Wang, Q. & Xu, J. (2023). Neural machine translation in AVT teaching in China: An in-depth analysis from the readability perspective. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *22*, 161–180.

Neural machine translation in AVT teaching in China: An in-depth analysis from the readability perspective

Qingran Wang
China University of Political Science and Law
cu192019@cupl.edu.cn
https://orcid.org/0009-0006-5861-9967

Jun Xu
China University of Political Science and Law
xujun289@163.com
https://orcid.org/0009-0005-2304-6997

Abstract

As audiovisual translation (AVT) becomes more complex and diverse, the need for advanced machine learning techniques has been increasing sharply, driving the widespread adoption of neural machine translation (NMT) technology in the field. This study contributes to the literature by evaluating the performance of NMT technology in AVT teaching. Based on readability theory, we constructed an evaluation framework with 12 indicators, built comparable corpora consisting of human and post-edited subtitle translations of corporate videos, and used them to examine the performance of four online NMT systems (Google Translate, Baidu Translate, Bing Translator, and Youdao Translate) in AVT teaching. Our statistical analyses and case studies show that Google Translate outperforms the other three platforms in all the readability tests, and it can enhance the readability of post-edited subtitles at five levels (word, syntax, textbase, situation model, genre and rhetorical). The performance of the other three platforms varies across different tests. Concrete examples are provided to substantiate the statistical analyses. Our study adds value to existing research both by examining the application and performance of NMT in AVT teaching and by suggesting potential directions for the refinement of current NMT systems.

Keywords: neural machine translation, NMT, AVT teaching, MT, machine translation evaluation, readability test, Chinese–English translation

Wang, Q. & Xu, J. (2023). Neural machine translation in AVT teaching in China: An in-depth analysis from the readability perspective. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *22*, 161–180.

1. Introduction

The field of machine translation (MT) has witnessed three major technological changes to date: the widespread use of computer-assisted translation tools (CAT) since the late 1980s, the rapid development of statistical machine translation (SMT) technology in the late 1990s, and the current promotion and application of neural machine translation (NMT) technology for both academic research and commercial use (Wang & Xu, 2022). Since its introduction in 2016, NMT has been empirically proven to outperform SMT and has even replaced SMT in various online MT systems.

MT evaluation is essential to MT research and development because the assessment results can reveal the degree of output reliability and help to refine MT systems. Chatzikoumi (2020) classifies MT evaluation metrics in the international literature into two categories: automated evaluation and human evaluation. In practice, automated metrics play a crucial role in the development of MT systems (Giménez & Márquez, 2010) and are more popular than human metrics because of their low cost, time-saving characteristics, repeatability, and consistency (Banerjee & Lavie, 2005; Olive et al., 2011). With the help of parallel corpora, NMT output is primarily evaluated by automatic metrics. However, Popović (2017) notes that most of the current evaluation studies present overall scores for the products of NMT systems, which tell us only about the general performance of a system and fail to provide information that is more detailed. To the best of our knowledge, no existing studies have examined the application and performance of NMT in AVT teaching. In this article, we complement the above research by reporting on the development of an evaluation framework for assessing the performance of NMT systems that should help students to translate subtitles in the AVT environment.

Specifically, we used the Chinese videos of the “About Us” sections on the official websites of Chinese companies included in the 2017 Fortune Global 500 as our research sample. We then adopted various evaluation indicators to compare the human subtitle translations produced by students without the use of MT with the post-edited subtitle translations produced by both students and four popular online MT systems (Google Translate, Bing Translator, Baidu Translate, and Youdao Translate). Our study intended to conduct performance evaluation for NMT output at a number of levels in the AVT environment instead of giving an overall score, as has been presented in previous research on MT evaluation. The results demonstrate how subtitle translations can be improved at the five text levels (word, syntax, textbase, situation model, genre and rhetorical) identified by Graesser et al. (2011) in the AVT environment with the help of the four NMT systems.

2. Related Literature

2.1 Application of Neural Machine Translation in the AVT environment

Although the AVT industry has had a long history of working with technology, it has been relatively slower than other translation fields in adopting and using technological tools to improve the efficiency of the translation process (Bywood et al., 2017). MT is commonly employed in the translation industry due to the repetitive nature of text, and research has demonstrated that it can enhance translators' productivity through post-editing the MT output (Sanchez-Torron & Koehn, 2016). NMT is based on deep learning in neural networks and is considered a major breakthrough in MT. As AVT becomes more complex and diverse, the need for advanced machine learning techniques has become increasingly apparent, driving as it has the widespread adoption of neural machine learning in the field. In recent years, the use of NMT in combination with interlingual subtitling has been investigated.

Bellés-Calvera and Quintana (2021) used an NMT system to translate Spanish audio into English subtitles and then to evaluate the quality of the generated subtitles. They concluded that NMT has the potential to be a valuable tool for subtitling in the AVT industry while emphasizing the importance of human post-editing to ensure the quality of the final product. Using well-established methods from research on the translation process, Tardel (2021) examined the way in which the integration of language technology, specifically NMT, has an impact on the sub-processes of interlingual subtitling when used in an indirect translation or pivot setup. The study adds to the limited empirical research on the process of subtitling and sheds light on the role of NMT in the post-editing of audiovisual content. Matusov et al. (2019) customized an NMT system for translating subtitles in the domain of entertainment. Their novel subtitle segmentation algorithm resulted in a notable productivity increase of up to 37% compared to translating from scratch and also in significant reductions in the rate of editing of human translation. Overall, these studies highlight the likelihood that the application of NMT in AVT has great promise and potential.

