*transLogos* 2024 Vol 7 Issue 2
Zhou, Xu, and Jinghua Zhou, pp. 29–56
https://dx.doi.org/10.29228/transLogos.68



# A Study on the Quality Assessment of the Chinese-English Subtitle Translation among Four Machine Translation Tools - A Case Study of *Wreaths at the Foot of the Mountain*

Xu ZHOU* and Jinghua ZHOU**

As artificial intelligence (AI) continues to evolve, machine translation (MT), particularly neural MT, is experiencing rapid development and innovation, significantly influencing leading MT applications in China. This evolution comes at a pivotal moment when Chinese films are going global at a faster pace, highlighting the importance of effective subtitle translation. However, despite these advancements, there remains a considerable gap in the quality of subtitle translations. The existing two major subtitle translation assessment models—Multidimensional Quality Metrics (MQM) and Functional Equivalence, Acceptability, and Readability (FAR)—face several challenges that hinder their effectiveness. In light of these issues, this study aims to integrate the strengths of both MQM and FAR to develop an enhanced assessment framework, referred to as the FAR 2.0 model. This new model will facilitate a more comprehensive comparison and analysis of four prominent MT applications discussed in this paper, focusing on three critical aspects: functional equivalence (F), acceptance (A), and readability (R). By applying this model, the study seeks to illuminate the respective strengths and weaknesses of each MT tool, offering insights that can guide quality improvements in subtitle translation both now and in the future. Ultimately, this research not only aims to enhance the understanding of current MT capabilities, but also strives to contribute to the broader goal of elevating the global competitiveness of Chinese films through improved subtitle quality.
Keywords: machine translation; subtitle translation; MQM; FAR 2.0 model; MT tool

## 1. Introduction

Machine translation (MT) refers to the automated conversion of a source language into a target language using computer technologies (Liu, Lü, and Liu 2024, 82). The concept of

* Post-graduate of Master Degree on English Written Translation major at the School of Law and Humanities, China University of Mining and Technology-Beijing.
E-mail: zxbyqx@163.com; ORCID ID: https://orcid.org/0009-0001-5133-3724.
** Post-graduate of Master Degree on English Written Translation major at the School of Law and Humanities, China University of Mining and Technology-Beijing.
E-mail: zjh3062494@163.com; ORCID ID: https://orcid.org/0009-0002-9615-8736.
(Received 17 October 2024; accepted 13 December 2024)

machine translation was first proposed by Georges B. Artsrouni in 1930, followed by the emergence of ENIAC, the world's first modern electronic digital computer, at the University of Pennsylvania in 1946. Since then, rule-based and example-based systems laid the foundation for modern machine translation, with statistical models significantly improving translation accuracy and ushering in an era of rapid growth.

Following the introduction of deep belief networks (DBNs) in 2006, deep neural networks began to transform machine translation. The 2014 proposal of the encoder-decoder model marked the advent of the deep learning era for machine translation (Feng and Shao 2020). In 2017, Ashish Vaswani's introduction of the Transformer model, which employs attention mechanisms, further established neural networks as the dominant approach in machine translation.

Released on November 30, 2022, ChatGPT, a natural language processing tool built on the Transformer framework, exhibited impressive translation capabilities in several languages. On May 13, 2024, OpenAI launched ChatGPT-4o, capable of processing and generating text, audio, and images. It can respond to audio inputs in as little as 232 milliseconds, with an average response time of 320 milliseconds, matching human conversational speeds and marking significant progress in machine translation technology.

Huang Youyi noted that China has experienced two translation booms since the initiation of the reform and opening-up policy in 1978. The first boom sought to introduce the wider world to China, while the second aims to facilitate the implementation of the "going global" strategy. A new trend has since emerged—"translating China," with a view to translating Chinese culture, history, and economic developments for global audiences. The shift from "translating the world" to "translating China" reflects an inevitable trend (Youyi 2022, 158). For China, it symbolizes the nation's growing composite strength, and for Chinese translators, it signifies an era-defining mission.

As China continues to open itself to the world, its film and television industry has increasingly attracted international audiences, aided by relevant policies. Films such as *My People and My Country*, *Wolf Warrior II*, and *The Wandering Earth* offer foreign viewers

insights into China's politics, economy, and culture. According to the *2023-2029 Market Operation Status and Investment Strategic Planning Report for China's Machine Translation Industry* (2023) by Chyxx, a renowned Chinese think tank, the value of China's machine translation industry totaled 460 million RMB in 2016. This figure rose to 7.94 billion RMB in 2021 and reached 16.6 billion RMB by 2023.

The potential of the subtitle translation market for Chinese films and television is immense. However, the number of human translators is insufficient to meet the growing demand. At the same time, machine translation faces persistent challenges, such as word-for-word translation, that limit its quality. Against this backdrop, this paper compares four popular translation applications—LingoCloud, DeepL, Baidu Translate, and Youdao Translate—using the classic Chinese film *Wreaths at the Foot of the Mountain* (1984) as the research corpus. The objective is to provide insights into improving the quality of machine translations.

## 2. Literature Review

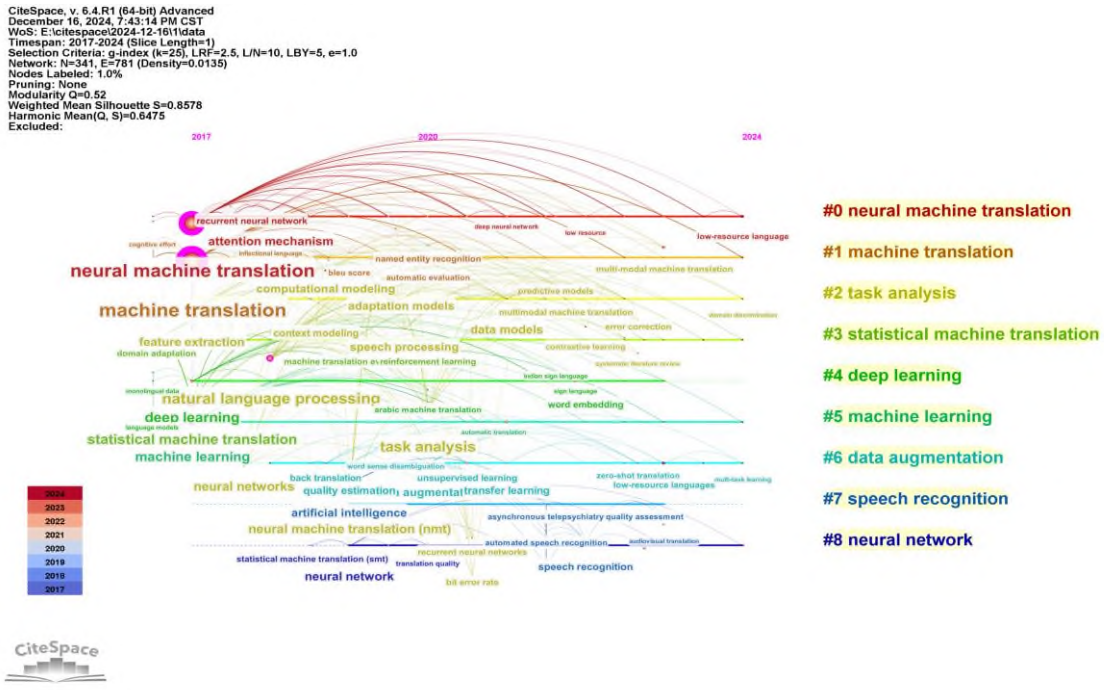
In recent years, both domestic and international scholars have increasingly explored the field of machine translation, proposing various methods to enhance translation models and improve evaluation systems for translation quality. To obtain research data from the past eight years, two statistical graphs were generated using CiteSpace (version 6.3.R1), a visual scientific research software. These graphs draw on data from CNKI, a major Chinese academic journal database, and Web of Science, an international counterpart, spanning the period from January 1, 2017, to May 26, 2024. The data collection employed "keywords" from Chinese journals and both "keywords" and "titles" from English journals as thematic search terms.

