google Translate to obtain 90 machine translated paragraphs for later analysis. To conduct a quantitative analysis, three linguistic indicators Type Token Ratio, Content Words Density and Average Sentence Length were calculated in a python program and then the calculated data were put into SPSS 20.0 to carry out Pearson Correlation analysis.

4. Data Analysis and Results

This chapter presents calculation results of three linguistic indicators: Type Token Ratio (TTR), Content Words Density and Average Sentence Length. Data and figures that reflect the differences between machine translation performance on literary and non-literary texts are given below.

4.1 Results of Type-Token Ratio (TTR)

The input TTR values of literary and non-literary texts are at the same level (M = 70.746 and M = 71.205 respectively), which avoids the possible influence caused by original textual differences. The TTR outputs of literary and non-literary texts are 62.686 and 65.307 respectively, both much lower than inputs. The TTR calculation results of each paragraph and its machine translation are shown in two scatterplot diagrams below. On the basis of indicators data, the Pearson coefficients were computed between each pair (pair 1: literary text and its machine translation; pair 2: non-literary text and its machine translation). As shown in Table 4.1, in terms of the Type Token Ratio, significant positive relationships are found in both correlation researches, namely between literary text and its Google translation ($r = 0.534$, $p = 0.000$) as well as between non-literary text and its Google translation ($r = 0.757$, $p = 0.000$). The literary pair ($r = 0.534$) has a moderate positive relationship as shown by Figure 4.1 and the non-literary pair ($r = 0.757$) has a strong positive relationship as shown by Figure 4.2. It is clear from these two scatterplot diagrams that TTR values of both original literary and non-literary texts and that of machine translated versions increase and decrease together. The decrease in TTR values of both texts may illustrate that whether translating literary text or non-literary text machine is inclined to reduce the vocabulary richness.

Figure 4.1 TTR Calculation Data for Literary Text

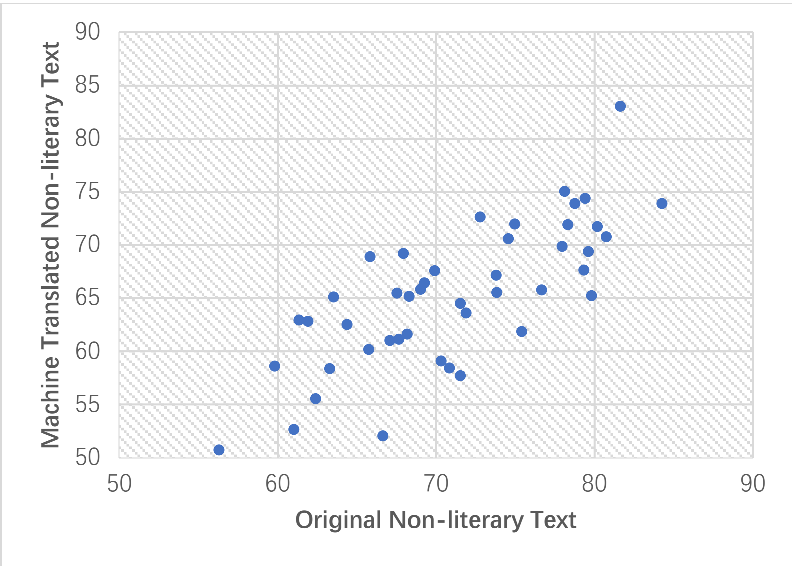

Figure 4.2 TTR Calculation Data for Non-literary Text

Table 4.1 Pearson Correlation Coefficients for TTR Values

	1.1 – 1.45	2.1-2.45	3.1-3.45	4.1-4.45
mean	70.746	62.686	71.205	65.307
r		0.534		0.757
p		0.000		0.000

4.2 Results of Content Words Density

The content words density of source literary text (M = 0.709) is similar to that of source non-literary text (M = 0.797), which diminishes the possible influence by machines translating different difficulty-level texts, while the figures of translated literary and non-literary texts are 0.593 and 0.691 respectively. The specific Content Words Density values for each paragraph and its machine translation are marked in Figure 4.3 and Figure 4.4. It is seen from Table 4.2 that a significantly moderate correlation is found between the non-literary text and its machine translation (r= 575, p = 0.000), and no significant correlation is confirmed between the sample literary text and its machine translation (r = 150, p = 0.324). The machine translated literary text (M = 0.593) and non-literary text (M = 0.691) may be considered "thinner" than their source texts (M = 709 and M = 0.797 respectively) because the latter contain a lower proportion of concrete and meaningful words.

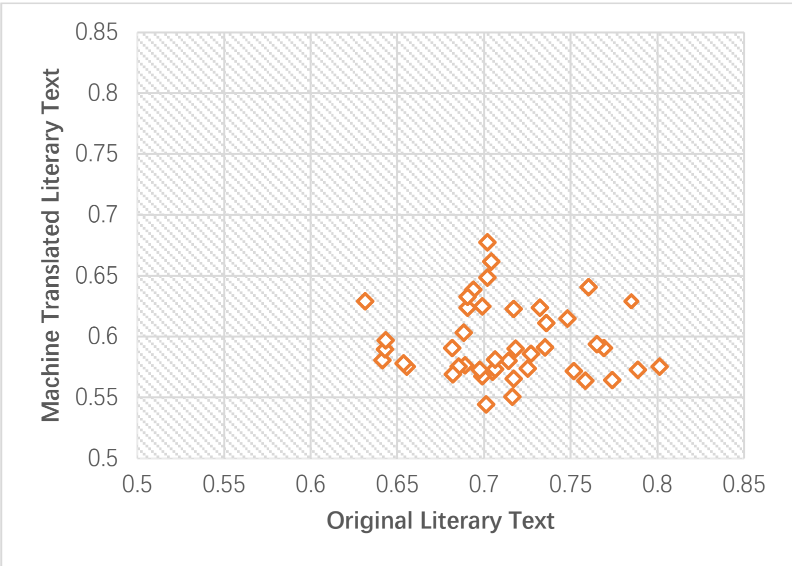

Figure 4.3 Content Words Density Calculation Data for Literary Text

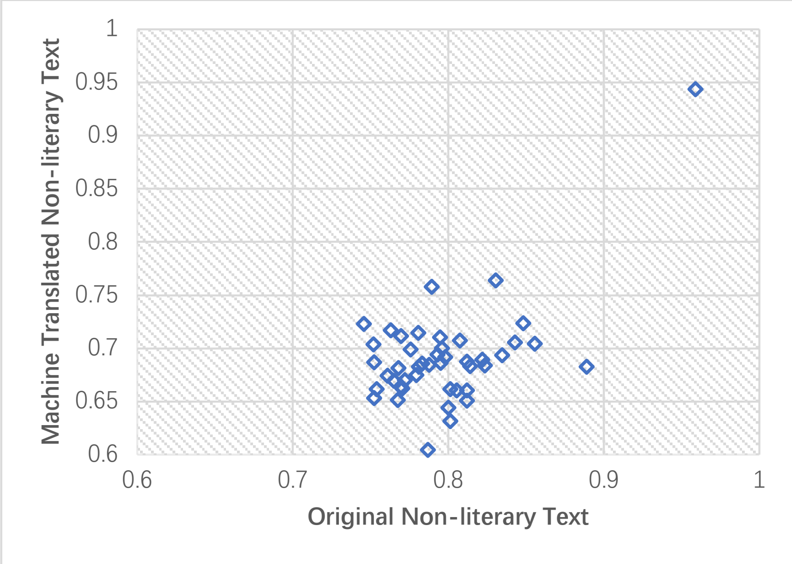

Figure 4.4 Content Words Density Calculation Data for Non-literary Text

Table 4.2 Pearson Correlation Coefficients for Content Words Density Values

	1.1 – 1.45	2.1-2.45	3.1-3.45	4.1-4.45
mean	0.709	0.593	0.797	0.691
r		0.150		0.575
p		0.324		0.000

4.3 Results of Average Sentence Length

The computational results of Average Sentence Length are shown in Table 4.3. The same as the above two linguistic indicators, the average sentence length values of both chosen original texts are on the same level (M = 22.798 and M = 22.566 respectively). The intention of using controlled sentence length is in consideration of a suggestion made by Jiang and Liu (2015) that "it is more desirable to use a parallel corpus with controlled sentence length, same genre and similar syntactic annotation schemes as we did…English and Chinese at great length". The average sentence lengths of chosen paragraphs and corresponding translated ones are shown in following two scatterplot diagrams. A significantly very strong relationship exists between the non-literary text and its machine translation (r = 0.824, p = 0.000), while no significant correlation is found in the pair of literary text (r = 0.178, p = 0.242). As Figure 4.6 shows, there is a clear positive linear correlation in the non-literary pair. Overall, the gap between the values of literary text and its machine translation (M = 22.798 and M = 14.584 respectively) shows that machine intends to simplify the sentences structure when dealing with literary text, and similar values of non-literary text and its machine translation (M = 22.566 and M = 24.802 respectively) reflect that machine retains the writing style to the maximum extent when dealing with non-literary text in addition to having a significantly high positive correlation in average sentence length.