2.2 Quality Assessment for Neural Machine Translation

NMT has become increasingly popular in recent years due to its ability to produce high-quality translations in various domains, such as business, education, and international relations (Wang & Xu, 2023). The performance of MT systems has improved significantly over the years, particularly with the advent of NMT models, which have shown remarkable translation accuracy in several language pairs (Revanuru et al., 2017; Song et al., 2020). However, the quality of the translation output can vary significantly depending on the NMT model, training data, and input text (Östling & Tiedemann, 2017). Therefore, it is essential to have reliable methods for evaluating the quality of NMT output.

Various methods are available to assess the quality of NMT output. Researchers have proposed both automated and human-based evaluation metrics (Chatzikoumi, 2020), and with the help of parallel corpora NMT outputs can be evaluated by both metrics (Popović, 2017). Specifically, evaluating the quality of Chinese–English translations produced by NMT is still a challenging task, and several scholars have conducted research in this area. Hassan et al. (2018) adopted the human metric to evaluate the performance of NMT in a news translation task from Chinese to English and they claimed that NMT output has reached human parity in this specific task. However, their human error analysis also indicates that there is still room for improvement in NMT output. Similarly, Han and Meng (2022) evaluated the quality of Chinese–English online translation using the BP neural network algorithm. They found that Baidu translation and iFLYTEK translation have a much higher error rate than Google Translate. Jia et al. (2019) compared the output quality of PBSMT and NMT systems and found that NMT produces higher-rated translations for both simple and more complex text. These findings suggest that NMT is a promising direction to move in for improving Chinese–English translation.

2.3 Readability and Machine Translation Evaluation

Readability is considered one of the most important factors that influence translation quality (Mobarakeh & Sardareh, 2016). Previous studies have employed various approaches to assess the readability of MT output. Jones et al. (2005) conducted experiments on Arabic-to-English text-based and audio-based MT systems and found that MT systems significantly affected readability, with text-based MT systems outperforming their audio-based counterparts. Alva-Manchego and Shardlow (2022) focused on COVID-19-related text and investigated the capabilities of MT models in generating translations with varying levels of readability. Furthermore, van Toledo et al. (2023) proposed a novel method based on fluency and readability indicators to predict when Google Translate is superior to other MT systems in Dutch translation. The above findings underscore the importance of considering readability in evaluating MT output.

Previous research has developed various formulas for measuring the readability of English text. Based on the assumption that the readability of a text is significantly influenced by shallow features (such as word length and sentence length), traditional readability approaches pay much attention to these shallow text properties (Ciobanu, 2015). However, these shallow factors often reflect only part of the superficial characteristics of the text, that is, part of the difficulty of the text. As a result, these traditional formulas tend to ignore other multidimensional levels involved in the reading comprehension process (McNamara et al., 2014).

Among the formulas that have been used to assess or measure readability, the LIX index and the Flesch–Kincaid measures are viewed as the most reliable and widely used metrics (Smith & Taffler, 1992). The most important parameters of these readability metrics are the lexical and morpho-syntactic features of a text (Loughran & McDonald, 2014). However, McNamara

et al. (2014) claimed that discourse-level analysis is needed to evaluate readability and that textual cohesion and coherence cannot be fully tested at the sentence level. As a supplement to previous readability evaluation approaches, the Coh-Metrix system has been developed to provide multilevel analyses of text readability from the word level to the discourse level (Graesser et al., 2011). In addition to the surface code (word and syntax), the system considers three other levels: the textbase, the situation model, and the genre and rhetorical structure. The principal component analysis conducted by Graesser and McNamara (2011) shows that eight principal components corresponding to these five levels can explain 67.3% of the readability variation in their sample.

It is worth mentioning that NMT technology was not available until recently, so the readability of machine-generated text from NMT models has not been studied extensively, especially in the field of Chinese–English translation. Further research is therefore needed to investigate the readability of NMT-generated text and to develop methods for measuring and enhancing readability which can help to improve the overall quality and usability of MT outputs. Our study aims to fill this gap by examining the application of NMT in enhancing Chinese-to-English translation in the AVT environment from the perspective of readability.

3. Data and Methodology

To conduct an evaluation of NMT in AVT teaching, we selected Chinese companies on the Fortune Global 500 list in 2017 and obtained the “About Us” sections in Chinese from their official websites. We then built comparable corpora that consist of human subtitle translations (HT corpus) of the “About Us” videos made by students without the use of MT and post-edited subtitle translations (PT corpus) acquired using popular online MT platforms based on NMT technology. Next, we selected 12 evaluation indicators from the literature and used WordSmith 5.0 and Coh-Metrix 3.0 to score the performance of the PT corpus and the HT corpus for each indicator. Finally, we used SPSS 24.0 to test the statistical significance of the difference between the NMT post-edited output and the HT output across the 12 indicators, in addition to case analysis to explain the results further.

3.1 Online Machine Translation Systems

We selected four popular online MT systems – Google Translate, Bing Translator, Baidu Translate, and Youdao Translate – to generate our NMT post-edited corpora. All four systems were used in February 2022. The reasons for selecting these four are:

- First, they all have adopted NMT technology, so by evaluating their output quality, we can test the effectiveness of the most cutting-edge MT technology.
- Second, according to the data from Similarweb, a web analytics company specializing in web traffic and performance analysis, the selected systems enjoy wide acceptance among

users of MT services worldwide, given their early launches, free access, and high level of popularity in the world.

- Third, the four systems are developed and maintained by top high-tech companies in the world. As they invest heavily in the research and development of NMT, their services represent the highest level of NMT technology.