This chapter reviews the current state of machine translation research in both foreign and domestic contexts.

## 2.1 Review of Foreign Literature

A dataset was generated from the "Core Collection" section of Web of Science using the keywords "machine translation" and "neural MT." After cleaning and filtering the data, 785 representative entries were retained. The data were visualized in CiteSpace using its keyword clustering function, which revealed that foreign research primarily focuses on "machine translation," "neural machine translation," "deep learning," and "artificial intelligence" (see figure 1).

Figure 1. Hot topics in the field of machine translation in foreign research (2017–2024)




As shown in figure 1, machine translation research has evolved since 2017, shifting from focusing on individual technologies to exploring multiple dimensions. The rise of neural machine translation (NMT) models marked a significant shift, challenging traditional methods that relied on large parallel corpora for training NMT and statistical machine translation (SMT) systems (Artetxe, Labaka, and Agirre 2019). Compared with SMT, NMT improves translation accuracy, with the Transformer model becoming the standard for high-quality translations (Garg et al. 2019).

Additionally, Song Linfeng et al. (2019) proposed that abstract meaning representation (AMR) could enhance attention-based sequence-to-sequence NMT models by improving meaning preservation and addressing data sparsity issues.[1] Although SMT dominated the field for decades, it has been overtaken by NMT's neural network architecture (Stahlberg 2020). However, large multilingual NMT models still face challenges, such as underperformance compared with bilingual models and difficulties in zero-shot translation[2] (Zhang et al. 2020).

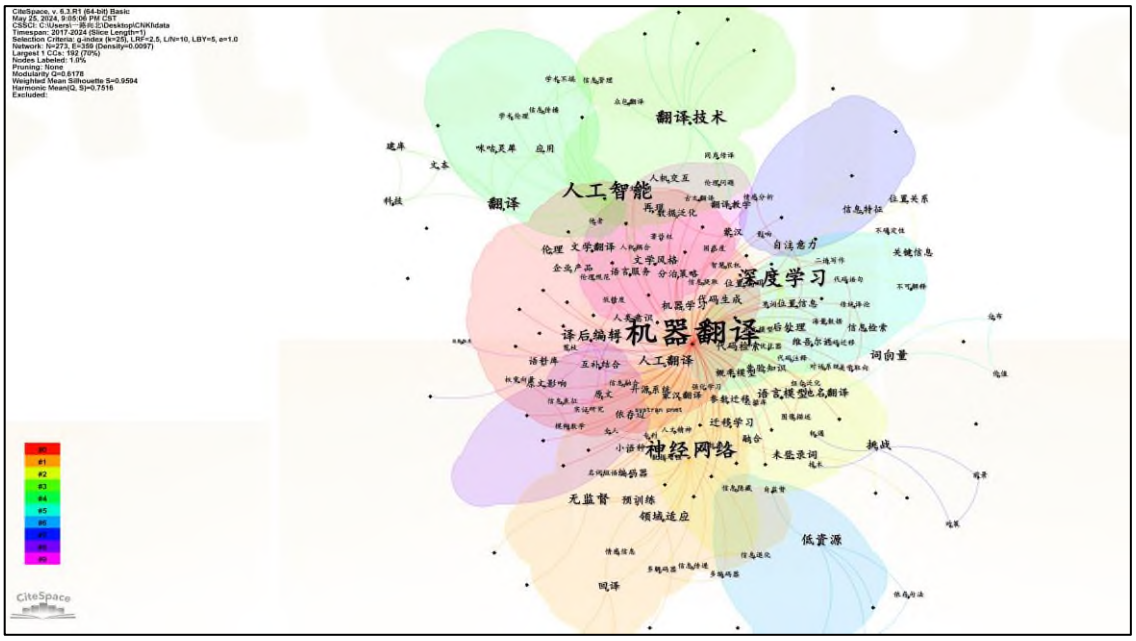
Machine translation evaluation remains a critical area in need of improvement. Current metrics are highly sensitive to translations for evaluation, which could result in misleading conclusions (Mathur, Baldwin, and Cohn 2020). Systems like CUBBITT have greatly improved translation quality, and, in some cases, even outperformed human translators, raising the possibility that deep learning may eventually replace human translators in certain contexts (Popel et al. 2020). However, low-resource languages and multilingual translations continue to present challenges due to the lack of robust evaluation benchmarks, prompting Naman Goyal et al. (2021) to introduce the FLORES-101 benchmark to address these issues.[3] Furthermore, Amr Hendy et al. (2023) argued that GPT models show potential in machine translation evaluation and that integrating GPT with other systems could further enhance translation quality.

## 2.2 Review of Domestic Literature

A similar search was conducted using CNKI with the keywords "machine translation" and "neural MT." After filtering, 551 valid entries were obtained from an initial set of 611. These data were then visualized using CiteSpace, producing the graph shown in figure 2.

[1] 'Data sparsity' refers to several sentences matching one meaning.
[2] 'Zero-shot translation' refers to the ability of a machine translation (MT) system, typically a neural machine translation (NMT) model, to translate between language pairs it has not explicitly been trained on.
[3] 'Low-resource languages' refer to languages that lack sufficient linguistic data, such as parallel corpora, dictionaries, or annotated texts.
'FLORES-101 benchmark' is a multilingual evaluation dataset developed by Facebook AI Research (FAIR) to assess the performance of MT models across low-resource and diverse languages, which provides a comprehensive benchmark for testing translation quality between 101 languages, including many low-resource languages that are often underrepresented in MT research.

Figure 2. Hot topics in the field of machine translation in Chinese research (2017–2024)




As indicated in figure 2, Chinese research on machine translation has expanded from a narrow focus to encompass broader themes, including 'machine translation,' 'neural networks,' 'deep learning,' 'translation technology,' and 'artificial intelligence.'

In the field of neural networks, Li Fengqi (2021) conducted a quantitative study, evaluating five major online machine translation systems by testing them with narrative, descriptive, explanatory, and argumentative texts to assess their performance in Chinese-English translation. Wang Jinquan and He Bojia (2024) incorporated Word2vec and Doc2vec models, along with N-grams, semantic points, and text similarity metrics, into their evaluation system to assess the quality of Chinese-English translations across various text types.

In the field of deep learning, Xu Wei (2024) developed an optimized translation model for the smart agricultural machinery sector, improving translation efficiency and accuracy. Back-translation methods have also been applied to segmental training of translation models using monolingual data, leading to improved Uygur-Chinese and Mongolian-Chinese translations under low-resource conditions (Zhang, Zhang, and Yang 2021).

In the area of artificial intelligence, Wen Xu and Tian Yaling (2024) used the report to the 20th National Congress of the Communist Party of China as their corpus, comparing the

translation quality of ChatGPT, Google Translate, Youdao Translate, and DeepL. Their study employed both BLEU and TER metrics, along with manual evaluation, to assess ChatGPT's effectiveness in translating political documents. Additionally, Zhou Xinghua and Wang Chuanying (2020) evaluated Memsource and SDL Trados Studio by focusing on their non-translation element functions, machine translation quality evaluation capabilities, and adaptive translation features.