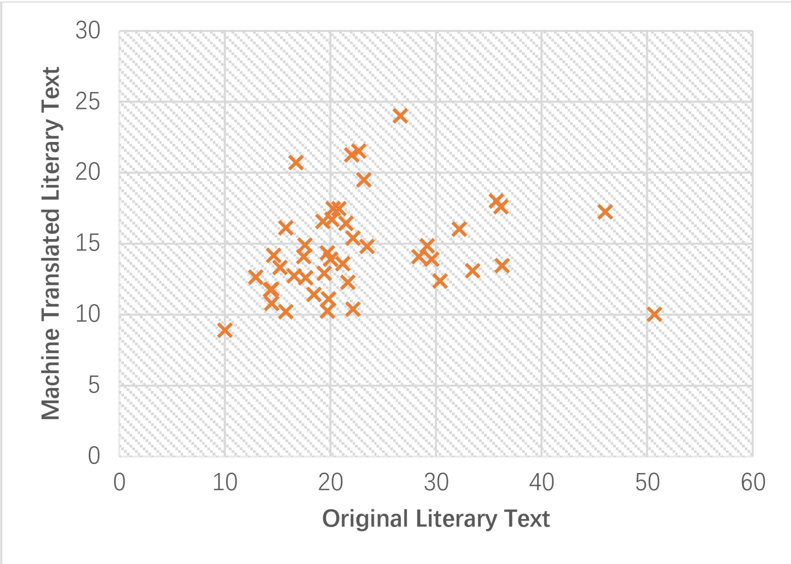

Figure 4.5 Average Sentence Length Calculation Data for Literary Text

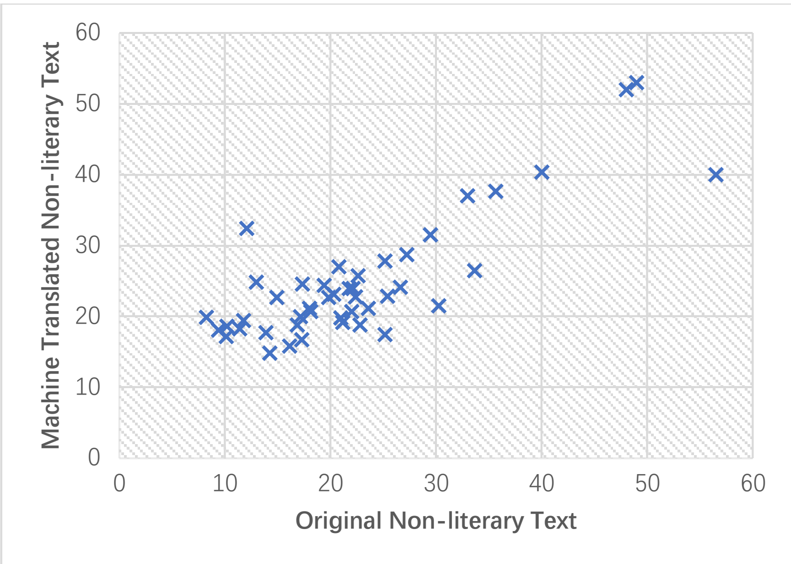

Figure 4.6 Average Sentence Length Calculation Data for Non-literary Text

Table 4.3 Pearson Correlation Coefficients for Average Sentence Length Values

	1.1 – 1.45	2.1-2.45	3.1-3.45	4.1-4.45
mean	22.798	14.584	22.566	24.802
r		0.178		0.824
p		0.242		0.000

4.4 Summary

This chapter first lists the calculated values of both literary and non-literary texts and according machine translated texts, and then shows the Pearson correlation coefficients between each pair (pair 1: literary text and its machine translation; pair 2: non-literary text and its machine translation). For the literary pair, a significantly moderate correlation is found in terms of TTR while no significant correlation is found in either Content Words Density or Average Sentence Length. For the non-literary pair, significant correlation is found in all three indicators with TTR in strong level, Contents Words Density in moderate level and Average Sentence Length in very strong level.

5. Case Studies

This chapter selects five paragraphs from literary text and five paragraphs from non-literary text as well as their corresponding google translations to conduct case analysis. As mentioned in the Introduction part, although it is a corpus-driven study, case analysis is sorely needed as an important single part. The data results concerning Pearson correlation analysis do provide some insights on whether machine translation can ensure reliable and consistent performance. Besides, the data results of each linguistic parameter are helpful in observing how stylistic features and writing style of texts vary after being processed by machine translation. Although data could provide powerful evidence for general patterns, sometimes they are limited in convincing readers of this type of general pattern. The translation quality can never only be judged by quantitative data results. To dive deeper into the machine translation performance on literary and non-literary texts, a few cases are listed below to see how these three indicators results change on a single paragraph and how machine translation deals with specific words and sentences. Comment and evaluation on machine translation quality are given at the end of each case analysis.

5.1 Literary Text Cases and Machine Translations

Case 1:

Source Text: 上海为了“节省天光”，将所有的时钟都拨快了一小时，然而白公馆里说：“我们用的是老钟，”他们的十点钟是人家的十一点。他们唱歌唱走了板，跟不上生命的胡琴。胡琴咿咿哑哑拉着，在万盏灯的夜晚，拉过来又拉过去，说不尽的苍凉的故事——不问也罢！……胡琴上的故事是应当由光艳的伶人来搬演的，长长的两片红胭脂夹住琼瑶鼻，唱了、笑了，袖子挡住了嘴……然而这里只有白四爷单身坐在黑沉沉的破阳台上，拉着胡琴。正拉着，楼底下门铃响了。这在白公馆是一件稀罕事，按照从前的规矩，晚上绝对不作兴出去拜客。

Machine Translated Text: Shanghai in order to "save sky", all the clocks are set forward an hour, but the white male shop, said: "We use the old bell," they are people of eleven ten o'clock. They sang and sang the board and couldn't keep up with the life of Hu Qin. Hu Qin was dumb and dumb, on the night of the lantern, pulled over and pulled over, telling the endless story of desolation - no matter what! ... Hu Qin’s story should be performed by a glorious monk. The long two pieces of red rouge sandwiched Qiong Yao's nose, sang and laughed, and the sleeves blocked his mouth... However, only Bai Siye sat alone in the dark. On the broken balcony, pull the huqin. Pulling, the doorbell rang at the bottom of the building. This is a rare thing in white public hall, according to the former rules, go out at night is definitely not for Xing Outback.

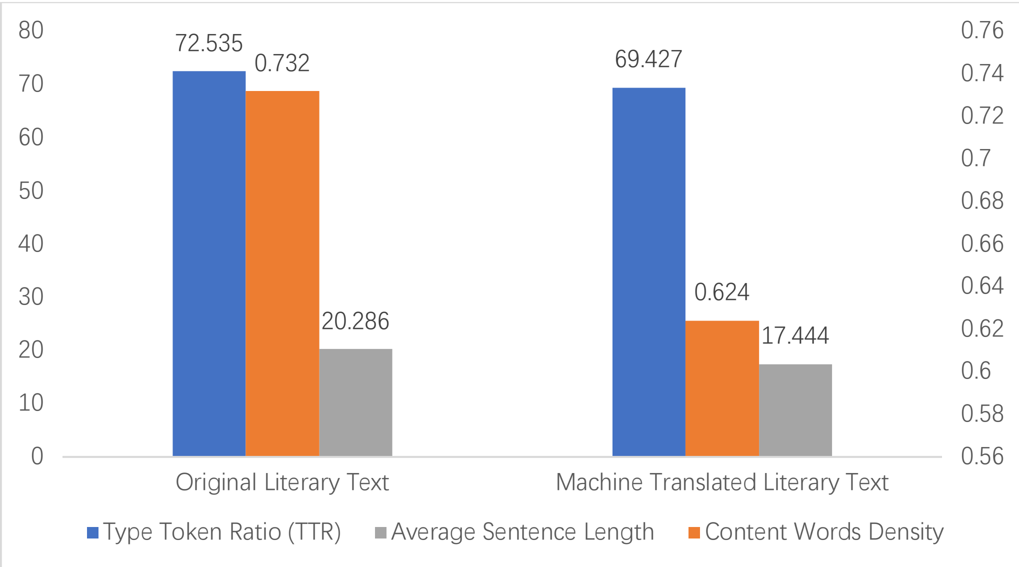

Figure 5.1 Calculation Data for Literary Text Case 1

The bar chart above presents three linguistic indicators of a single paragraph and its machine translated version. All three indicators data of machine translated version show a drop down trend compared with that of original paragraph (TTR: 69.427 < 72.535, Average Sentence Length: 17.444 < 20.286, Content Words Density: 0.624 < 0.732), which indicates that after being processed by machine translation, the words richness and lexical density decrease and sentence structure becomes simpler. When dealing with concrete words, the computer seems to be in a big mess. For instance, the phrase "节省天光" in the first line originally means shifting to daylights saving time, but machine translation cannot understand such literary expression and literally translates this phrase into "save sky" which definitely leaves the readers wondering in what way the author wants to save the sky. Besides, facing unique vocabularies in Chinese culture like "白公馆 (white residence)", "伶人 (opera actor or actress)" and "琼瑶 (precious jade)", because of lacking relevant Chinese cultural knowledge input, Google Translate who turns them into "white male shop", "monk" and "Qiong Yao" makes a spectacle of itself. Besides, correctly understanding complete sentences written in a literary way is a difficult task for computers as well. In Case 1, a complete Chinese sentence (underlined above) is divided into two parts with the last one lacking a subject, which is an obvious grammatical error that could never be accepted by human translators. As seen from these selected parts, the machine translated quality of literary text is certainly unsatisfactory.