According to Similarweb's data, the number of users of these four MT platforms far exceeds that of other platforms worldwide. By integrating the information published by these MT platforms and their traffic analysis information from Similarweb, Table 1 presents a comprehensive comparison of the selected platforms from five aspects, including major technology in generating Chinese–English translation, monthly visit, average visit duration, pages per visit, and bounce rate. Although these four platforms all adopt NMT technology, they are quite different in their website traffic. As shown in Table 1, Google Translate has the highest number of monthly visits, followed by Baidu Translate and Youdao Translate, while Bing Translator has the lowest number of monthly visits.

This is slightly different for pages per visit: Google still ranks highest, followed by Baidu and Bing, and Youdao ranks lowest. As supplementary indicators, visit duration represents the average time spent on the website of each user for the selected period, whereas bounce rate indicates the percentage of visitors who view only one page on the website before leaving. For both indicators, Google performs best, followed by Baidu, Youdao, and Bing. The web traffic analysis suggests that, among MT users worldwide, Google and Baidu are used more frequently and for longer times than Youdao and Bing.

Table 1

Basic information about the selected online MT systems

	Google Translate	Bing Translator	Baidu Translate	Youdao Translate
Technology (C–E translation)	Neural machine translation			
Monthly visits	727.7 M	8.12 M	71.72 M	10.43 M
Visit duration	00:07:57	00:00:47	00:06:35	00:01:56
Pages per visit	35.44	2.04	7.85	1.72
Bounce rate	21.24%	64.24%	25.90%	60.73%

Data source: information published by each platform, and https://www.similarweb.com. Visits are based on data as at 3 March 2023.

3.2 Corpora and Participants

We used the videos of the "About Us" sections of Chinese corporate websites as our research sample, mainly based on the following three considerations. First, because the teaching experiment was conducted in the class of "Audiovisual Translation of Business English" and the students were taking the course major in business administration, the selected videos needed to be business-related to meet the teaching objectives of the course. Second, as the purpose of the "About Us" section is to establish a corporate image by communicating with various stakeholders, it is an important platform for corporate marketing and the promotion of corporate image (Scott, 2015). Recently, the "About Us" section has become a hot topic for both business discourse research and business English teaching (Abdullah et al., 2013; Shi & Shan, 2019). Third, previous research on AVT teaching (Chinese to English) has rarely paid attention to business discourse. By constructing test corpora that fall within the business domain, our study could fill this gap.

Ten sample companies were selected from among those Chinese companies on the Fortune Global 500 list of 2017. To build comparable corpora, we first collected relevant Chinese text from the "About Us" sections of the selected companies' websites. Next, we asked a native Chinese speaker to render the relevant information into a video for each company which was 3–4 minutes long and contained both narration and footage of the company.

After that, we recruited 40 students from the class of "Audiovisual Translation of Business English" to participate in our teaching experiment. They were all first-year undergraduates of a prestigious university in Beijing with an average score of 134.2 (out of 150) in English in China's national college entrance examination. We randomly assigned these students into four groups of equal size so that each group had a similar level of English proficiency on average. The students were asked to produce English subtitles for the videos without the use of MT, which formed the corpus of the human translation. Each group was then assigned to one of the four online NMT systems. Finally, the students in each group used the NMT system assigned them to generate subtitle translations and then post-edited them. This process produced four corpora of post-edited translations. In this experiment, all the students worked individually.

3.3 Evaluation Methods

We employed 12 indicators (to be introduced below) to evaluate the readability of the corpora and used two corpus analysis tools, WordSmith 5.0 and Coh-Metrix 3.0, to acquire the indicator values for each corpus. We then used SPSS 24.0 to conduct statistical analyses of the evaluation indicators between the human translation and the post-edited translation corpora. WordSmith 5.0 has been developed by Oxford University Press and is the most widely used software in corpus-based studies (Wang & Liang, 2007). This study used WordSmith 5.0 mainly to score the word length and sentence length of the corpora and calculate the LIX index

accordingly.

Four of the 12 indicators are word length, sentence length, the LIX index, and the Flesch Reading Ease index. As discussed in section 2.3, a limitation of traditional readability measures is that they consider only the superficial features of text. As a complement to traditional readability metrics, Coh-Metrix 3.0 provides information about text difficulty at multiple levels, including word and sentence characteristics and the discourse relationships between ideas in text (McNamara et al., 2014). Through principal component analysis, Graesser et al. (2011) found eight principal components (narrativity, syntactic simplicity, word concreteness, referential cohesion, deep cohesion, verb cohesion, connectivity, and temporality) at five text levels (words, syntax, textbase, situation model, genre and rhetorical structure). Because our study aimed to assess the readability of self-built corpora at multiple levels of discourse, we included these eight indicators in our evaluation framework in addition to the four previously mentioned indicators. Their values were obtained from the easability module of Coh-Metrix 3.0.

4. Results and Discussion

In this next section, we report on the evaluation results for the corpora of human and post-edited subtitle translations based on the 12 indicators and discuss their implications.

4.1 Bing Translator

Table 2 presents the indicator values and evaluation results for the Bing corpus and the HT corpus for the two basic descriptive indicators (word length and sentence length) and the two traditional readability indicators (LIX index and Flesch Reading Ease index). In particular, the last column shows the statistical test results for the differences in these indicators between the two corpora and reports on the p-values in parentheses. At the 1% significance level, the Bing corpus exhibits no difference from the HT corpus on the four evaluation indicators.