Overall, domestic research on machine translation quality evaluation has made considerable progress. However, there remains a gap in research focusing specifically on the quality of subtitle translation, particularly in Chinese-English contexts. Moreover, the existing evaluation frameworks do not fully address the unique characteristics of subtitle texts.

## 3. Subtitle Processing and Evaluation Framework

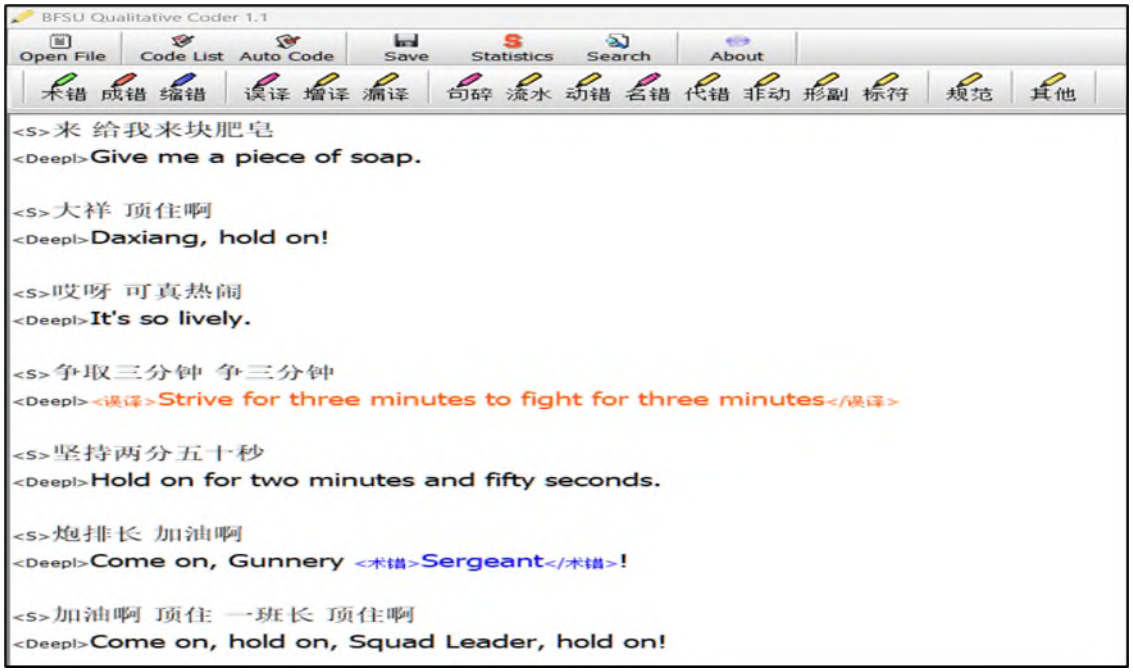
This chapter details the preparation, annotation, and evaluation processes used to analyze the film subtitle corpus. Both qualitative and quantitative approaches were adopted to assess translation quality systematically.

### 3.1 Corpus Preparation and Annotation

This study selects the 1984 Chinese film *Wreaths at the Foot of the Mountain* as the research corpus, as it currently lacks official English subtitles. The original film subtitles contain 2,077 segments, totaling 12,485 characters. After formatting was removed, the cleaned text was processed through four MT tools: LingoCloud, DeepL, Baidu Translate, and Youdao Translate.

The translated outputs were aligned into bilingual texts with parallel Chinese-English lines, as illustrated in figure 4. This alignment was performed to facilitate subsequent manual annotation using the BFSU Qualitative Coder 1.2 software, a qualitative data analysis software developed by Beijing Foreign Studies University (BFSU) (figure 3).

Figure 3. The surface of the BFSU Qualitative Coder 1.2 software

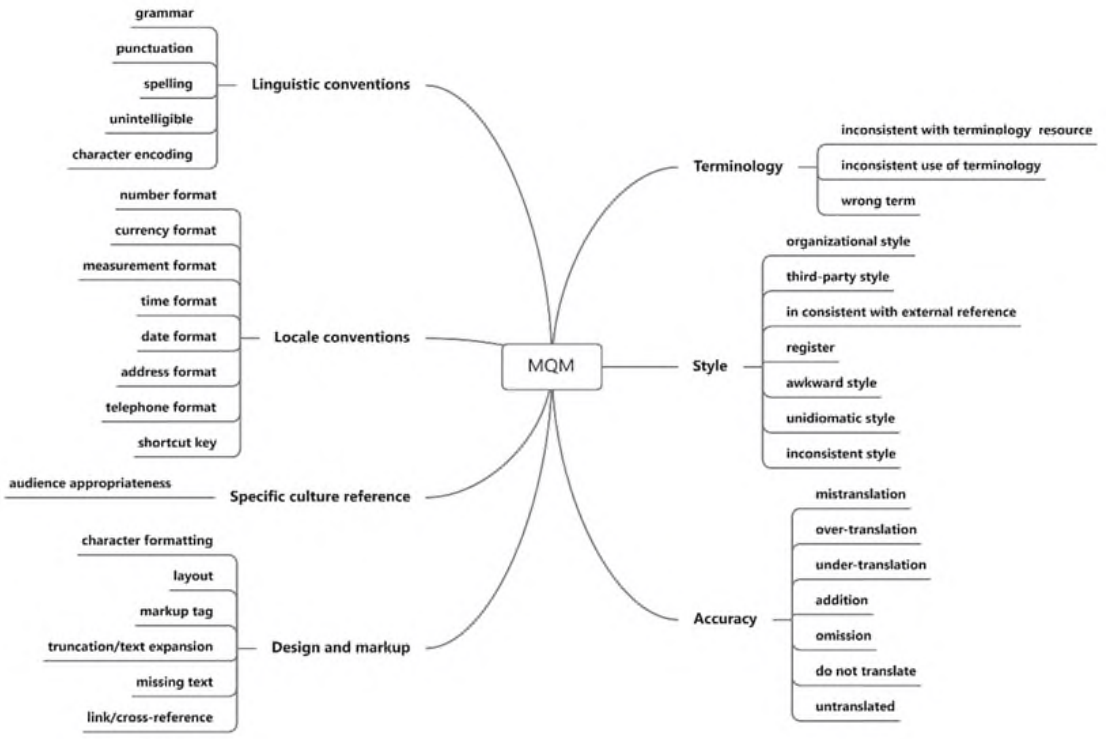



3.2 Methodology and Evaluation Framework

This study combines both qualitative and quantitative analyses to evaluate the accuracy and performance of the four MT tools. After annotating the outputs using the manual annotation tool, four statistical reports were generated to provide detailed performance metrics for each tool.

The Multidimensional Quality Metrics (MQM) framework, developed by the German Research Center for Artificial Intelligence (DFKI), evaluates translation performance across seven dimensions: terminology, accuracy, linguistic conventions, style, locale conventions, audience appropriateness, and design and markup (figure 4). It accounts for 38 specific error types, offering a highly flexible and comprehensive system for translation quality assessment (Lommel, Burchardt, and Uszkoreit 2013).