Case 2:

Source Text: 屋里没有灯，影影绰绰的只看见珠罗纱帐子里，她母亲躺在红木大床上，缓缓挥动白团扇。流苏走到床跟前，双膝一软，就跪了下来，伏在床沿上，哽咽道："妈。"白老太太耳朵还好，外间屋里说的话，她全听见了。她咳嗽了一声，伸手在枕边摸索到了小痰罐子，吐了一口痰，方才说道："你四嫂就是这样碎嘴子，你可不能跟她一样的见识。你知道，各人有各人的难处，你四嫂天生的强要性儿，一向管着家，偏生你四哥不争气，狂嫖滥赌，玩出一身病来不算，不该挪了公账上的钱，害得你四嫂面上无光，只好让你三嫂当家，心里咽不下这口气，着实不舒坦。

Machine Translated Text: There was no light in the room, and the shadows only saw the ball in the jewels. Her mother was lying on the big mahogany bed and slowly waving the white fan. The tassel walked to the front of the bed, and the knees were soft. He fell down and fell on the edge of the bed. He whimpered: "Mom." The old lady's ears were fine. She said everything in the outside room. She coughed and reached out to find a small pot on the pillow and spit a sip. Then she said, "You are so broken, you can't have the same knowledge as her. You know, everyone has their own Difficulties, you are born with a strong personality, have always been in charge of home, partial to your four brothers do not live up to expectations, arrogant gambling, play a disease is not counted, should not move the money on the public account, hurt you There is no light on the four faces, so I have to let you go to the top of my house. I can't swallow this breath in my heart. It is really uncomfortable.

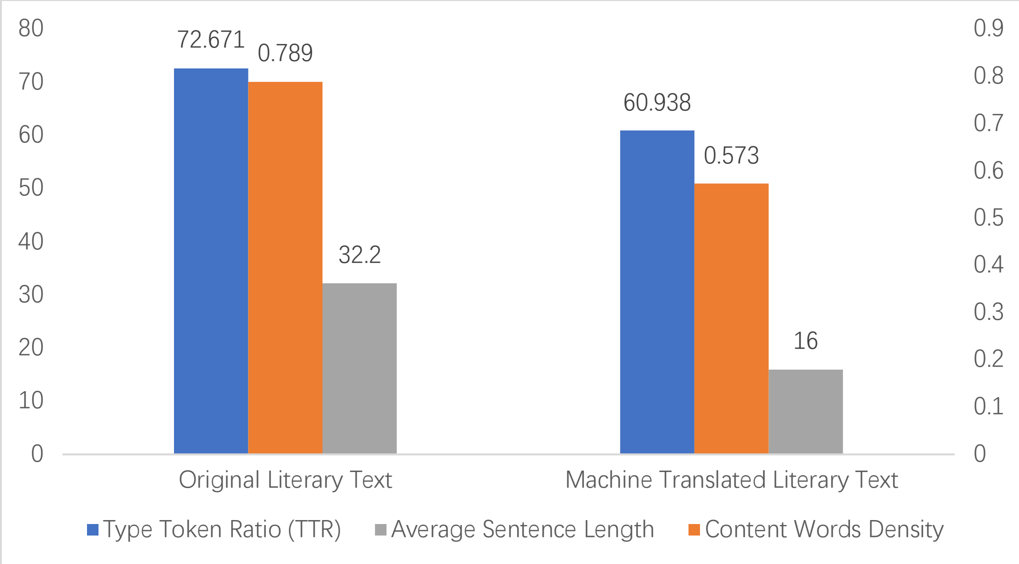

Figure 5.2 Calculation Data for Literary Text Case 2

As shown from Figure 5.2, the TTR, Average Sentence Length and Content Words Density values of the original paragraph (72.671, 32.2 and 0.789 respectively) are obviously higher than that of its machine translation (60.938, 16 and 0.573 respectively), which reveals that to some degree machine decreases the lexical richness and sentence structure complexity. Then, a detailed look at translated words and sentences reflects that there are misunderstanding and omissions. The Chinese idiom "面上无光 (feel ashamed)" is literally translated into "no light on the face" and the content words that are mentioned several times in the original paragraph "四嫂（wife of the fourth elderly brother)" and "三嫂(wife of the third elderly brother)" are all missing in the translated version. As for "流苏(Liusu)" the leading female character name that appears thousands of times in this literary novel, machine just translates it into "tassel" without being aware of how special and important it is at all. According to the underlined sentence and its machine translation, machine translation prefers cutting long Chinese sentence into shorter ones, which might be one reason for shortened average sentence length. Besides of errors like misunderstanding and omissions, computers often make grammar mistakes and thus cannot ensure smooth output when dealing with literary text. The machine translation quality of Case 2 is unable to meet readers' expectation.

Case 3:

Source Text: 为了宝络这头亲，却忙得鸦飞雀乱，人仰马翻。一样是两个女儿，

一方面如火如荼，一方面冷冷清清，相形之下，委实使人难堪。白老太太将全家的金珠细软，尽情搜括出来，能够放在宝络身上的都放在宝络身上。三房里的女孩子过生日的时候，干娘给的一件巢丝衣科，也被老太太逼着三奶奶拿了出来，替宝络制了旗袍。老太太自己历年攒下的私房，以皮货居多，暑天里又不能穿着皮子，只得典质了一件貂皮大袄，用那笔款子去把几件首饰改镶了时新款式。珍珠耳坠子、翠玉手镯、绿宝戒指，自不必说，务必把宝络打扮得花团锦簇。

Machine Translated Text: For the sake of Baoluo's relatives, but they are so busy that they are flying around. The same is the two daughters, on the one hand, in full swing, on the one hand cold and clear, in contrast, it is really embarrassing. Mrs. Bai's golden pearls of the whole family are soft and easy to find, and those who can be placed on Baoluo are placed on Baoluo. When the girl in the three-bedroom had a birthday, a nest of silk clothes was given by the mother, and the old lady forced the grandmother to take it out and made a cheongsam for Baoluo. The old lady's own private house has been in the past for many years. She can't wear leather in her summer days. She only has a sable skin, and used the money to change a few pieces of jewelry into a new style. Pearl earrings, jade bracelets, and green treasure rings, you don't have to say, you must dress up Baoluo.

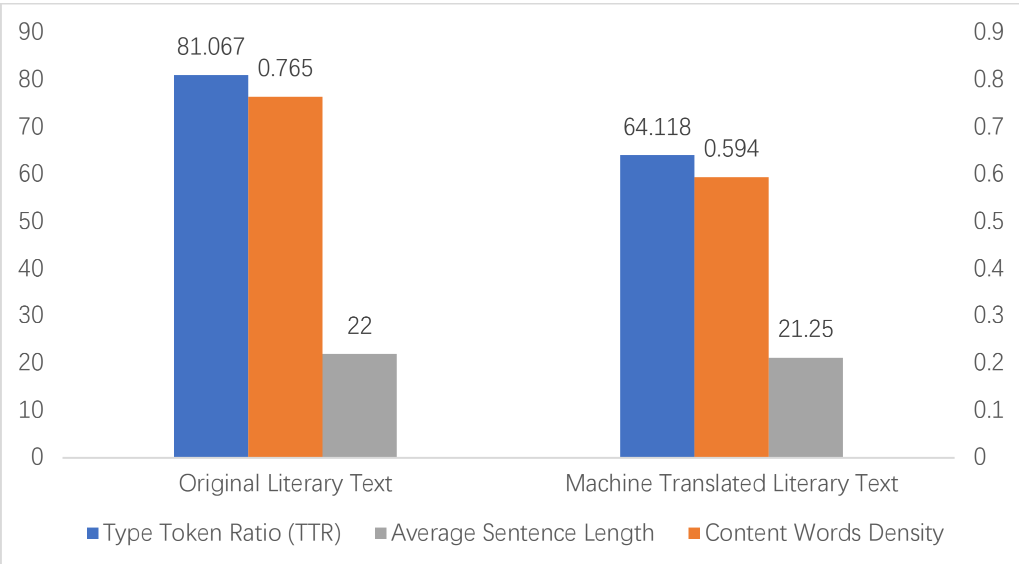

Figure 5.3 Calculation Data for Literary Text Case 3

A similar trend is found in the figure for Case 3, in which all indicators results of machine translated literary text are lower than that of original one and TTR of original literary text (81.067), in particular, markedly declines to 64.118. It is noticed that Chinese metaphorical idioms like “鸦飞雀乱”,“人仰马翻”, and “如火如荼” are beyond computer’s apprehension, so that it cannot clearly describe the images shown by these idioms but has to simplify or even neglect them, which would definitely lead to decreasing lexical density. The machine translation of items like “珍珠耳坠子”,“翠玉手镯” and “绿宝戒指” is accurate. The underlined long sentence in the original paragraph is divided into three. In general, although the machine performance on specific terms is acceptable, the messy grammatical structure and incorrect translation of most Chinese idioms leave many sentences meaningless.