Table 2

Basic descriptive indicators and traditional readability indicators (Bing Translator)

Evaluation Indicator	Students (HT)	Bing+Students (Post-editing)	Comparison of Evaluation Indicators (p-value in parentheses)
Word Length (number of letters)	5.51	5.63	HT < Post-editing (.094)
Sentence Length (number of words)	23.21	22.87	HT > Post-editing (.083)
LIX Index	52.54	52.62	HT < Post-editing (.375)
Flesch Reading Ease	22.53	22.63	HT < Post-editing (.574)

Table 3 presents the evaluation results for the Bing corpus and the HT corpus for the eight Coh-Metrix readability indicators. As shown in the last column, at the 1% significance level, the Bing corpus exhibits no significant difference from the HT corpus on all the indicators except for

word concreteness. The score of the Bing corpus on word concreteness (1.71) is higher than that of the HT counterpart (1.08) at the 1% significance level. McNamara et al. (2014) pointed out that a higher word concreteness score indicates a larger fraction of content words that are concrete, meaningful, and imaginable. Jiang and Han (2018) emphasized that text with more content words is easier to process than that with more abstract words. Example (1) illustrates this point.

Table 3

Coh-Metrix readability indicators (Bing Translator)

Evaluation Indicator	Students (HT)	Bing+Students (Post-editing)	Comparison of Evaluation Indicators (p-value in parentheses)
Narrativity	–1.91	–1.79	HT < Post-editing (.051)
Syntactic Simplicity	–1.33	–1.20	HT < Post-editing (.442)
Word Concreteness	1.08	1.71	HT < Post-editing (.000)***
Referential Cohesion	0.91	1.01	HT < Post-editing (.515)
Deep Cohesion	–0.83	–0.72	HT < Post-editing (.345)
Verb Cohesion	–0.18	–0.06	HT < Post-editing (.443)
Connectivity	–2.75	–2.67	HT < Post-editing (.567)
Temporality	–0.46	–0.67	HT > Post-editing (.291)

*** represents the 1% significance level.

Example (1): Word concreteness

Chinese text in the video (Alibaba): 我们的业务包括核心电商、云计算、数字媒体和娱乐以及创新项目。

Human Subtitle Translation: We provide e-commerce services.

Post-edited Translation (Bing): Our businesses include core commerce, cloud computing, digital media and entertainment, and innovation projects.

As shown in Example (1), when stating the company's main business, the student translated

hé xīn diànshāng yún jì suàn shù zì méi tǐ hé yú lè yǐ jí chuàng xīn xiàng mù
核心电商、云计算、数字媒体和娱乐以及创新项目 as "e-commerce services" in the human translation to describe the business field of the enterprise abstractly without explaining the specific type of business. The same participant translated it as "core commerce, cloud computing, digital media and entertainment, and innovation projects" in the post-edited translation (Bing). With the help of Bing Translator, the student was able to adopt concrete words and give a specific explanation of the company's main business. This case shows that the human translation has a weaker word concreteness compared to the post-edited translation (Bing), which is consistent with the empirical results and may be attributed to the limited extent of the student's vocabulary. Thus, our analysis suggests that Bing Translator can

help students to improve the readability of their subtitle translation in the AVT environment at the word level.

4.2 Youdao Translate

Table 4 presents the evaluation results for the Youdao corpus and the HT corpus for the two basic indicators and the two traditional readability indicators. The test results in the last column show that, according to the basic indicators, the Youdao corpus exhibits no significant difference from the HT corpus at the 1% significance level. However, the indicators differ significantly on the evaluation of the Flesch Reading Ease index, because the corresponding index value of the Youdao corpus is 23.47, higher than that of the HT corpus (22.53) at the 1% significance level. As discussed above, the Flesch Reading Ease index is one of the most widely used readability measures for English text and a higher value indicates a higher level of readability of the text. The statistical test results therefore suggest that the post-edited text in the Youdao corpus is more readable than the human translation in the HT corpus.

Table 4

Basic descriptive indicators and traditional readability indicators (Youdao Translate)

Evaluation Indicator	Students (HT)	Youdao+Students (Post-editing)	Comparison of Evaluation Indicators (p-value in parentheses)
Word Length (number of letters)	5.51	5.57	HT < Post-editing (.326)
Sentence Length (number of words)	23.21	23.50	HT < Post-editing (.102)
LIX Index	52.54	52.37	HT > Post-editing (.186)
Flesch Reading Ease	22.53	23.47	HT < Post-editing (.000) ***

*** represents the 1% significance level.

Table 5 presents the evaluation results for the Youdao corpus and the HT corpus for the eight Coh-Metrix indicators. As shown in the last column, the Youdao corpus exhibits no significant difference from the HT counterpart on the Coh-Metrix indicators, except for referential cohesion and deep cohesion. Specifically, the referential cohesion indicator and the deep cohesion indicator of the Youdao corpus are higher than those of the HT counterpart at the 1% significance level. The results suggest that Youdao can help students increase overlaps (as suggested by the result on referential cohesion) and the use of causal and intentional connectives to express the logical relationship of text explicitly (as suggested by the result on deep cohesion). As Baidu Translate and Google Translate also exhibit a similar pattern, we illustrate the readability difference in the cohesion indicators when we discuss Google Translate in section 4.4.

Table 5

Coh-Metrix readability indicators (Youdao Translate)

Evaluation Indicator	Students (HT)	Youdao+Students (Post-editing)	Comparison of Evaluation Indicators (p-value in parentheses)
Narrativity	–1.91	–1.86	HT < Post-editing (.251)
Syntactic Simplicity	–1.33	–1.55	HT > Post-editing (.232)
Word Concreteness	1.08	1.15	HT < Post-editing (.648)
Referential Cohesion	0.91	1.66	HT < Post-editing (.000)***
Deep Cohesion	–0.83	–0.42	HT < Post-editing (.006)***
Verb Cohesion	–0.18	–0.36	HT > Post-editing (.081)
Connectivity	–2.75	–2.62	HT < Post-editing (.334)
Temporality	–0.46	–0.39	HT < Post-editing (.712)

*** represents the 1% significance level.

The results for the two sentence indicators (sentence length, syntactic simplicity) suggest that Youdao cannot increase readability at the sentence level in the post-editing mode. Below, we use an example to illustrate this finding. Specifically, we select a sample video and compare the human subtitle translation to the post-edited translation based on Youdao Translate as follows.