Figure 4. The content of the MQM



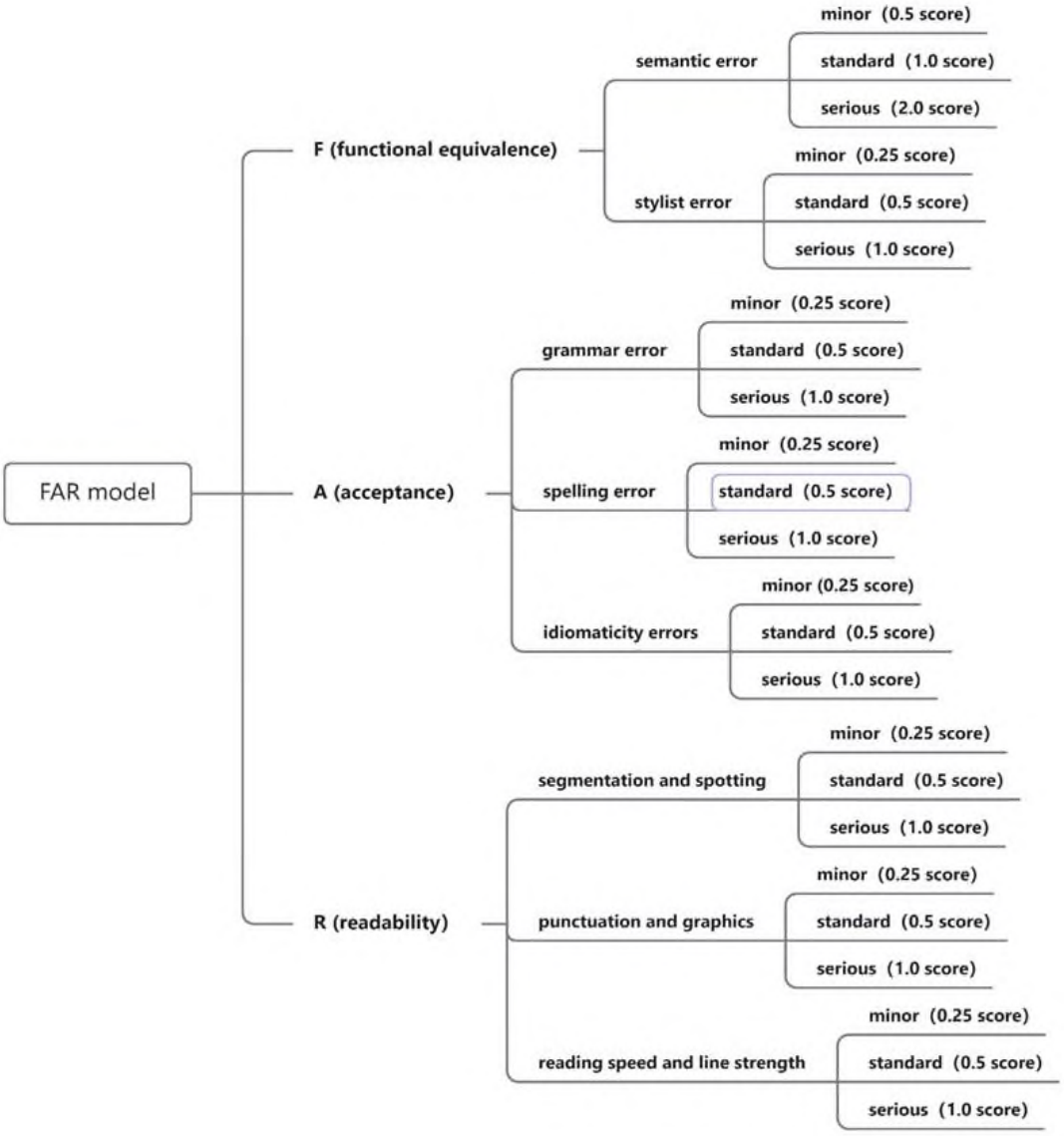

While the MQM framework offers detailed insights, it introduces overlapping error categories that complicate annotation. For example, "under-translation" and "mistranslation" fall under the "accuracy" dimension, and "awkward" and "unidiomatic" style errors are both part of the "style" dimension. Such overlaps can result in double-counting during annotation.

To address the limitations of MQM, this study adopts the FAR model developed by Jan Pedersen (2017) for interlingual subtitling. The FAR model evaluates subtitle translations across three error levels:

- minor errors: small issues like misspellings or missing punctuation.
- standard errors: errors that disrupt most readers' experience.
- serious errors: issues that impair comprehension of multiple subtitles.

The original FAR model applies error deductions on a sentence-by-sentence basis, making it suitable for assessing machine-translated subtitles (figure 5). However, to better align it with the MQM framework and improve clarity, this study introduces a modified version—FAR 2.0.

Figure 5. The content of the FAR model

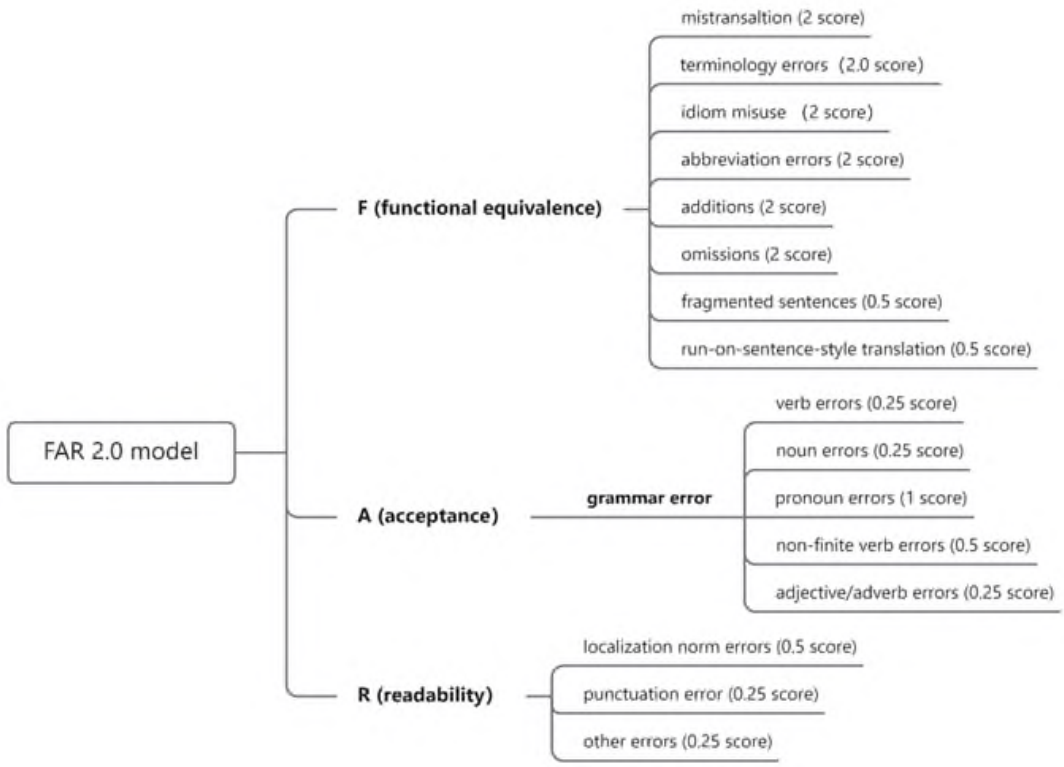



The adjustments to FAR 2.0 reflect both practical translation insights and MQM's detailed framework. Changes were made to three dimensions: functional equivalence, acceptability, and readability.