Case 4:

Source Text: 徐太太徐先生带着孩子一同乘车来接了她上船，坐的是一只荷兰船的头等舱。船小，颠簸得厉害，徐先生徐太太一上船便双双睡倒，吐个不休，旁边儿啼女哭，流苏倒着实服侍了他们好几天。<u>好容易船靠了岸，她方才有机会到甲板上看看海景，那是个火辣辣的下午，望过去最触目的便是码头上围列着的巨型广告牌，红的、橘红的、粉红的，倒映在绿油油的海水里，一条条，一抹抹刺激性的犯冲的色素，窜上落下，在水底下厮杀得异常热闹。</u>流苏想着，在这夸张的城市里，就是栽个跟斗，只怕也比别处痛些，心里不由得七上八下起来。

Machine Translated Text: Mrs. Xu, with her child, took the bus and took her to the ship. She was in the first class of a Dutch ship. The boat was small and bumpy. Mrs. Xu’s husband fell asleep as soon as he got on the boat. He spit incessantly, and the niece cried next to him. The tassels served them for several days. <u>It’s easy to get to the shore, and she has a chance to see the sea on the deck. It’s a hot afternoon. The most striking thing in the past is the giant billboards on the pier, red, orange and pink. Reflected in the green waters of the sea, a strip of scented irritating pigments, smashed up and down, killing under the water is very lively.</u> The tassel thought, in this exaggerated city, it was to plant a fight, I am afraid that it hurts more than anywhere else, and my heart could not help but get up and down

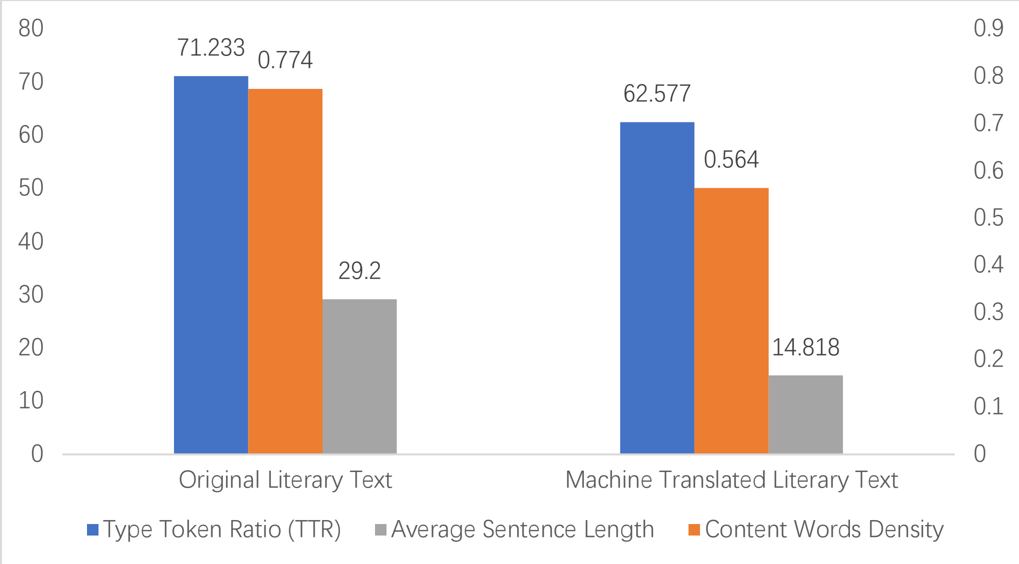

Figure 5.4 Calculation Data for Literary Text Case 4

The figure above shows three declining indicators values with TTR value from 71.233 to 62.577, Average Sentence Length value from 29.2 to 14.818 and Content Words Density value from 0.774 to 0.564, which is a clear demonstration that for this paragraph the words and sentences structure have been simplified by machine translation. The underlined long sentence of the original Chinese paragraph is divided into four sentences in which the last one is a bit more confusing because it is difficult for machines to identify subjects and objects. Besides, several translations seem to be quite absurd such as "栽个跟斗 (to suffer a setback)" being translated into "plant a fight". Some machine translated sentences of this paragraph are complete and correct while some still need extensive modification for finally achieving high quality output.

Case 5:

Source Text: 柳原靠在墙上，流苏也就靠在墙上，一眼看上去，那堵墙极高极高，望不见边。墙是冷而粗糙，死的颜色。她的脸，托在墙上，反衬着，也变了样——红嘴唇、水眼睛、有血、有肉、有思想的一张脸。柳原看着她道："这堵墙，不知为什么使我想起地老天荒那一类的话。……有一天，我们的文明整个的毁掉了，什么都完了——烧完了、炸完了、坍完了，也许还剩下这堵墙。流苏，如果我们那时候在这墙根底下遇见了……流苏，也许你会对我有一点真心，也许我会对你有一点真心。"

Machine Translated Text: Liuyuan leaned against the wall, and the tassels leaned against the wall. At first glance, the wall was extremely high and high. The wall is cold and rough, with a dead color. Her face, on the wall, against, also changed - red lips, water eyes, blood, flesh, a face with thoughts. Liu Yuan looked at her and said: "This wall, I don't know why it reminds me of the old days.... One day, our civilization was completely destroyed, and everything was finished." After it was burned, the bomb was finished, and it was finished. Maybe there is still this wall. Fringe, if we met under the wall at that time... tassels, maybe you will have a little heart for me, maybe I will have a little heart for you."

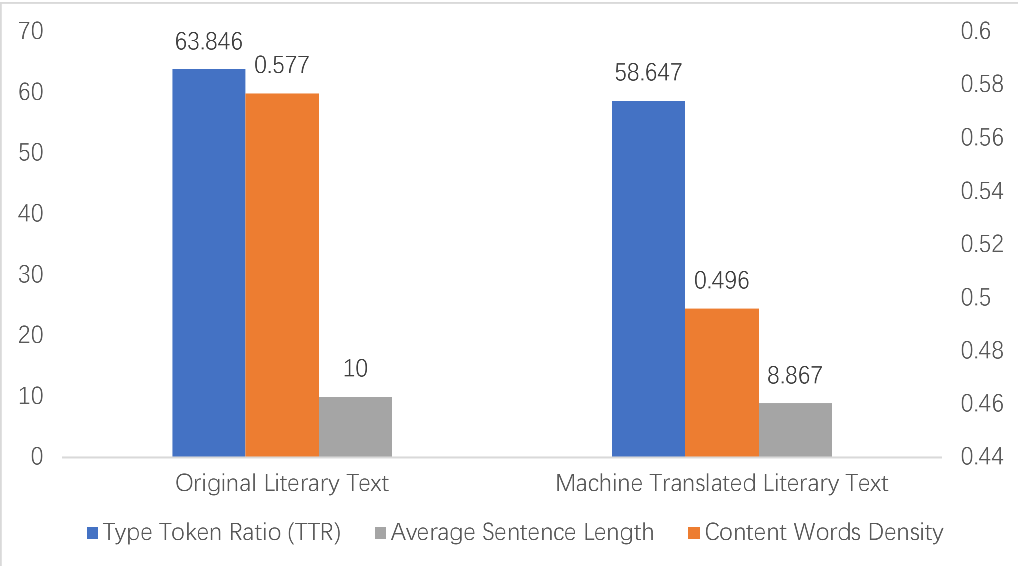

Figure 5.5 Calculation Data for Literary Text Case 5

As the four cases above, the same declining trend is shown in all three indicators data in Figure 5.5 (TTR: 58.647 < 63.846, Average Sentence Length: 8.867 < 10, Content Words Density: 0.496 < 0.577). However, the machine translation accuracy of this paragraph is much higher than that of former cases. Except for a few misunderstandings of Chinese idioms like “地老天荒(the end of time)” being translated into “old times“, the translation is acceptable on the whole and only needs little post-editing.

5.2 Non-literary Text Cases and Machine Translations

Case 1:

Source Text: 经济增长放缓是如今的热门话题，尽管如此，我们的研究发现，快速消费品消费依然强劲。2018 年，快速消费品消费总额继续反弹，增速达到 5.2%，略高于上一年的 4.7%。我们从《2016 年中国购物者报告》开始提出的“双速增长”现象依然存在：食品饮料行业、个人护理和家庭护理行业继续以不同的速度扩张，其中个人护理和家庭护理保持领先增速。受益于销量的稳健增长，2018 年包装食品销售额增长 4.7%，而饮料增速停留在 1.5%，这是销量下降和平均售价上涨共同作用的结果。相比之下，2018 年个人护理品类的销售额增长 10.3%，主要归功于高端化的推动，而家庭护理品类的销售额增长 7.2%，销量增长是主要的推动力。

Machine Translated Text: The slowdown in economic growth is a hot topic today, but our research has found that fast-moving consumer goods are still strong. In 2018, the total consumption of fast-moving consumer goods continued to rebound, with a growth rate of 5.2%, slightly higher than the 4.7% in the previous year. The “two-speed growth” phenomenon we started with the “2016 China Shopper Report” still exists: the food and beverage industry, the personal care and home care industries continue to expand at different rates, with personal care and home care maintaining a leading growth rate. Benefiting from steady growth in sales, packaged food sales increased by 4.7% in 2018, while beverage growth stayed at 1.5%, which was the result of a combination of lower sales and higher selling prices. In contrast, sales of personal care products increased by 10.3% in 2018, mainly due to the promotion of high-end, while sales of home care products increased by 7.2%, and sales growth was the main driving force.