Example (2): Syntactic Simplicity

Chinese text in the video (Industrial and Commercial Bank of China): 中国工商银行经过持续努力和稳健发展，已经迈入世界领先大银行行列，拥有优质的客户基础、多元的业务结构、强劲的创新能力和市场竞争力。

Human Subtitle Translation: Through continuous efforts and steady development, the Industrial and Commercial Bank of China has entered the ranks of the world's leading banks, with a high-quality customer base, diversified business structure, strong innovation capability and market competitiveness.

Post-edited Translation (Youdao): Through its continuous efforts and steady development, the Industrial and Commercial Bank of China has become one of the world's leading banks, possessing a solid customer base, a diversified business structure, strong innovation capabilities and market competitiveness.

The syntactic simplicity index reflects the degree to which sentences in text contain fewer words and use simpler syntactic structures that are less challenging to process (McNamara et al., 2014). As shown in Example (2), when stating the advantages of the bank (优质的客户基础、多元的业务结构、强劲的创新能力和市场竞争力), the differences in the human and post-edited translations are mainly reflected in the use of adjectives, singular and plural forms, and

indefinite articles. In both modes, the student basically used the same syntactic structure and adopted several noun phrases to describe the advantages of the bank. In addition, the sentence lengths of the two translations are almost the same. Therefore, there is no difference in syntactic complexity between the human and the post-edited translations, which is consistent with the empirical results in Table 5.

4.3 Baidu Translate

Table 6 presents the evaluation results for the Baidu corpus and the HT corpus for the two basic indicators and the two traditional readability indicators. The test results in the last column show that the two corpora exhibit no significant difference in the basic indicators, but they differ significantly on Flesch Reading Ease. In particular, the Flesch index of the Baidu corpus is 27.18, significantly higher than that of the HT corpus (22.53) at the 1% significance level. The results suggest that the post-edited translation is more readable than the human translation. In addition, Smith and Taffler (1992) showed that reading materials that have a Flesch index lower than 30 are difficult to read and more suited to readers with a higher education. However, Leong et al. (2002) found that the Flesch index for corporate website text is generally low, so our evaluation results on this index are limited to some extent by the text genre.

Table 6

Basic descriptive indicators and traditional readability indicators (Baidu Translate)

Evaluation Indicator	Students (HT)	Baidu+Students (Post-editing)	Comparison of Evaluation Indicators (p-value in parentheses)
Word Length (number of letters)	5.51	5.46	HT > Post-editing (.274)
Sentence Length (number of words)	23.21	23.48	HT < Post-editing (.113)
LIX Index	52.54	52.81	HT < Post-editing (.068)
Flesch Reading Ease	22.53	27.18	HT < Post-editing (.000)***

*** represents the 1% significance level.

Table 7 presents the evaluation results for the Baidu corpus and the HT corpus for the eight Coh-Metrix indicators. The test results in the last column indicate that the Baidu corpus is significantly different from the HT corpus at the 1% significance level for the following three indicators: syntactic simplicity, referential cohesion, and deep cohesion.

Table 7

Coh-Metrix readability indicators (Baidu Translate)

Evaluation Indicator	Students (HT)	Baidu+Students (Post-editing)	Comparison of Evaluation Indicators (p-value in parentheses)
Narrativity	–1.91	–1.97	HT > Post-editing (.298)
Syntactic Simplicity	–1.33	–0.79	HT < Post-editing (.005)***
Word Concreteness	1.08	0.86	HT > Post-editing (.177)
Referential Cohesion	0.91	1.40	HT < Post-editing (.008)***
Deep Cohesion	–0.83	–0.48	HT < Post-editing (.004)***
Verb Cohesion	–0.18	–0.19	HT > Post-editing (.919)
Connectivity	–2.75	–2.63	HT < Post-editing (.207)
Temporality	–0.46	–0.13	HT < Post-editing (.096)

*** represents the 1% significance level.

These results suggest that Baidu can improve the readability of the students' translations by adjusting the syntactic simplicity and text cohesion of their translations. We describe the cohesion measures when we examine Google Translate in section 4.4 as the reasoning is the same. For the syntactic simplicity test, Graesser et al. (2011) pointed out that lower syntactic simplicity signifies a longer sentence length and a more complex sentence structure. We examined several examples in our corpora[i] and found that to express the same Chinese information, the students used longer and more complex sentences in the human translation mode than in the post-edited translation mode (Baidu). This is consistent with the statistical analysis presented in Table 7 that the syntactic simplicity of the Baidu corpus (–0.79) is significantly higher than that of the HT corpus (–1.33). The results suggest that Baidu can help the students to pay more attention to the sentence segmentation method so as to avoid too many complex sentences and make the reading process less challenging.

4.4 Google Translate

Table 8 presents the evaluation results for the Google corpus and the HT corpus for the two basic indicators and the two traditional readability indicators. The results in the last column show that there are significant differences between the two corpora in several dimensions. In particular, the sentence length of the Google corpus is significantly shorter and the LIX index is significantly lower than those of their counterparts in the HT corpus, while the Flesch Reading Ease index of the Google corpus is significantly higher.