1. Functional Equivalence:
    a. The "stylist error" category was removed. In Chinese-English subtitle translation, differences between formal and informal address (e.g., 你 [*ni*, you] and 您 [*nin*, you]) are often neutralized in English as "you" (Xiao 2017).
    b. The "semantic errors" category was refined into subcategories: terminology errors, idiom misuse, abbreviation errors, additions, omissions, fragmented sentences, run-on-sentence-style translation, and mistranslations. These were assigned severity-based scores to reflect their impact on comprehension.
2. Acceptability:
    a. The "idiomaticity errors" category was removed, as very few errors of this type were identified during annotation.
    b. The "grammar errors" category was subdivided into more specific subcategories: verb errors, noun errors, pronoun errors, non-finite verb errors, and adjective/adverb errors. Scores were assigned based on the severity of the identified errors.
3. Readability:
    a. Since this study does not analyze subtitle-video synchronization, the categories "segmentation and spotting" and "reading speed and line length" were excluded.
    b. Instead, two new categories—"localization norm errors" and "other errors"—were added to capture any additional issues encountered during the annotation process.

The structure of the FAR 2.0 model is shown in figure 6.

Figure 6. The structure of the FAR 2.0 model




The quality of subtitle translations is evaluated using the following formula:

$$FAR = \frac{\mathrm{N} - \mathrm{F} - \mathrm{A} - \mathrm{R}}{N} \times 100\%$$

"N" represents the total number of subtitles, and F, A, and R refer to the total deductions for Functional Equivalence, Acceptability, and Readability errors, respectively. The deduction for each dimension is calculated as:

$$F/A/R = \frac{\mathrm{N} - F/A/R}{\mathrm{N}} \times 100\%$$

This scoring system ensures a precise comparison of the four MT tools, with the refined FAR 2.0 model addressing the overlapping error categories identified in the MQM framework.

## 4. Results and Analysis

To ensure the accuracy of error annotation, all four machine translations in this study were manually annotated. The scores were given according to "FAR Translation Quality

Assessment Model 2.0," followed by thorough proofreading to avoid omissions or mistakes. The results are illustrated in figure 7. This part will analyze in detail the scores of these four tools across the three dimensions under the "FAR Translation Quality Assessment Model 2.0," along with the causes of the errors.

Figure 7. The results of the four MT tools

| | subcatgories | score | LingoCloud | | | | DeepL | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | number of errors | total scores | F/A/R | F/A/R (%) | number of errors | scores | F/A/R | F/A/R (%) |
| F | mistransaltion | 2 | 248 | 496.00 | 1317.50 | 63.40% | 145 | 290.00 | 769.00 | 62.98% |
| | terminology error | 2 | 36 | 72.00 | | | 89 | 178.00 | | |
| | idiom error | 2 | 43 | 46.00 | | | 111 | 222.00 | | |
| | abbreviation error | 2 | 0 | 0.00 | | | 0 | 0.00 | | |
| | additions | 1 | 18 | 18.00 | | | 19 | 19.00 | | |
| | omissions | 2 | 41 | 82.00 | | | 30 | 60.00 | | |
| | fragmented sentences | 0.5 | 11 | 5.50 | | | 0 | 0.00 | | |
| | run-on-sentence-style translation | 0.5 | 0 | 0.00 | | | 0 | 0.00 | | |
| A | verb error | 0.25 | 17 | 4.25 | 44.00 | 9.80% | 27 | 6.75 | 76.75 | 96.30% |
| | noun error | 0.25 | 8 | 2.00 | | | 4 | 1.00 | | |
| | pronoun error | 1 | 35 | 35.00 | | | 67 | 67.00 | | |
| | non-finite verb error | 0.5 | 1 | 0.50 | | | 0 | 0.00 | | |
| | adjective/adverb error | 0.25 | 9 | 2.25 | | | 8 | 2.00 | | |
| R | locationization norm error | 0.5 | 3 | 1.50 | 36.75 | 9.82% | 13 | 6.50 | 8.25 | 99.60% |
| | punctuation error | 0.25 | 71 | 17.75 | | | 7 | 1.75 | | |
| | other errors | 0.25 | 2 | 0.50 | | | 0 | 0.00 | | |
| F+A+R | | | 1398 | | | | 854.00 | | | |
| FAR | | | 32.70% | | | | 58.88% | | | |

| | subcatgories | score | Baidu Translate | | | | Youdao Translate | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | number of errors | total scores | F/A/R | F/A/R (%) | number of errors | total scores | F/A/R | F/A/R (%) |
| F | mistransaltion | 2 | 179 | 358.00 | 611.00 | 70.58% | 187 | 374.00 | 628 | 69.79% |
| | terminology error | 2 | 68 | 136.00 | | | 45 | 90.00 | | |
| | idiom error | 2 | 52 | 104.00 | | | 63 | 126.00 | | |
| | abbreviation error | 2 | 0 | 0.00 | | | 0 | 0.00 | | |
| | additions | 1 | 5 | 5.00 | | | 15 | 15.00 | | |
| | omissions | 2 | 2 | 4.00 | | | 9 | 18.00 | | |
| | fragmented sentences | 0.5 | 6 | 3.00 | | | 9 | 4.50 | | |
| | run-on-sentence-style translation | 0.5 | 2 | 1.00 | | | 0 | 0.00 | | |
| A | verb error | 0.25 | 8 | 2.00 | 74.50 | 96.41% | 25 | 6.25 | 76 | 96.35% |
| | noun error | 0.25 | 0 | 0.00 | | | 6 | 1.50 | | |
| | pronoun error | 1 | 72 | 72.00 | | | 65 | 65.00 | | |
| | non-finite verb error | 0.5 | 0 | 0.00 | | | 1 | 0.50 | | |
| | adjective/adverb error | 0.25 | 2 | 0.50 | | | 10 | 2.50 | | |
| R | locationization norm error | 0.5 | 0 | 0.00 | 1.75 | 99.92% | 12 | 6.00 | 13 | 99.39% |
| | punctuation error | 0.25 | 0 | 0.00 | | | 26 | 6.50 | | |
| | other errors | 0.25 | 7 | 1.75 | | | 1 | 0.25 | | |
| F+A+R | | | 687.25 | | | | 716 | | | |
| FAR | | | 66.91% | | | | 65.53% | | | |

4.1 Functional Equivalence

The first dimension, "Functional Equivalence," in the "FAR Translation Quality Assessment Model 2.0" examines whether the core meaning of the source language is preserved in the translation. As shown in figure 7, the four MT tools exhibit varying degrees of error, signaling the need for improvements in translation accuracy. Upon deeper analysis, the most significant score deductions occurred under the "mistranslation" category, with LingoCloud incurring the most severe penalties. This section will explore the most frequent error categories within this dimension, with a focus on those most impactful to meaning transmission.

*4.1.1 Mistranslation.* Mistranslation is the most common error type across all four translation tools, inflicting the most serious influence on the accuracy of translation. The primary causes of mistranslation included improper sentence segmentation and word-for-word translation.

*Example 1:*

ST: 冠军奖根大黄瓜 (*guanjun jiang gen da huang gua*)

Context: This sentence occurs during an arm-wrestling contest between two soldiers, with the winner receiving a big cucumber as a reward.

LingoCloud MT: The winner gets a cucumber root.

Baidu MT: Champion Award Root Cucumber.

Youdao MT: The winner will be a cucumber.

DeepL: The winner gets a cucumber.

In this case, DeepL correctly conveys the intended meaning. However, both LingoCloud and Baidu mistranslate "a big cucumber" as "cucumber root" and "root cucumber," respectively, which leads to significant errors. While Youdao translates the cucumber accurately, the sentence structure leads to the mistaken implication that "the winner will be a cucumber."

