Human Decision-Making and Machine Assistance in Subtitle Translation Through the Lens of Viewer Experience

Sijin Xian(✉)

Translaxian LLC, Silver Spring, MD 20906, USA
projects@xiansijin.com

Abstract. Subtitle translation, which entails a unique blend of language skills and technical know-how, has benefited from the productivity boost enabled by various forms of technology, such as machine translation, automatic reading speed calculation, waveform display, and shot change detection. While these tools provide helpful assistance, the success of audiovisual translation projects is predicated upon the human linguist that interacts with the subtitling interface. By documenting the working and decision-making process of the author, an experienced English-to-Chinese subtitler and quality controller (QCer) of streaming and entertainment content, this paper explores what the machine can and cannot yet do for us in audiovisual translation, highlights the importance of empathy and judgment in creating top-notch viewer experience, and informs potential directions for future machine development in aid of human expertise.

Keywords: Audiovisual Translation · Viewer Experience · Technology Humanization · Multimodality · User Centricity · User Experience Design

1 Introduction

Thanks to technological advancements, entertainment and streaming content such as movies, drama series, unscripted reality shows, and documentaries can be localized and subtitled with accelerated productivity. The efficiency boost has been made possible in two ways. The first is the integration of the various components of audiovisual translation workflow into one tool or interface. Long gone were the days when "subtitlers needed a desktop computer, an external video player in which to play the VHS tapes with the material to be translated, and a television monitor to watch the audiovisual productions" [1]. The second is a suite of machine assistance capabilities that remove unnecessary labor and tedium on the subtitler's part. Even with heavyhanded edits, working with machine-translated output saves time and keystrokes compared to typing every word from scratch. Automatic reading speed calculation alerts the linguist when the subtitle exceeds the suggested length for the display time without requiring manual calculation. Waveform display enables precise identification of audio timeframes, avoiding reaction delay from auditory perception to manual marking. Shot change detection removes the need to examine a video frame by frame.

H. Degen and S. Ntoa (Eds.): HCII 2023, LNAI 14051, pp. 192–204, 2023.
https://doi.org/10.1007/978-3-031-35894-4_14

Against the backdrop of technological efficiency, how are human subtitlers interacting with the machine? Where are we in terms of building a professional translation world where the machine does what it does best while human expertise shines in the arena of its incomparable mastery? What potential new use cases could we explore for future technological developments? As a professional English-Chinese translator who has worked in the streaming content localization field since 2018 both as a subtitle translator and quality controller (QCer), the author investigates these questions by documenting her reflections on and observations of the thoughtful and deliberate considerations needed to accommodate audio, visual, and storytelling elements. By presenting the pitfalls and best practices through the lens of viewer experience, the author highlights the beautiful minuet between human decision-making and machine assistance.

The viewer experience angle is meant to provide a unifying focal point for the intricate balance a subtitle translator needs to strike when attempting to accommodate reading speed, shot changes, and linguistic particularities while preserving readability and honoring creative intent. This perspective is also in line with the focus on cognitive and empirical research identified by Chaume [2] as one of the four methodological turns in the discipline of audiovisual translation, where "the interest is geared not only towards the translator's mental processes…but also, and mainly, on the audience's response to audiovisual translation."

Furthermore, this perspective is optimal for highlighting the most valuable asset of human expertise in audiovisual translation: empathy and judgment. Empathy puts the translator in the shoes—or in the context of audiovisual translation, seats—of the viewers they serve. This awareness requires the translator, who is bilingual and bicultural themselves, to be actively cognizant of what it is like to navigate an audiovisual asset as someone unfamiliar with the source language and culture. Judgment is especially salient when different priorities are in conflict, and an informed choice needs to be made to sort out the least of multiple evils. These two essential skills are the central thread that runs through the human expertise that we will explore later.

This paper is organized by the four cornerstones of viewer experience: accurate translation, scriptwriting mindset, effortless reading, and equivalent experience. These are derived from the four sets of unique characteristics of audiovisual translation compared to traditional text-based translation: the heavy presence of idiomatic expressions, slang terms, and cultural references; the essential components of plot, lines, and characters; the limited display time and continuous play of subtitles; and the artistry of direction, audience reaction, and creative intent. Each section covers the role of machine assistance in terms of benefits and challenges, best practices to achieve optimal viewer experience, and potential directions for future machine advancements.

2 First Cornerstone: Accurate Translation

Accurate translation is the first cornerstone of the audiovisual viewer experience. All other considerations are reduced to pointlessness when plot lines and intended messages are skewed to the point of impeding authentic storytelling. The adoption of machine translation in audiovisual translation has been slower compared to traditional text-based translation projects [3], and it was fairly recent when one of the author's clients incorporated pre-translated subtitles into the workflow. It needs to be highlighted that the use of

machine translation in the professional world—contrary to those who turn to the likes of Google Translate to sidestep hiring a trained linguist—comes with the awareness that it is a productivity enhancement tool with inaccurate output and that it is the human expert's job to catch and fix the errors.