Table 8

Basic descriptive indicators and traditional readability indicators (Google Translate)

Evaluation Indicator	Students (HT)	Google+Students (Post-editing)	Comparison of Evaluation Indicators (p-value in parentheses)
Word Length (number of letters)	5.51	5.54	HT < Post-editing (.748)
Sentence Length (number of words)	23.21	17.38	HT > Post-editing (.000)***
LIX Index	52.54	49.33	HT > Post-editing (.000)***
Flesch Reading Ease	22.53	29.39	HT < Post-editing (.000)***

*** represents the 1% significance level.

In this study, of the four corpora of post-edited subtitle translations, the Google corpus is the only one that is significantly different from the HT corpus on sentence length and the two traditional readability indicators. In general, the results indicate that the Google corpus has the highest readability level. Specifically, for sentence length, a low score is consistent with a high level of readability. Leong et al. (2002) found that a sentence comprising 20–25 words is more suitable for readers at an advanced reading level. Considering the inconvenience of reading from a screen, it is generally believed that the sentences in subtitle translations should contain fewer than 18 words each. Whereas the sentence average length of the Google corpus is 17.38, it is 23.21 for the HT corpus. This indicates that the students used relatively long and complex sentences in their translations without the use of MT, sentences which are difficult for general users to read. For the LIX index, a high score is consistent with a low level of readability and a score over 50 is considered to indicate a difficult read (Courtis, 1995), so the HT corpus's score on this index (52.54) indicates an inappropriate readability level. For the Flesch indicator, the reasoning is the same as for the Baidu corpus.

Table 9 presents the evaluation results for the Google corpus and the HT corpus for the eight Coh-Metrix indicators. The results show that the Google corpus differs significantly from the HT counterpart in all of these indicators.

Table 9

Coh-Metrix readability indicators (Google Translate)

Evaluation Indicator	Students (HT)	Google+Students (Post-editing)	Comparison of Evaluation Indicators (p-value in parentheses)
Narrativity	–1.91	–1.42	HT < Post-editing (.000)***
Syntactic Simplicity	–1.33	–0.90	HT < Post-editing (.008)***
Word Concreteness	1.08	1.55	HT < Post-editing (.000)***
Referential Cohesion	0.91	1.45	HT < Post-editing (.000)***
Deep Cohesion	–0.83	–0.36	HT < Post-editing (.000)***
Verb Cohesion	–0.18	0.11	HT < Post-editing (.008)***
Connectivity	–2.75	–1.72	HT < Post-editing (.000)***
Temporality	–0.46	0.12	HT < Post-editing (.009)***

*** represents the 1% significance level.

The evaluation scores indicate that Google Translate can help students to enhance their performance on the eight indicators. According to Jiang and Han (2018), these indicators correspond to the five levels identified by Graesser et al. (2011): word level (word concreteness); syntax level (syntactic simplicity); textbase (referential cohesion); the situation model (deep cohesion, verb cohesion, connectivity, temporality); and the genre and rhetorical structure (narrativity). For indicators at the word and syntax levels, the reasoning is the same as for the Bing and Baidu corpora. Example (3) below therefore focuses on the level of the textbase and the situation model.[ii] Specifically, we illustrate the differences in the cohesion and connectivity indicators between the HT corpus and its post-edited counterpart (Google) by comparing their outputs from a sample subtitle translation.

Example (3): Cohesion indicators and connectivity

Chinese text in the video (Tencent): 腾讯希望成为各行各业的数字化助手，助力数字中国建设。我们希望在制造业、医疗、零售、教育等各个领域，使用数字创新提升每个人的生活品质。随着数字经济的发展，我们通过战略合作与开放平台，与合作伙伴共建数字生态共同体，推进云计算、大数据、人工智能等前沿科技与各行各业的融合发展及创新共赢。多年来，腾讯的开放生态带动社会创业就业人次达数千万。

Human Subtitle Translation: (1) Tencent hopes to become a digital assistant in all walks of life and help build digital China. (2) And, we hope to use digital innovation in the fields of industry, medical care, retail and education to enhance the quality of life for everyone. (3) With the development of the digital economy, our company has built a digital ecological community with partners through strategic cooperation and open platforms, and promoted the cutting-edge technologies such as cloud computing, big data, and artificial intelligence. (4) Over the years, Tencent's open ecology has brought success to tens of millions of social entrepreneurs and employers.

> Post-edited Translation (Google): (6) Tencent strives to be the driving force behind China's digitalization in various industries. (7) Moreover, the company is passionate about improving people's quality of life through digital innovation in the fields of manufacturing, healthcare, retail and education. (8) In addition, as the digital economy gathers pace, Tencent is co-creating an ecosystem with partners through strategic collaboration and open platforms. (9) These partnerships have advanced the integration and development of cutting-edge technologies, including cloud computing, big data, and artificial intelligence. (10) As a result, individuals and businesses of our economy are already realizing the benefits of Tencent's Open Ecosystem, which has helped tens of millions of entrepreneurs establish new businesses and create new jobs over the years.

The three cohesion indicators (referential cohesion, deep cohesion and verb cohesion) reflect the degree to which the text contains words and ideas that overlap across sentences and the entire text, while the connectivity index reflects the degree to which the text contains connectives that express relations in the text (McNamara et al., 2014). As shown in Example (3), when discussing future development goals, the overlapping frequency of arguments between sentences in post-edited translation is much higher than that in human translation. For example, the argument "Tencent" appears in sentences (6), (8) and (10), and the word "economy" overlaps in both sentences (8) and (10). However, the same student used a relatively small number of argument overlaps in the human translation, and the argument "Tencent" is used repeatedly only in sentences (1) and (4).

Second, the post-edited translation also uses overlapping synonyms between sentences, such as "collaboration" in sentence (8) and "partnerships" in sentence (9); "company" in sentence (7) and "business" in sentence (10). However, there is no overlapping of synonyms between sentences in the human translation.