*Example 2:*

ST: 娶个老婆当月亮看呢 (*qu ge laopo dang yueliang kan ne*)

Context: The protagonist is being teased by a colleague for delaying his trip home with

his leave having been granted for some time.

LingoCloud MT: Marry a wife and watch the moon

Baidu MT: Marry a wife and watch the moon

Youdao MT: Take a wife like the moon

DeepL: Marrying a wife is like looking at the moon.

In this example, both LingoCloud and Baidu render the phrase as a parallel structure between "marry" and "watch," resulting in an illogical translation that implies one must both marry a wife and watch the moon. This translation deviates from the original meaning in context. Although Youdao and DeepL capture the essence more accurately, they still fail to show the playful tone of the original sentence.

*Example 3:*

ST: 我水平不高 (*wo shuiping bugao*)

Context: This sentence is from a newly appointed political instructor during his introductory speech to the soldiers, where he refers to his level of leadership.

LingoCloud MT: I'm not very good.

Youdao MT: I'm not good at it.

Baidu MT: My level is not high.

DeepL: I'm not very good.

In this context, "水平" (*shuiping*) refers to the instructor's leadership competence. Both LingoCloud and DeepL interpret the phrase as "I'm not very good," which primarily conveys a judgment about personal character rather than leadership capability, thus misrepresenting the original meaning. Baidu translates the sentence literally as "My level is not high," which, while grammatically correct, does not appropriately capture the intended message. Youdao avoids this issue by using the phrase "good at it," which more accurately expresses the speaker's intention.

*4.1.2 Terminology Misuse.* As this study uses the subtitles of a military film as the source text, numerous military terms appear. The quality of their translation across the four machine translation tools varies significantly.

*Example 4:*

ST: 他为我们九连的指导员 (*ta wei women jiulian de zhidaoyuan*)

Context: This sentence introduces the role of a political instructor in the company.

LingoCloud MT: He was the instructor of our ninth company.

Youdao MT: He's our instructor in the ninth company.

Baidu MT: He is the instructor of our Ninth Company.

DeepL MT: He's the commander of our 9th Company.

The term "指导员" (*zhidaoyuan*) refers to a unique role in the Chinese military system, specifically responsible for political education. The appropriate English translation is "political instructor." LingoCloud, Youdao, and Baidu all translate the term as "instructor," failing to describe the actual role of that position, which can lead to confusion. DeepL mistranslates it as "commander," introducing an error in the rank and duties of the character.

*Example 5:*

ST: 你去的那个师的师长 (*ni qude nage shi de shizhang*)

Context: The speaker is discussing the commander of a military division.

LingoCloud MT: The teacher of the division you went to.

Youdao MT: The head of the division you went to.

Baidu MT: The teacher of the division you went to.

DeepL MT: The commander of the division you're going to.

The standard translation of "师长" (*shizhang*) is "division commander." Youdao's "the head of the division" and DeepL's "the commander of the division" are accurate translations, while both LingoCloud and Baidu incorrectly translate "师长" (*shizhang*) as "teacher," which distorts the meaning.

*4.1.3 Idiom and Colloquial Expression Misuse.* The dialogue in this film is predominantly colloquial, and it includes many Chinese idioms and cultural expressions, posing challenges for machine translation tools.

*Example 6:*

ST: 咱们牛郎该会织女了(*zanmen niulang gai hui zhinv le*)

Context: A character is teasing another about his reunion with his wife, referencing the mythological Cowherd and Weaver Girl, implying a long-awaited meeting between a couple.

LingoCloud MT: Our cowherd should know how to weave.

Youdao MT: We cowherd should meet the weaver girl.

Baidu MT: Our Cowherd should be able to Weaver Girl now.

DeepL MT: It's time for the Cowherd to meet the Weaver.

This expression refers to the Chinese myth of the Cowherd and the Weaver Girl, symbolizing a reunion between a man and his wife. LingoCloud translates it incorrectly, suggesting that the "cowherd" needs to know how to weave, which is a serious error. Youdao, Baidu, and DeepL better convey the metaphor, with DeepL providing a particularly fitting translation by using the phrase "It's time for…," matching the original tone and context well.

*Example 7:*

ST: 都是属猪的 (*doushi shuzhu de*)

Context: The speaker humorously refers to someone born in the Year of the Pig, based on the Chinese zodiac.

LingoCloud MT: They're all pigs.

Youdao MT: They're all pigs.

Baidu MT: All belong to the pig genus.

DeepL MT: We're both pigs.

In this instance, the characters are referring to someone born in the Year of the Pig, a part of the Chinese zodiac. LingoCloud, Youdao, and DeepL translate it literally as "pigs," which may confuse audiences unfamiliar with the cultural context, as it could imply an insult. Baidu attempts to clarify this by translating it as "belong to the pig genus," which also results in a significant mistranslation.

These examples illustrate that machine translation tools struggle with idiomatic and culturally specific expressions, resulting in frequent errors.

*4.1.4 Omission Errors.* In the "Omission" category, LingoCloud and DeepL committed the most errors, with 41 and 30 instances, respectively, while Baidu and Youdao performed better. Below is an example highlighting errors common across all four tools.

*Example 8:*

ST: 别听他说话难听(*bie ting ta shuohua nanting*)

Context: The character is commenting on another's blunt speech, explaining that although his words sound harsh, his intentions are good.

LingoCloud MT: Don't listen to him

Youdao MT: Don't listen to what he says

Baidu MT: Don't listen to him speak harshly

DeepL MT: Don't listen to him.

Here, the omitted phrase "难听" (*nanting*) means "harsh" or "unpleasant to hear." Both LingoCloud and DeepL translate this simply as "Don't listen to him," missing the implication that the character's speech is unpleasant but not ill-intentioned. Youdao provides a similar translation, while Baidu comes closer to the original meaning by including "harshly," though it introduces a grammatical issue.

This section provided examples from the four most frequent and serious error categories in the "Functional Equivalence" dimension. The main causes of these errors can be attributed to two factors: (1) insufficient training with culturally specific data, leading to misinterpretations of Chinese culture-laden terms, and (2) difficulty in handling subjectless sentences, causing literal translations that distort the original meaning.

## 4.2 Acceptability

Acceptability concerns whether the translation conforms to the linguistic norms of the target language. Among the four machine translation tools, the most frequent errors occurred in

the use of verb, pronoun, and adjective/adverb. Noun-related errors, as well as errors with non-finite verbs and infinitives, were rare. This section will primarily analyze the first three categories.

*4.2.1 Verb Errors.* Verb errors primarily involve issues related to tense, voice, and number. These errors are often due to the high density of cuts between segments in the Chinese subtitles and the lack of punctuation, making it difficult for the translation tools to accurately translate the meaning based on the context. The following examples illustrate this issue.

*Example 9:*

ST: 天天为你当牛做马 (*tiantian weini dangniu zuoma*)

Context: A character recalls a conversation where he promised his wife that after he was sent to a civilian position, he would do all household chores for her.

LingoCloud MT: I'll work for you every day.

Youdao MT: Every day for you to be a cow.

Baidu MT: Every day I will be a cow and a horse for you.

DeepL MT: I've been working my ass off for you every day.

In this context, the action described will occur in the future, after the character changes to a civilian position, meaning the verb should be in the future tense. LingoCloud, Youdao, and Baidu use the right tense. However, DeepL incorrectly uses the present perfect continuous tense, suggesting that the action has been ongoing, which contradicts the original context. This is likely due to the segmented nature of the subtitles, which caused the tool to overlook the timeline.