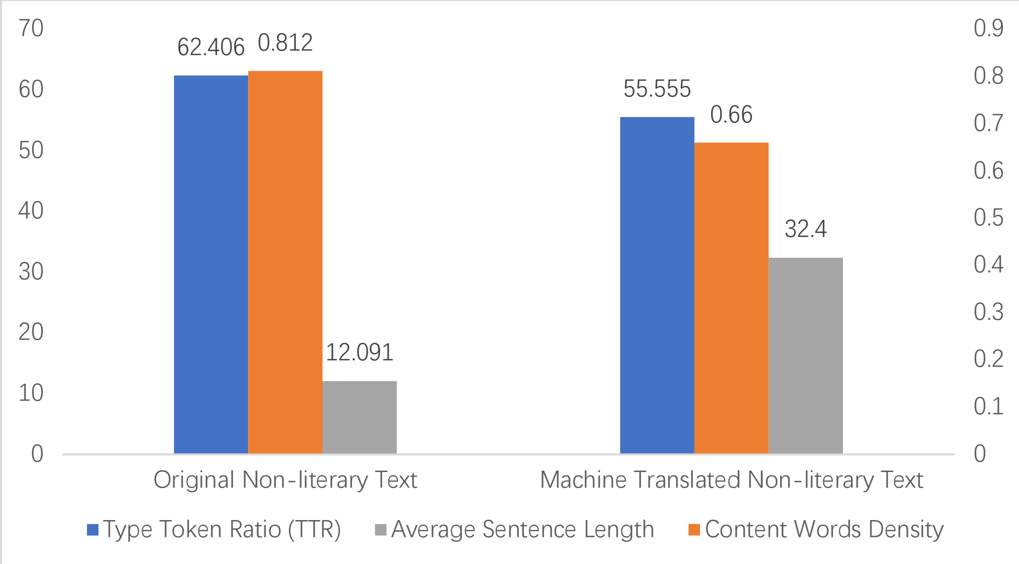

Figure 5.6 Calculation Data for Non-literary Text Case 1

Figure 5.6 shows three linguistic indicators of a single non-literary paragraph and its machine translation. After being translated into English in Google Translate, TTR and Content Words Density values show a downward trend (From 62.406 to 55.555 and from 0.812 to 0.66 respectively) which indicates that the words richness and lexical density of original non-literary text decrease, while another one Average Sentence Length value sees an increase from 12.091 up to 32.4 which means that translated sentences may be more complex. The overall translation quality of this paragraph, judged by its specific words usage, grammar and sentence structure, is high without any misunderstanding or data missing.

Case 2:

Source Text: 第二组以个人护理品类为主，包括个人清洁用品、洗发护发和口腔护理。这些品类的电商相对渗透率处于中等水平，而且电商渗透率从 2016 年以来保持高速增长。渗透率的显著增长在很大程度上归功于电商平台的促销活动，以及小品牌和领先品牌在电商渠道的巨大投资。第三组包括大多数食品和家庭护理品类，包括方便面、糖果、巧克力和厨房清洁用品。这些品类的电商相对渗透率处于中等水平，但电商渗透率增速较低。最后一组与去年报告相同：电商相对渗透率较低、电商渗透率增长有限的品类。

Machine Translated Text: The second group focuses on personal care categories, including personal cleansers, shampoos and oral care. The relative penetration rate of e-commerce in these categories is at a medium level, and e-commerce penetration has maintained a high growth rate since 2016. The significant increase in penetration is due in large part to the promotion of e-commerce platforms and the huge investment of small brands and leading brands in e-commerce channels. The third group includes most food and home care categories, including instant noodles, candy, chocolate and kitchen cleaning products. The relative penetration rate of e-commerce in these categories is at a medium level, but the penetration rate of e-commerce is low. The last group is the same as last year's report: the category of e-commerce with relatively low penetration rate and limited growth in e-commerce penetration.

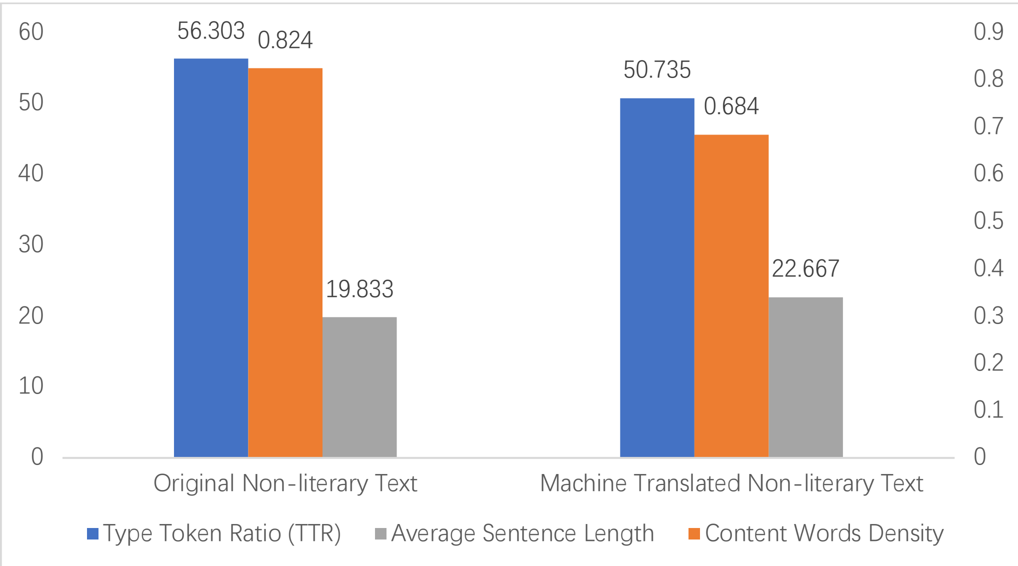

Figure 5.7 Calculation Data for Non-literary Text Case 2

As shown from Figure 5.7, TTR and Content Words Density values of the original paragraph (56.303 and 0.824 respectively) are higher than that of its machine translation (50.735 and 0.684 respectively) which reveals that the words and phrases used in the translated version are less rich and varied, while Average Sentence Length value experiences an increase from 19.833 up to 22.667. After a detailed look at translated words and sentences, it can be said that most information in the original non-literary text is accurately and clearly conveyed through machine translation though the last underlined sentence might be a little confusing for readers.

Case 3:

Source Text: 大卖场市场份额连年下滑，从 2014 年的 23.6%下降到 2018 年的 20.2%。然而，小型业态下滑正在放缓，在某些情况下甚至实现了增长。例如，在 2018 年，超市和小超市渠道的市场份额虽然下滑，但销售额实现 1.9%的增长。此外，基于我们追踪的 10 个品类，在传统渠道（杂货店）中，在外消费的食品饮料销售额自 2016 年以来年均增长 14%，占 2018 年杂货店销售额比例达到近 80%。在外消费也拉动了便利店销售额的增长，2016 年以来，便利店在一二线城市的销售额年均增长了 16%。我们发现，大型业态也具备增长潜力，但需要转型。

Machine Translated Text: Hypermarket market share has declined year after year, from 23.6% in 2014 to 20.2% in 2018. However, the decline in small formats is slowing and in some cases even growing. For example, in 2018, although the market share of supermarkets and small supermarket channels declined, sales increased by 1.9%. In addition, based on the 10 categories we tracked, in traditional channels (grocery stores), the sales of food and beverages consumed outside the country increased by 14% annually since 2016, accounting for nearly 80% of the grocery store sales in 2018. Out-of-town consumption has also boosted the growth of convenience store sales. Since 2016, sales of convenience stores in first- and second-tier cities have grown at an average annual rate of 16%. We have found that large formats also have growth potential but need to be transformed.

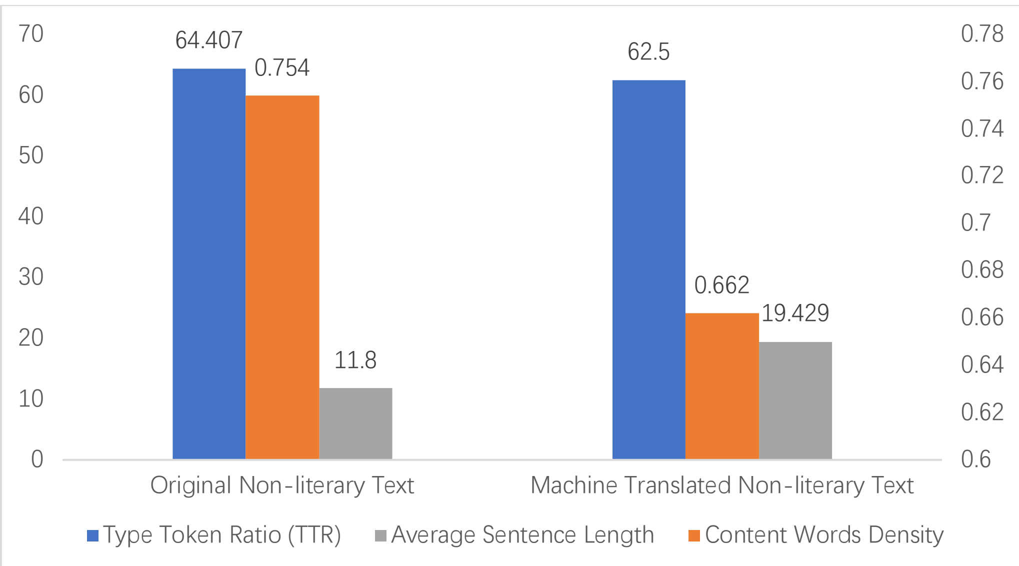

Figure 5.8 Calculation Data for Non-literary Text Case 3

As the above two cases, TTR and Content Words Density values decrease after machine translation (from 64.407 to 62.5 and from 0.754 to 0.662 respectively), and Average Sentence Length value increases from 11.8 to 19.429, which indicates that for this paragraph words have been simplified by machine translation and sentence structure might be more complicated. This machine translated non-literary text, on the basis of correctly expressing the original content, ensures the diversity of words usage and sentences structure, and achieves the best possible translation.