Third, in the post-edited translation, the student used many connectives, such as "moreover" in sentence (7), "in addition" in sentence (8), and "as a result" in sentence (10), while only "and" appears in sentence (2) in the HT translation.

To sum up, Google Translate helps the students to use more measures in a translation to ensure cohesion and improve the readability of a text. The statistical test results in Table 9 are consistent with the finding in the case study that Google Translate significantly improves the performance of student translations based on the indicators of connectivity and cohesion. In addition, Google Translate also helps the students to improve the accuracy of terminology translation. For instance, in sentence (2), the student mistakenly translated 制造业 (zhì zào yè) as "industry" in the human translation mode, but with the help of Google Translate, they changed the translation to "manufacturing" in sentence (7), which is the correct term in English.

5. Conclusion and future research

Previous studies suggest that both students and teachers can benefit from the use of technology in AVT education. However, owing to the short launch time, the existing research has paid little attention to the application of NMT in AVT teaching. To contribute towards filling this gap, this study examined the effectiveness of NMT in enhancing the readability of Chinese–English subtitle translations in AVT teaching. Specifically, we constructed parallel corpora that contained human subtitle translations and the corresponding post-edited translations provided by both students and four NMT-based online platforms. With the help of two corpus analysis tools, we then calculated 12 indicators for each corpus. Finally, we used SPSS 24.0 to conduct statistical analyses of the evaluation indicators between the human and the post-edited subtitle translations.

Although all four systems are NMT-based, our statistical analysis suggests that Google Translate outperforms the other three platforms in all readability tests, and it can enhance the readability of post-edited subtitle translation across the five levels (word, syntax, textbase, situation model, genre and rhetorical) identified by Graesser et al. (2011). The remaining three platforms performed differently in the tests. Specifically, Bing Translator can improve the readability of students' subtitle translation at the word level, Youdao Translate can improve the readability at the textbase level, and Baidu Translate can improve the readability at the syntax and the textbase levels. Further analyses based on concrete examples verified the empirical results.

The study has two major implications. First, the empirical results confirmed the significant role that NMT technology can play in enhancing the readability of students' Chinese–English subtitle translation. Therefore, teachers should consider using online MT systems based on NMT technology as an auxiliary tool in AVT teaching. Regarding improving the readability of Chinese-to-English subtitle translation, based on our findings, Google Translate is the best choice. However, if Google Translate cannot be accessed, teachers can select an MT system according to the focus of their teaching. If the focus of subtitle translation teaching is to improve readability at the word level, Bing Translator is a better choice than Youdao and Baidu, whereas if the focus is on the syntax level, Baidu is a better choice than Bing and Youdao. What is equally important is that the empirical findings and case studies together provide directions for improvements to MT systems. Specifically, to enhance the readability performance of MT systems, the developers of Bing, Youdao and Baidu may consider working on adjusting the algorithms at the genre and rhetorical levels, and also the level of the situation model. Moreover, the developer of Youdao should pay additional attention to algorithms at the lexical and syntax levels, and the developer of Bing should focus on the syntax and textbase levels, while the developer of Baidu may consider working on adjusting the algorithms at the word level.

Wang, Q. & Xu, J. (2023). Neural machine translation in AVT teaching in China: An in-depth analysis from the readability perspective. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *22*, 161–180.

Our analysis has two limitations, however. First, the text genre and language type of the research sample are relatively homogeneous, involving only Chinese–English corporate text. Second, there are many methods and indicators for translation quality evaluation; this study has selected only 12 indicators to assess the performances of four NMT-based systems in AVT teaching from the readability perspective; these may not fully reflect the NMT quality.

In the future, we plan to perform the tests on other language pairs and to develop new evaluation frameworks pertaining to different theoretical perspectives in order to test other text genres in the AVT environment. Translations in different languages and genres vary greatly in style, method, and technique, and therefore the evaluation frameworks cannot be identical. Consequently, future research could expand the genre and language ranges to examine NMT outputs in the AVT Environment comprehensively.

References

Abdullah, Z., Md Nordin, S., & Abdul Aziz, Y. (2013). Building a unique online corporate identity. *Marketing intelligence & planning, 31*(5), 451–471. https://doi.org/10.1108/MIP-04-2013-0057

Alva-Manchego, F., & Shardlow, M. (2022). Towards readability-controlled machine translation of COVID-19 texts. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pp. 287–288. https://aclanthology.org/2022.eamt-1.33

Banerjee S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Valuation Measures for MT and/or Summarization*, pp. 65–72. https://aclanthology.org/W05-0909.pdf

Bellés-Calvera, L., & Quintana, R. C. (2021). Audiovisual translation through NMT and subtitling in the Netflix series 'cable girls'. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pp. 142–148. https://doi.org/10.26615/978-954-452-071-7_015

Bywood, L., Georgakopoulou, P., & Etchegoyhen, T. (2017). Embracing the threat: Machine translation as a solution for subtitling. *Perspectives*, *25*(3), 492–508. https://doi.org/10.1080/0907676X.2017.1291695

Chatzikoumi, E. (2020). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering, 26*(2), 137–161. https://doi.org/10.1017/S1351324919000469

Ciobanu, M., Dinu, L., & Pepelea, F. (2015). Readability assessment of translated texts. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pp. 97–103. https://aclanthology.org/R15-1014.pdf

Courtis, K. (1995). Readability of annual reports: Western versus Asian evidence. *Accounting, Auditing & Accountability Journal, 8*(2), 4–17. https://doi.org/10.1108/09513579510086795

Graesser, A. C., & McNamara, D. S. (2011). Computational analysis of multilevel discourse comprehension. *Topics in Cognitive Science, 3*(2), 371–398. https://doi.org/10.1111/j.1756-8765.2010.01081.x

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223–234. https://doi.org/10.3102/0013189X11413260

Giménez, J. & Márquez, L. (2010). Linguistic measures for automatic machine translation evaluation. *Machine Translation, 24*, 209–240. https://link.springer.com/article/10.1007/s10590-011-908

Wang, Q. & Xu, J. (2023). Neural machine translation in AVT teaching in China: An in-depth analysis from the readability perspective. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *22*, 161–180.