*Example 10:*

ST: 摊上个俊媳妇儿啊 (*tanshang ge jun xifu er*)

Context: A character jokingly remarks about another character's wife, implying that he is fortunate to have a beautiful wife.

LingoCloud MT: He had a beautiful wife

Youdao MT: You got a beautiful wife

Baidu MT: Spread out a handsome daughter-in-law

DeepL MT: He's got a handsome daughter-in-law.

The phrase "摊上" (*tanshang*) means "to have" in a colloquial sense. In this context, it should be rendered in the present perfect or simple present tense, as the character and his wife are in the second year of their marriage. LingoCloud and Youdao use the past tense, which is incorrect. Baidu's translation, "spread out," is an erroneous literal translation. DeepL correctly captures the meaning with the oral English "He's got" and accurately uses the present tense with the right number.

*4.2.2 Pronoun Errors.* Pronoun errors primarily occur when the translated pronoun does not correspond correctly to the referent in the ST. These mistakes often result from the fragmented nature of subtitles, making it difficult for the machine to track the context, and from a lack of understanding of dialect-specific expressions. The examples below highlight these issues.

*Example 11:*

ST: 让人家天天在家等你 (*rang renjia tiantian zaijia dengni*)

Context: The speaker criticizes the protagonist for leaving his wife waiting at home every day. In this context, "人家" (*renjia*, others) refers to the wife.

LingoCloud MT: Let them wait for you at home every day.

Youdao MT: Let people wait for you at home every day.

Baidu MT: Make them wait for you at home every day.

DeepL MT: Make her wait for you at home every day.

Here, "人家" (*renjia,* others) is a colloquial term meaning "she" in this context, referring to the wife. Only DeepL correctly translates this pronoun as "her." Both LingoCloud and Baidu use "them," which causes confusion by suggesting a plural subject. Youdao's use of "people" is vague and fails to convey the specific referent, making it unclear.

*Example 12:*

ST: 报效祖国，咱也愿留点儿血 (*baoxiao zuguo, zan ye yuan liudian er xie*)

Context: A soldier expresses his willingness to shed blood for his country. The word "咱" is used as a dialectal self-reference, meaning "I."

LingoCloud MT: Serve the motherland we are willing to leave some blood

Youdao MT: We'll shed some blood to serve our country

Baidu MT: To serve our country we are willing to leave some blood behind

DeepL MT: We're willing to spare some blood to serve our country.

In this case, "咱" (*zan*) refers to "I" in a regional dialect. However, all the tools translate it as "we," which misrepresents the speaker's personal statement. This shows that the tools struggle with understanding dialect and tend to default to a more general translation.

*4.2.3 Adjective/Adverb Errors.* Errors involving adjectives and adverbs arise mainly from incorrect use of comparative or superlative forms and poor word choice. Here are a few examples to illustrate such errors.

*Example 13:*

ST: 提前了五分钟 (*tiqian le wu fenzhong*)

Context: A character is explaining that the training session started five minutes earlier than planned.

LingoCloud MT: Five minutes early

Youdao MT: Five minutes early

Baidu MT: Five minutes ahead of schedule

DeepL MT: Five minutes early.

The correct translation should use "earlier" to reflect that the action is earlier compared to a set time. LingoCloud, Youdao, and DeepL translate it simply as "early," losing the comparative nuance. Baidu uses "ahead of schedule," which conveys the correct sense of timing.

*Example 14:*

ST: 老规矩啊，罚 (*laoguiju a, fa*)

Context: A character jokingly says they should follow the customary rules and make the other character do something funny in front of many colleagues.

LingoCloud MT: The old rules.

Youdao MT: The old rules. Punishment.

Baidu MT: Old rules punishment.

DeepL MT: The usual rules apply.

In this context, "老" (*lao*) refers to something habitual or customary, not literally "old." LingoCloud, Youdao, and Baidu mistakenly translate it as "old," meaning the rules that are obsolete. DeepL provides the correct version with "usual," effectively capturing the implied context.

This section focused on errors related to verbs, pronouns, and adjectives/adverbs. These mistakes are often due to the highly fragmented structure of Chinese subtitles and a large number of non-subject sentences. Without sufficient context, machine translation tools are unable to correctly combine meaning across segment boundaries.

## 4.3 Readability

Given that the accuracy rates of all four machine translation tools in handling "Localization Norm Errors," "Punctuation Errors," and "Other Errors" exceeded 98%, this section will focus on representative examples of these types of errors.

*4.3.1 Localization Norm Errors.* Such errors emerge when a translation fails to adapt to cultural, social and linguistic norms familiar to the target audience. Typical examples are as follows.

*Example 15:*

ST: 那个老娘们能给你什么消息？(*nage laoniangmen neng geini shenme xiaoxi?*)

Context: This statement is spoken with a slightly condescending tone, questioning the reliability of the information an old woman could provide.

LingoCloud MT: What can that Old Bitch Tell You?

Youdao MT: What news can that old lady give you?

Baidu MT: What news can that old lady give you?

DeepL MT: What kind of information could that old woman have given you?

In this case, “老娘们” (*laoniangmen*) conveys a mildly derogatory meaning in Chinese, but LingoCloud overtranslates it as “Old Bitch,” which is too harsh and inappropriate. Youdao, Baidu, and DeepL opt for the neutral “old lady” or “old woman,” which better reflects the original tone without being overly aggressive. DeepL’s use of “what kind of information” also introduces a more appropriate level of subtlety.

*Example 16:*

ST: 五百五十块 (*wubai wushi kuai*)

Context: This refers to an amount of money, with “块” being a colloquial term for “yuan.”

LingoCloud MT: Five hundred and fifty dollars.

Youdao MT: Five hundred and fifty dollars.

Baidu MT: 550 yuan.

DeepL MT: Five hundred and fifty dollars.

Here, “块” is the casual term for “yuan” in Chinese currency. Baidu accurately translates it as “yuan,” although this may confuse non-Chinese speakers unfamiliar with the currency. LingoCloud, Youdao, and DeepL localize it as “dollars,” but they fail to account for the fact that the amount is in yuan, not dollars, leading to a serious mistake in meaning transmission.

*4.3.2 Punctuation Errors.* LingoCloud and Youdao had the highest occurrence of errors in this category, primarily related to periods and question marks, while Baidu and DeepL performed more consistently, making fewer mistakes.

*Example 17:*

ST: 你不进城？(*ni bu jincheng*?)

Context: A character asks whether someone is planning to go into town.

LingoCloud MT: You’re not going to town.

Youdao MT: You don’t go to town.

Baidu MT: You don’t go to the city.

DeepL MT: You’re not going to the city?

The original sentence is a question, but LingoCloud, Youdao, and Baidu render it as a statement, losing the interrogative tone of the original. Only DeepL preserves the question mark and accurately reflects the intended inquiry. The failure to include the correct punctuation significantly alters the tone and disrupts the flow of the dialogue.

In this section, typical examples of "localization norm errors" and "punctuation errors" were provided. These issues mainly arise due to the lack of punctuation in the original Chinese subtitles, leading to misinterpretation by machine translation tools. Moreover, these tools still lack sufficient training in handling non-verbal elements in dialogue. As a result, machine translation tools struggle when faced with sentences that lack explicit punctuation, which negatively impacts the translation's overall quality.

## 5. Conclusion

This study applied quantitative analysis to a full set of Chinese film subtitles and compared the output from four popular machine translation tools. By integrating two prevailing evaluation frameworks for subtitle translation and refining them, this report proposed the "FAR Translation Quality Assessment Model 2.0" and analyzed typical examples to highlight the performance of the four machine translation tools.