Case 4:

Source Text: 新的市场现实是什么？小品牌在满足消费者原始需求方面的表现令人印象深刻，从研发到数字营销，它们在每个环节都充分展现了速度和敏捷，相比之下，大品牌似乎表现平平。无论是专注于发展大品牌，还是建立不同品牌组合、服务不同客户细分，都是令每一位快速消费品公司高管头疼的问题。解决这个问题有时需要重大战略转型；市值 10 亿美元的品牌与价值 2500 万美元的品牌大不相同，需要截然不同的管理方法。新零售是我们解读的另一大新兴趋势。和往年一样，我们深入调查了食品饮料品类的电商渠道和在外消费市场。

Machine Translated Text: What is the new market reality? The performance of small brands in meeting the original needs of consumers is impressive. From research and development to digital marketing, they show speed and agility at every stage. In contrast, big brands seem to be flat. Whether it is focusing on the development of big brands, or establishing different brand combinations and serving different customer segments, it is a headache for every fast-moving consumer goods company executive. Solving this problem sometimes requires a major strategic transformation; brands with a market value of $1 billion are very different from brands with a value of $25 million and require very different management methods. New retail is another big emerging trend we have interpreted. As in previous years, we conducted in-depth investigations into the e-commerce channels of food and beverage categories and the external consumer market.

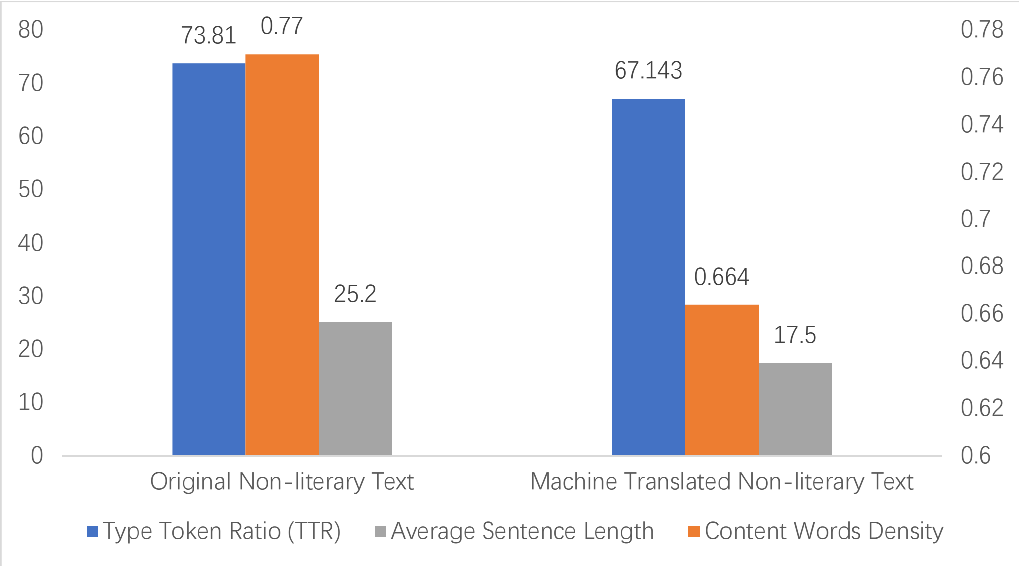

Figure 5.9 Calculation Data for Non-literary Text Case 4

A slightly different trend is found in the figure of Case 4, in which all indicators results of machine translated non-literary text (67.143, 0.664 and 17.5 respectively) are lower than original ones (73.81, 0.77 and 25.2 respectively). After a proof reading, it can be said that the words and grammar that Google Translate uses are clear and accurate to present the original content.

Case 5:

Source Text: 尽管人们普遍担心经济增长放缓，但 2018 年中国快速消费品市场表现依然强劲。快速消费品消费总额增长 5.2%，略高于上一年 4.7%的增速。个人护理品类的增长最健康，增速从 2017 年的 10.1%提高到 10.3%。高端化是这一出色表现背后的重要原因：由于消费者的消费升级意愿不断加强，商品平均售价上涨 9.8%。例如，护发素品类的销量下滑 0.8%，但平均售价却上涨 8.1%。洗发水销量仅增长 1%，但价格上涨 5%；欧雅和潘婷的平均售价分别增长 6%和 4%。家庭护理品类强劲增长 7.2%，较 2014-2017 年 3%-4%的年均增长率有所反弹。在家庭护理领域，销售额的增长主要来自销量增长，而不是价格上涨。

Machine Translated Text: Despite widespread concerns about slowing economic growth, China’s fast-moving consumer goods market remained strong in 2018. The total consumption of fast- moving consumer goods increased by 5.2%, slightly higher than the growth rate of 4.7% in the previous year. The growth of personal care categories was the healthiest, with growth rates increasing from 10.1% in 2017 to 10.3%. High-end is an important reason behind this outstanding performance: the average selling price of goods has increased by 9.8% due to the increasing consumer willingness to upgrade. For example, sales of hair conditioners fell 0.8%, but average selling prices rose 8.1%. Shampoo sales increased by only 1%, but prices rose by 5%; L'Oreal and Pantene's average selling price increased by 6% and 4% respectively. Home care products grew strongly by 7.2%, rebounding from the average annual growth rate of 3%-4% in 2014-2017. In the home care sector, sales growth is mainly driven by sales growth, not price increases.

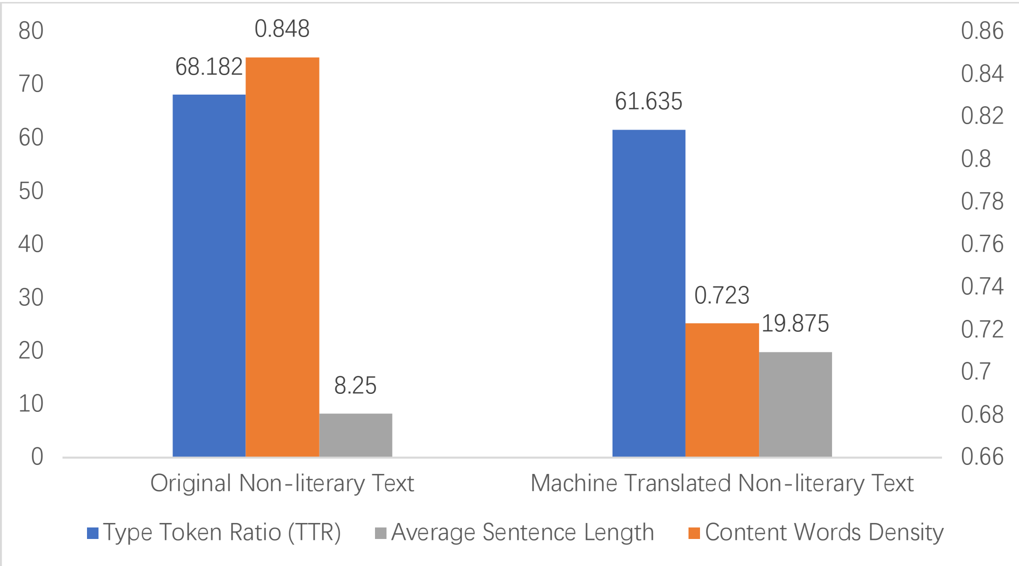

Figure 5.10 Calculation Data for Non-literary Text Case 5

The figure above shows two declining values with TTR value from 68.182 to 61.635 and Content Words Density value from 0.848 to 0.723, and one largely increasing value of Content Words Density from 8.25 to 19.875. A small mistake is found in this translation that the adjective "最健康 (referring to be 'the most reasonable' in current context)" is translated into "the healthiest" since computer only understands this vocabulary in its most literal sense. But beyond that, the machine translation performance is commendable.

5.3 Summary

This chapter chooses five literary text cases and five non-literary text cases to conduct text analysis. The calculation data of three indicators Type Token Ratio, Average Sentence Length and Content Words Density of each chosen paragraph and its machine translation are first shown in Figure 5.1-5.10. Then the next is the analysis on specific words and sentences. For literary text cases, the machine translation quality fails to live up to expectation. For non-literary text cases, Google Translate delivers high quality translation.

6. Discussions

Generally, this corpus-driven study has two research approaches. First, the study calculated linguistic indicators values of chosen literary and non-literary texts and carried out the Pearson correlation analysis for each indicator to assess the machine translation performance. Second, on the basis of data analysis, altogether ten cases from both literary and non-literary texts were selected for detailed assessment of machine translation quality. Thus, discussions are given towards two approaches and corresponding results in the following two sections.

6.1 Discussions on Corpus-driven Analysis Results

The present study first collected data on three linguistic indicators of original texts (literature and non-literature) and their Google translations, and then computed the correlation coefficients of each pair (pair 1: literary text and its translation; pair 2: non-literary text and its translation).

Table 6.1 Pearson Correlation Coefficients Result

	TTR	Content Words Density	Average Sentence Length
Literary Text	0.534**	0.150	0.178
Non-literary Text	0.757**	0.575**	0.824**

(** $p < 0.01$)

As shown in Table 6.1, the data demonstrate that, except the correlation analysis on Type Token Ratio (TTR) in which significant correlation exists in both pairs, the other two show similar results with a significant positive correlation existing between non-literary text and its machine translation and no significant correlation is found between literary text and its machine translation. The correlation existence provides the evidence that two variables are correlated with a very high statistical significance. In other words, there is a linear relationship between two variables. Google translation of non-literary text shares strong positive correlation with the source text in all three linguistic indicators in this study, indicating that machine translation performance on non-literary text is stable and consistent. By comparison, the result that only one significant correlation exists between the literary pair shows that machine translation performance on literary translation is basically not constant. The evaluation that machine translation performance on literary text is less stable and consistent than non-literary text is also supported by the standard deviations of the difference between the value of original text and that of machine translation for each indicator (see Table 6.2):

Table 6.2 Standard Deviations of the Difference for Each Indicator

	TTR	Content Words Density	Average Sentence Length
Literary Text	4.76	8.49	0.05
Non-literary Text	4.66	5.94	0.04

As shown in Table 6.2, all three standard deviations of literary text (4.76, 8.46 and 0.05 respectively) are higher than that of non-literary text (4.66, 5.94 and 0.04 respectively), which indicates that there is a larger fluctuation in the data of literary text.