8-7

Han, Y., & Meng, S. (2022). Machine English translation evaluation system based on BP neural network algorithm. *Computational Intelligence & Neuroscience, 2022*, 1–10. https://doi.org/10.1155/2022/4974579

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., … Zhou, M. (2018). Achieving human parity on automatic Chinese to English news translation. arXiv:1803.05567. https://arxiv.org/abs/1803.05567

Jia, Y., Carl, M., & Wang, X. (2019). Post-editing neural machine translation versus phrase-based machine translation for English-Chinese. *Machine Translation*, *33*, 9–29. https://doi.org/10.1007/s10590-019-09229-6

Jiang, J., & Han, B. (2018). A study of reading text difficulty of CET6, TOEFL and IELTS based on Coh-Metrix. *Foreign Languages in China, 3*, 86–95.

Jones, D., Gibson, E., Shen, W., Granoien, N., Herzog, M., Reynolds, D., & Weinstein, C. (2005). Measuring human readability of machine generated text: Three case studies in speech recognition and machine translation. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1009–1012. https://ieeexplore.ieee.org/document/1416477

Leong, K., Michael T., & Leyland, F. (2002). E-comprehension: Evaluating B2B websites using readability formulae. *Industrial Marketing Management, 31*(2), 125–131. https://doi.org/10.1016/S0019-8501(01)00184-5

Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. *Journal of Finance, 69*(4), 1643–1671. https://doi.org/10.1111/jofi.12162

Matusov, E., Wilken, P., & Georgakopoulou, Y. (2019). Customizing a neural machine translation system for the translation of subtitles in the entertainment domain. In *Proceedings of the Fourth Conference on Machine Translation,* pp. 82–93. https://aclanthology.org/W19-5209.pdf

McNamara, D., Graesser, A., McCarthy, P, & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

Mobarakeh, M. D., & Sardareh, S. A. (2016). The effect of translation shifts on the level of readability of two Persian translations of novel “1984” by George Orwell. *International Journal of Humanities and Cultural Studies*, *2016*, 1418–1427.

Olive, J., Christianson, C, & McCary, J. (2011). *Handbook of natural language processing and machine translation.* Springer Science & Business Media. https://doi.org/10.1007/978-1-4419-7713-7

Östling, R., & Tiedemann, J. (2017). Neural machine translation for low-resource languages. arXiv:1708.05729. https://doi.org/10.48550/arXiv.1708.05729

Popović, M. (2017). Comparing language related issues for NMT and PBMT between German and English. *The Prague Bulletin of Mathematical Linguistics*, *108*, 209–220. https://doi.org/10.1515/pralin-2017-0021

Revanuru, K., Turlapaty, K., & Rao, S. (2017). Neural machine translation of Indian languages. In *Proceedings of the 10th annual ACM India compute conference*, pp. 11–20. https://doi.org/10.1145/3140107.3140111

Sanchez-Torron, M., & Koehn, P. (2016). Machine translation quality and post-editor productivity. In *Proceedings of the Association for Machine Translation in the Americas: MT Researchers' Track*, pp. 16–26. https://aclanthology.org/2016.amta-researchers.2.pdf

Scott, J. (2015). A re-examination of Fortune 500 homepage design practices. *IEEE Transactions on Professional Communication*, *58*(1), 20–44. https://doi.org/10.1109/TPC.2015.2420371

Shi, X., & Shan, X. (2019). A corpus-based study on linguistic and cross-cultural adaptation of Chinese corporate websites. *Foreign Languages in China*, *2*, 71–80.

Smith, M., & Taffler, R. (1992). Readability and understandability: Different measures of the textual

complexity of accounting narrative. *Accounting, Auditing & Accountability Journal, 5*(4), 84–98. https://doi.org/10.1108/09513579210019549

Song, K., Wang, K., Yu, H., Zhang, Y., Huang, Z., Luo, W., & Zhang, M. (2020). Alignment-enhanced transformer for constraining NMT with pre-specified translations. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(05), 8886–8893. https://doi.org/10.1609/aaai.v34i05.6418

Tardel, A. (2021). Measuring effort in subprocesses of subtitling: The case of post-editing via pivot language. In M. Carl (Ed.), *Explorations in empirical translation process research* (pp. 81–110). Springer. https://doi.org/10.1007/978-3-030-69777-8_4

van Toledo, C., Schraagen, M., van Dijk, F., Brinkhuis, M., & Spruit, M. (2023). Readability metrics for machine translation in Dutch: Google vs. Azure & IBM. *Applied Sciences*, *13*(7), 1–14. https://doi.org/10.3390/app13074444

Wang, L., & Liang, M. (2007). Applying WordSmith tools in studies of second languages. *Technology Enhanced Foreign Language*, *3*, 3–7.

Wang, Q., & Xu, J. (2022). Analysis of the impact of machine translation technology on the language service industry from the perspective of technological progress. *Foreign Languages in China*, *1*, 21–29.

Wang, Q., & Xu, J. (2023). Interdisciplinary research on translation teaching and talent training in the new era: Taking business translation teaching as an example. *Foreign Languages in China*, *4*, 89–97.

[i] Owing to limited space, examples are not presented here, but they are available from the authors upon request.

[ii] Owing to limited space, analyses of temporality and narrativity are not presented here, but they available from the authors upon request.