The results indicate that Baidu achieved the highest overall translation quality. This is attributed to the significant amount of Chinese-English training data incorporated into Baidu's system, resulting in more accurate translations overall. On the other hand, LingoCloud exhibited the weakest performance, with frequent and severe translation errors, pointing to the need for substantial improvements.

In the "F" (Functional Equivalence) dimension, "Mistranslation" emerges as the most serious error category, with LingoCloud showing the highest error rate and Baidu the lowest. In the "A" (Acceptability) dimension, all four MT tools perform well, with error rates below 4%. In this dimension, pronoun errors are the most frequent, largely due to the prevalence of non-subject sentences, which are difficult for machine translation tools to interpret correctly. In the "R" (Readability) dimension, the main areas of weakness are "Localization Norm Errors"

and "Punctuation Errors," suggesting that future advancements in machine translation should focus on addressing these specific challenges.

Based on the observations and reflections, the development of machine translation, both domestically and internationally, should prioritize the following areas:

Expanding the machine translation corpus, especially those that include Chinese cultural terms. Improving subtitle segmentation to ensure smoother transitions between sentences for more complete grammatical structure and cohesive meaning. Furthermore, building specialized corpora for fields such as politics, engineering, and law will better improve accuracy.

Enhancing human-machine collaboration. Human editors should play a vital role in pre-editing subtitles before they are processed by machine translation tools, resolving potential translation issues beforehand. Additionally, post-editing is crucial to polish the machine-translated text by correcting grammar, format, and fluency issues.

Nevertheless, this study has certain limitations. For instance, human annotators may have varying understanding on definitions of error categories, which could introduce subjectivity into the evaluation process. Furthermore, the use of only one film as a corpus limits the generalizability of the findings. Future research should consider using a wider variety of subtitle materials and standardize annotation criteria for more comprehensive comparisons.

# References

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2019. "An Effective Approach to Unsupervised Machine Translation." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 194–203. doi:10.18653/v1/P19-1019.

Chyxx Company. 2023. "2023-2029 Nian Zhongguo Jiqi Fanyi Hangye Shichang Yunyin Taishi Ji Touzi Zhanlue Guihua Baogao." [2023-2029 China machine translation industry market operation situation and investment strategy planning report.] July 24. Accessed May 5, 2024. https://www.sohu.com/a/708793117_120950077.

Feng, Yang, and Chenze Shao. 2020. "Shenjing Jiqi Fanyi Qianyan Zongshu." [Frontiers in neural machine translation: A literature review]. *Journal of Chinese Information Processing* 34 (7): 1–18. http://jcip.cipsc.org.cn/CN/Y2020/V34/I7/1.

Garg, Sarthak, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. "Jointly Learning to Align and Translate with Transformer Models." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 4453–4462. doi:10.48550/arXiv.1909.02074.

Goyal, Naman, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. "The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation." doi:10.48550/arXiv.2106.03193.

Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, Hany Hassan Awadalla. 2023. "How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation." doi:10.48550/arXiv.2302.09210.

Huang, Youyi. 2022. *Cong Fanyi Shijie Dao Fanyi Zhongguo* [From translating the world to translating China]. Beijing: Foreign Languages Press.

Li, Fengqi. 2021. "Jiyu Shenjing Wangluo De Zaixian Jiqi Fanyi Xitong Yinghan Huyi Zhiliang Duibi Yanjiu." [A comparative study on the quality of English-Chinese translation of online machine translation systems based on neural networks.] *Shanghai Journal of Translators*, no. 4, 46–52. doi:10.3969/j.issn.1672-9358.2021.04.013.

Linfeng, Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. "Semantic Neural Machine Translation Using AMR." *Transactions of the Association for*

*Computational Linguistics*, no. 7, 19–31. doi:10.1162/tacl_a_00252.

Liu, Jichao, Shisheng Lü, and Chengpan Liu. 2024. "Shenjing Wangluo Jiqi Fanyi De Renzhi Yinyu Jiegou." [Deconstructing cognitive metaphors in neural network machine translation.] *Translation Research and Teaching*, no 1, 82–88. https://d.wanfangdata.com.cn/periodical/fyyjyjx202401013.

Lommel, Arle Richard, Aljoscha Burchardt, and Hans Uszkoreit. 2013. "Multidimensional Quality Metrics: A Flexible System for Assessing Translation Quality." In *Proceedings of Translating and the Computer 35*. London: Aslib. https://aclanthology.org/2013.tc-1.6.pdf.

Mathur, Nitika, Timothy Baldwin, and Trevor Cohn. 2020. "Tangled Up In BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4984–4997. doi:10.18653/v1/2020.acl-main.448.

Pedersen, Jan. 2017. "The FAR Model: Assessing Quality in Interlingual Subtitling." *Journal of Specialised Translation*, no. 28. https://jostrans.soap2.ch/issue28/art_pedersen.php.

Popel, Martin, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. "Transforming Machine Translation: A Deep Learning System Reaches News Translation Quality Comparable to Human Professionals." *Nature Communications*, no. 11. doi:10.1038/s41467-020-18073-9.

Stahlberg, Felix. 2020. "Neural Machine Translation: A Review." *Journal of Artificial Intelligence Research*, no. 69, 343–418. doi:10.1613/jair.1.12007.

Wang, Jinquan, and Bojia He. 2024. "Jiyu Shenjing Wangluo Moxing De Fanyi Yuyi Zhiliang Lianghua Pingjia." [Quantified assessment of translation semantic quality based on neural network model.] *Chinese Foreign Languages* 21 (1): 92–101. https://benjamins.com/online/etsb/publications/58188.

Wen, Xu, and Yaling Tian. 2024. "ChatGPT Yingyong Yu Zhongguo Tese Huayu Fanyi De Youxiaoxing Yanjiu." [The effectiveness of ChatGPT in translating China-specific discourse text.] *Shanghai Journal of Translators* 39 (2): 27–34+94–95. doi:10.3969/j.issn.1672-9358.2024.02.005.

Xiao, Weiqing. 2017. *Yinghan Yingshi Fanyi Shiyong Jiaocheng* [A practical guide to English-Chinese audiovisual translation]. Shanghai: East China University of Science and Technology Press.

Xu, Wei. 2024. "Jiyu Shendu Xuexi De Nongji Jiqi Yingyu Yuliaoku De Sheji." [Design of an English corpus for agricultural machines based on deep learning.] *Journal of Agricultural Mechanization Research* 46 (10): 208–212. doi:10.3969/j.issn.1003-188X.2024.10.035.

Zhang, Biao, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. "Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1628–1639. doi:10.18653/v1/2020.acl-main.148.

Zhang, Wenbo, Xinlu Zhang, and Yating Yang. 2021. "Mianxiang Diziyuan Shenjing Jiqi Fanyi De Huiyi Fangfa." [Back translation for low resources neural machine translation.] *Journal of Xiamen University (Natural Science)* 60 (4): 675–679. doi:10.6043/j.issn.0438-0479.202011025.

Zhou, Xinghua, and Chuanying Wang. 2020. "Rengongzhineng Jishu Zai Jisuanji Fuzhu Fanyi Ruanjian Zhong De Yingyong Yu Pingjia." [Application and evaluation of artificial intelligence technology in computer-aided translation software.] *Chinese Translators Journal* 41 (5): 121–129. https://qikan.cqvip.com/Qikan/Article/Detail?id=7102938325.

**Filmography**

*Wreaths at the Foot of the Mountain*. 1984. Directed by Xie Jin. Shanghai Film Studios. War picture.