In terms of TTR, though both literary and non-literary pairs show significant correlation, the correlation strength is different between two pairs with the former in moderate level and latter in strong level. Hence, the machine translated literary text is less related to the original one on TTR. As for the other two indicators, the data show that the machine translated literary text has no correlation with the original text on either Average Sentence Length or Content Words Density, while the machine translated non-literary text is highly correlated with the original one.

The overall data analysis part of this study mainly lists the calculated Pearson correlation coefficients result, which confirms that machine translation performance is less stable and consistent when dealing with literary text than non-literary text.

6.2 Discussions on Case Analysis Results

In order to verify the correctness of the data result and assess the machine translation quality from a qualitative perspective, several paragraphs were randomly selected from literary and nonliterary text for carrying out case analysis.

The changing trend of calculated three indicators values for each case is basically consistent with the overall sample text. For literary text, after being processed by machine, TTR, Average Sentence Length and Content Words Density values show a drop down trend, which means that the words richness, lexical density and sentence structure of original text are all simplified by machine translation. Besides, since there are many grammatical mistakes, misunderstandings and omissions in the machine translated texts, the translation quality is unsatisfactory and more technological upgrades are needed in helping machines comprehend culture-loaded words and different cultural contexts in literary works. By comparison, for non-literary text, though the words richness and lexical density values show a certain decline, the overall translation is accurate and smooth without information loss or misunderstanding, which provides users very good reading experience and is certainly able to meet the needs of machine translation users.

In general, the final result that machine translation performance on literary text is barely satisfactory is in line with most people's original judgement that the profound literary connotation can never be easily understood by machines. The most advanced translation machines make it possible to the batch processing of non-literary texts like legal documents, company files or news reports, but still fail to read between the lines when handling literary works. For better translating literary works, machine translation developers could consider collecting more training data on literature and designing translation systems more adaptive to literary narrative structure. There is a great demand for translating foreign literary works and it is expected that in the near future machine translation may offer support to lift translation efficiency and help promote foreign cultures.

7. Conclusions

This study aimed to assess the machine translation performance on literary and non-literary texts. It first built a small corpus of both literary and non-literary texts, and then conducted a Pearson correlation analysis on the basis of calculated indicators data. In addition to the corpus research, it selected a few cases to assess the machine translation quality from a qualitative perspective. Major findings of this study are listed in the first section. Although the present study has come to a preliminary conclusion that translation machine performance on literary text is unsatisfactory, limitations still remain. Given that more studies may be carried out on this subject in the future, suggestions are offered in the last section of this chapter.

7.1 Major Findings

This study found answers to the three research questions raised in the beginning:

(1) Is there any significant positive correlation in three linguistic indicators between literary text and its machine translation?

There is a significant positive correlation on TTR between literary text and its machine translation while no significant correlation is found on either Average Sentence Length or Content Words Density.

(2) Is there any significant positive correlation in three linguistic indicators between non-literary text and its machine translation?

Significant positive correlation is found on all three indicators, namely TTR, Average Sentence Length and Content Words Density, between non-literary text and its machine translation.

(3) Is machine translation performance on literary translation equal to that of non-literary translation?

Regarding the computational results of three linguistic indicators as well as case analysis result, the drawn conclusion is that machine translation of non-literary text outperforms the machine translation of literary text. And when dealing with literary text, machine translation has trouble figuring out the sentence structure and understanding specific culture-loaded words, as a result of which there are many mistransaltions and omssions in the output text. So far, the erratic performance of machine translation on literary text is unable to meet users' expectations yet. Generally, it is best to bring more literature-related training data to the translation system and conduct more researches to help machine translation better understand literary works.

The conclusion that machine translation of literary text still has a long way to go might disappoint some people while giving some others a break. However, the intension of conducting the research is to keep watch on the development of machine translation and update stakeholders by informing them of the latest process in machine translation field rather than create a sense of crisis. Only by knowing the degree of machine translation development can developers improve related techniques in a targeted direction and translators fully enjoy the convenience brought by technological progress.

7.2 Limitations and Suggestions

Although this study has successfully addressed three proposed research questions, there are still some limitations.

The first limitation is on the corpus size. This study chose *Love in a Fallen City* and *Bain China Shopper Report 2019* (both in Chinese) as literary and non-literary texts respectively, so that the machine translation is only from Chinese to English. It is expected that, in future studies, more classical literary works like novels, proses and poetries and more valuable non-literary texts like commercial documents, official reports and public speeches will be sorted and brought into research to expand the existing research corpus. Besides, the language of source text should be diversified rather than only being restricted to Chinese. Many similar linguistic researches, to conduct a comprehensive approach, involve even more than twenty languages from different language families. Texts in different languages like English, Arabic, Spanish and Japanese should be taken into consideration as well, which may bring new discoveries and insights.

The second limitation concerns the number of indicators. Although this study chose three classical linguistic indicators Type Token Ratio, Average Sentence Length and Content Words Density to respectively represent words richness, lexical density and writing style, and the calculation data results gave full support to the final conclusion, it would be better to include more linguistic indicators like Readability into the research.

A well-known computing tool Coh-Matrix provides over one hundred linguistic metrics calculation and analyses texts generally from five aspects (word concreteness, syntactic simplicity, referential cohesion, deep cohesion and narrativity). But Coh-Matrix is only applicable to English texts, because for other languages the standard on how to calculate these linguistic metrics has not been unified. It is expected that in the future there will be more comprehensive linguistic metrics calculation tools that are suitable for different languages in the market and more linguistic indicators will be included into similar researches.

There are many possibilities for the future study. For instance, the research method the present study used might be applied into other translation studies to assess whether the translated text is highly correlated with the source text on the basis of calculated linguistic indicators result. Also, more researches related to the machine translation performance on literary and non-literary texts are needed to help find the reasons for the performance differences.

References

[1] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. *Sentiment Analysis of Twitter Data*[C]. Proceedings of the Workshop on Language in Social Media (LSM 2011). 2011: 30-38.

[2] Akoglu, H. User's Guide to Correlation Coefficients[J]. *Turkish Journal of Emergency Medicine*, 2018.

[3] Bahdanau, D., Cho, K., & Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. *arXiv: Computation and Language*, 2014.

[4] Baker, M. Corpus Linguistics and Translation Studies: Implications and Applications[J]. *Text and Technology: In Honour of John Sinclair*, 1993, 233: 250.

[5] Baker, M. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research[J]. *Target. International Journal of Translation Studies*, 1995, 7(2): 223-243.

[6] Baker, M. Corpus-based Translation Studies: The Challenges That Lie Ahead[J]. *Benjamins Translation Library*, 1996, 18: 175-186.

[7] Baker, M. Towards a Methodology for Investigating the Style of a Literary Translator[J]. *Target. International Journal of Translation Studies*, 2000, 12(2): 241-266.

[8] Bosseaux, C. Who’s Afraid of Virginia’s You: A Corpus-based Study of the French Translations of The Waves[J]. *Meta: Journal Des Traducteurs/Meta: Translators' Journal*, 2006, 51(3): 599-610.

[9] Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. Is Neural Machine Translation the New State of the Art? [J]. *The Prague Bulletin of Mathematical Linguistics*, 2017, 108(1): 109-120.

[10] Chen, X., Yan, Y., & Hu, J. A Corpus-based Study of Hillary Clinton’s and Donald Trump’s Linguistic Styles[J]. *International Journal of English Linguistics*, 2019, 9(3): 13-22.

[11] Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. On the Properties of

Neural Machine Translation: Encoder-Decoder Approaches[J]. *arXiv: Computation and Language*, 2014.

[12] Chung, Y. M., & Lee, J. Y. A Corpus‐based Approach to Comparative Evaluation of Statistical Term Association Measures[J]. *Journal of the American Society for Information Science and Technology*, 2001, 52(4): 283-296.

[13] Covington, M. A., & McFall, J. D. Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR)[J]. *Journal of Quantitative Linguistics*, 2010, 17(2): 94-100.

[14] De Vries, E., Schoonvelde, M., & Schumacher, G. No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications[J]. *Political Analysis*, 2018, 26(4): 417-430.

[15] Du, J., & Way, A. Pre-Reordering for Neural Machine Translation: Helpful or Harmful? [J]. *The Prague Bulletin of Mathematical Linguistics*, 2017, 108(1): 171-182.

[16] Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., ... & Tyers, F. M. Apertium: A Free/Open-source Platform for Rule-based Machine Translation[J]. *Machine Translation*, 2011, 25(2): 127-144.

[17] Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. Coh-Metrix: Analysis of Text on Cohesion and Language[J]. *Behavior Research Methods, Instruments, & Computers*, 2004, 36(2): 193-202.

[18] Graesser, A. C., Dowell, N., & Moldovan, C. A Computer's Understanding of Literature[J]. *Scientific Study of Literature*, 2011, 1(1): 24-33.

[19] Groves, M., & Mundt, K. Friend or Foe? Google Translate in Language for Academic Purposes[J]. *English for Specific Purposes*, 2015: 112-121.

[20] Hill, J., Ford, W. R., & Farreras, I. G. Real Conversations with Artificial Intelligence: A Comparison between Human-human Online Conversations and Human-chatbot Conversations[J]. *Computers in Human Behavior*, 2015: 245-250.

[21] Holmes, D. I. The Analysis of Literary Style—A Review[J]. *Journal of the Royal*

Statistical Society: Series A (General), 1985, 148(4): 328-341.

[22] Hutchins, W. J, & Somers, H. L. *An Introduction to Machine Translation*[M]. London: Academic Press, 1992.

[23] Hutchins, W. J. *Machine Translation: A Brief History*[M]. Concise History of the Language Sciences. Pergamon, 1995: 431-445.

[24] Hutchins, J. Machine Translation: A Concise History[J]. *Computer Aided Translation: Theory and Practice*, 2007, 13: 29-70.

[25] Jiang, J., & Liu, H. The Effects of Sentence Length on Dependency Distance, Dependency Direction and the Implications–based on a Parallel English–Chinese Dependency Treebank[J]. *Language Sciences*, 2015, 50: 93-104.

[26] Jones, R., & Irvine, A. Th*e (un)Faithful Machine Translator*[C]. Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. 2013: 96-101.

[27] Kazakova, T. A. Strategies of Literary Translation[J]. *Journal of Siberian Federal University*, 2015, 8(12): 2842-2847.

[28] Kelih, E., Grzybek, P., Antić, G., & Stadlober, E. *Quantitative Text Typology: The Impact of Sentence Length*[M]. From Data and Information Analysis to Knowledge Engineering. Springer, Berlin, Heidelberg, 2006: 382-389.

[29] Koehn, P., & Knowles, R. *Six Challenges for Neural Machine Translation*[C]. Meeting of the Association for Computational Linguistics, 2017: 28-39.

[30] Läubli, S., & Orrego-Carmona, D. When Google Translate is Better than Some Human Colleagues, Those People are no Longer Colleagues[J]. *Translating and the Computer 39*, 2017: 59.

[31] Laufer, B., & Nation, P. Vocabulary Size and Use: Lexical Richness in L2 Written Production[J]. *Applied Linguistics*, 1995, 16(3): 307-322.

[32] Laviosa, S. The Corpus-based Approach: A New Paradigm in Translation Studies[J]. *Meta: Translators' Journal*, 1998, 43(4): 474-479.

[33] Lefevere, A. *Translation, Rewriting, and the Manipulation of Literary Fame*[M]. Routledge, 2016.

[34] Liang, J., Fang, Y., Lv, Q., & Liu, H. Dependency Distance Differences Across Interpreting Types: Implications for Cognitive Demand[J]. *Frontiers in Psychology*, 2017, 8: 2132.

[35] Liu, H. Dependency Distance as a Metric of Language Comprehension Difficulty[J]. *Journal of Cognitive Science*, 2008, 9(2): 159-191.

[36] Liu, H. What Linguistics Could Do in the Era of Big Data (in Chinese) [N]. *Chinese Social Sciences Weekly*,2017-3-16.

[37] Loper, E., & Bird, S. NLTK: The Natural Language Toolkit[J]. *arXiv preprint cs/0205028*, 2002.

[38] Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. Computer-assisted Text Analysis for Comparative Politics[J]. *Political Analysis*, 2015, 23(2): 254-277.

[39] McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955[J]. *AI Magazine*, 2006, 27(4): 12-12.

[40] McKee, G., Malvern, D., & Richards, B. Measuring Vocabulary Diversity Using Dedicated Software[J]. *Literary and Linguistic Computing*, 2000, 15(3): 323-338.

[41] Min, K., Ma, C., Zhao, T., & Li, H. *BosonNLP: An Ensemble Approach for Word Segmentation and POS Tagging*[M]. Natural Language Processing and Chinese Computing. Springer, Cham, 2015: 520-526.

[42] Nagao, M. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle[J]. *Artificial and Human Intelligence*, 1984: 351-354.

[43] Nagpal, A., & Gabrani, G. *Python for Data Analytics, Scientific and Technical Applications*[C]. 2019 Amity International Conference on Artificial Intelligence (AICAI). IEEE, 2019: 140-145.

[44] Olohan, M., & Baker, M. Reporting That in Translated English. Evidence for Subconscious Processes of Explicitation?[J]. *Across Languages and Cultures*, 2000, 1(2): 141-158.

[45] Pastor, G. C., Mitkov, R., Afzal, N., & Pekar, V. *Translation Universals: Do They Exist? A Corpus-based NLP Study of Convergence and Simplification*[C]. 8th AMTA Conference. 2008: 75-81.

[46] Prates, M. O., Avelar, P. H., & Lamb, L. C. Assessing Gender Bias in Machine Translation: A Case Study with Google Translate[J]. *Neural Computing and Applications*, 2018: 1-19.

[47] Richards, B. Type/Token Ratios: What Do They Really Tell Us? [J]. *Journal of Child Language*, 1987, 14(2): 201-209.

[48] Smith, M. W. Recent Experience and New Developments of Methods for the Determination of Authorship[J]. *ALLC BULL*., 1983, 11(3): 73-82.

[49] Stubbs, M. *Words and Phrases: Corpus Studies of Lexical Semantics*[M]. Oxford: Blackwell Publishers, 2001.

[50] Sumita, E., Iida, H., & Kohyama, H. *Translating with Examples: A New Approach to Machine Translation*[C]. The Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language. 1990 (3): 203-212.

[51] Taivalkoski-Shilov, K. Ethical Issues Regarding Machine(-assisted) Translation of Literary Texts[J]. *Perspectives*, 2018:1-15.

[52] Toral, A., & Way, A. Machine-assisted Translation of Literary Text: A Case Study[J]. *Translation Spaces*, 2015a, 4(2): 240-267.

[53] Toral, A., & Way, A. *Translating Literary Text between Related Languages Using SMT*[C]. Proceedings of the Fourth Workshop on Computational Linguistics for Literature. 2015b: 123-132.

[54] Toral, A., & Way, A. What Level of Quality can Neural Machine Translation Attain on Literary Text[J]. *arXiv: Computation and Language*, 2018.

[55] Toral, A., Wieling, M., & Way, A. Post-editing Effort of a Novel with Statistical and Neural Machine Translation[J]. *Frontiers in Digital Humanities*, 2018: 1-11.

[56] Voigt, R., & Jurafsky, D. *Towards a Literary Machine Translation: The Role of Referential Cohesion*[C]. North American Chapter of the Association for Computational Linguistics, 2012: 18-25.

[57] Way, A. Traditional and Emerging Use-cases for Machine Translation[J]. *Proceedings of Translating and the Computer*, 2013, 35.

[58] Wong, K. F., Li, W., Xu, R., & Zhang, Z. S. Introduction to Chinese Natural Language Processing[J]. *Synthesis Lectures on Human Language Technologies*, 2009, 2(1): 1-148.

[59] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Klingner, J. Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation[J]. *arXiv preprint arXiv:1609.08144*, 2016.

[60] Yule, G. U. On Sentence-length as a Statistical Characteristic of Style in Prose: With Application to Two Cases of Disputed Authorship[J]. *Biometrika*, 1939: 363-390.

[61] Zanettin, F., Bernardini, S., & Stewart, D. *Corpora in Translator Education*[M]. Routledge, 2014.

[62] 陈吉荣. 基于自译语料的翻译理论研究：以张爱玲自译为个案[M].中国社会科学出版社，2009.

[63] 付巧玉.解构主义翻译观视域下文学作品不可译性探究[J].东北师大学报(哲学社会科学版),2019(02):55-60.

[64] 何莲珍.基于汉、英语平行语料库的翻译数据库设计[J].现代外语,2007(02):191-199+220.

[65] 胡开宝,毛鹏飞.国外语料库翻译学研究述评[J].当代语言学,2012,14(04):380-395+437-438.

[66] 胡开宝.基于语料库的莎剧《哈姆雷特》汉译文本中“把”字句应用及其动因研究[J].外语学刊,2009(01):111-115.

[67] 黄伟，梁君英. 语言研究的科学化与国际化[N].光明日报, 2017-3-26.

[68] 黎昌抱,李菁.基于语料库的文学自译语言特征考察[J].中国外语,2017,14(02):86-95.

[69] 冉诗洋.国际语料库翻译学研究新趋势——《语料库翻译学新趋向》评析[J].中国科技翻译,2019,32(01):59-61+39.

[70] 斯介生,宋大我,李扬.大数据背景下的谷歌翻译——现状与挑战[J].统计研究,2016,33(05):109-112.

[71] 孙艺风.翻译研究与世界文学[J].中国翻译,2019,40(1):5-18.

[72] 王克非. 语料库翻译学探索[M]. 上海：上海交通大学出版社,2012.

[73] 王克非,胡显耀.基于语料库的翻译汉语词汇特征研究[J].中国翻译,2008,29(06):16-21+92.

[74] 王克非,秦洪武,王海霞.双语对应语料库翻译教学平台的应用初探[J].外语电化教学,2007(06):3-8.

[75] 许钧. 论翻译的层次[J]. 现代外语, 1989 (3): 63-70.

[76] 许渊冲.文学翻译的心路历程[J].中国翻译,2003(04):32-35.

[77] 许渊冲.有中国特色的文学翻译理论[J].中国翻译,2016,37(05):93-99.

[78] 闫亮亮,朱健平.从“求达”到“信达雅”——严复“信达雅”成因钩沉[J].外语与外语教学,2017(05):122-131+151.

[79] 杨惠中.语料库语言学与机器翻译[J].上海交通大学学报(社会科学版),1993(01):98-111.

[80] 杨梅,白楠.国内语料库翻译研究现状调查——基于国内学术期刊的数据分

析(1993-2009)[J].中国翻译,2010,31(06):46-50.

[81] 郑海凌.钱钟书“化境说”的创新意识[J].北京师范大学学报(人文社会科学版),2001(03):70-76.