Achieving accuracy requires impeccable execution on three fronts: processing the source text, internalizing the message, and producing the end rendition**.** This process faces two notable challenges in audiovisual translation. First, the script or audio transcription of audiovisual content is only one-third of the "source text." Translating the source text alone in isolation from the audiovisual elements is a doomed attempt. Second, the abundance of idiomatic expressions, slang terms, and cultural references puts exceedingly high demands on the linguist's source language mastery to understand and convey the essence of the message.

2.1 Three-Dimensional Context

In translation, we say context is king. In audiovisual translation, we have three: the audio elements (such as tone, tempo, delivery of the speech, and background audio cues), the visual elements (such as on-screen information for storylines, facial expressions, and gestures), and the transcribed source text. For example, the same line of text, "I should report this to the boss," may need to be rendered differently based on the actor's delivery. If "I" is the focal point of this statement, then what should be conveyed is "I should be the one who reports this to the boss," not "I have the obligation to report this to the boss."

Similarly, the source text's translation is also affected by the on-screen visual cues. In one QC instance, two sequential lines both read "Hello?" in English. The character says the first "Hello?" to check if the person on the phone can still hear her, and the linguist assumes the second "Hello?" to be a repeated follow-up query. However, the latter utterance takes place a few seconds later, when the character gets up to scan the house upon hearing a strange sound. In Chinese, these should be rendered differently, and the oversight is a telltale sign that the linguist did not watch the movie closely.

The two examples above show the contextual variations within standard language usage: emphasis matters; a word has more than one meaning. Apart from these scenarios, there are cases where the standard dictionary definition and everyday usage differ. In another movie the author QCed, three women are being chased by some thugs, and one of the women makes it into a safe room first. The two other women arrive later, knocking on the door and requesting to be let in. The woman in the room asks, out of caution, "Are you alone?" to ensure the dangerous men are not around. The machine translation treated "alone" as one solitary person. However, from the context, we know the speaker is asking, "Are there just the two of you?" Again, the rendition would be different in Chinese, and this contextually jarring output was not corrected by the linguist. A similar case would be the phrase "a couple of." While its standard meaning is two, it is used in informal contexts to mean two or three. Simply having the source text available is not enough to produce an accurate translation.

As highlighted in the above examples, machine translation, besides its usual issues and challenges [4], encounters the critical issue that source text is not the whole context in audiovisual translation. Furthermore, the author's QC experience has shown that when the machine translation output is grammatically correct yet contextually unfitting, it is

not always detected by linguists if they are not watching the video with a keen eye and taking in the three-dimensional context themselves. In other words, audiovisual translators must not think of their job as simply checking the source text against the machine output. In future technological developments, while linguists are trained on the importance of grasping the entirety of the context, machines might be trained on the fronts of integrated audio, visual, and text-level comprehension. There could also be a flagging mechanism for contextual or usage uncertainty to alert linguists that additional confirmation is needed.

2.2 Tricky Translations

Machine translation tends to run into trouble regarding idioms [5], slang terms [6], and cultural references [7], and human expertise is essential to correcting these errors. In this realm of human-machine interaction, three issues are particularly salient in the author's QC experience: mistranslations hiding in plain sight, hidden messages uncaptured, and unidiomatic or unnatural renditions. The culprits behind these issues are usually a lack of mindful engagement with the context, inadequate language mastery causing difficulty understanding nuances, and getting swayed by source text phrasing. Below are some examples for each category.

Glaring errors can escape detection when the translator goes on autopilot. Knee-jerk reaction as such happens when the linguist has a go-to rendition or understanding of a phrase or expression when they first learned the source language and never paused to ponder other meanings. As a result, when the machine translation output is also based on the dominant interpretation, the mistake ends up flying under the radar. For example, the primary usage of "I don't know," meaning "I do not have the knowledge or information," can be so ingrained in translators that they fail to entertain the possibility of it having a second usage. However, frequently in everyday speech, "I don't know" is used to signal uncertainty, hesitation, or disagreement.

Similarly, when someone says "I don't blame you" in response to the other person's rant, this phrase has nothing to do with assigning blame and is instead an expression of sympathy and understanding. When the machine translates "I don't blame you" word for word, the linguist must have the contextual awareness and sensitivity to uncover the right message. In other words, translators need an inner alarm system for jarring renditions that prompts them to perform due diligence.

Difficulty understanding linguistic or cultural nuances leaves hidden messages uncaptured. In a notable QC instance, a translator rendered the expression "playing for the other team" literally in the sense of sports team allegiance—which is also something machine translation would do—instead of the intended reference to one's sexuality. The author has also seen the expression "good morning to you, too" rendered as a morning greeting instead of a sarcastic tease implying the other person did not greet them properly. Humor and sarcasm are typical tripwires in audiovisual translation and require an intimate understanding of the source language and culture, which further shows that in a workflow that intends to utilize machine translation to save time and increase productivity, higher priority must be placed on cultivating a talent pool that supplies highly qualified translators to make up the machine's shortcomings.

The third issue is unnatural translation. Being bilingual means constantly picking up new vocabulary, expressions, and linguistic nuances from both directions. This could lead to mixing up languages, forgetting words, or losing some sensibilities of what sounds natural and idiomatic. One of the challenges in translation is being able to think as if one only knew one language so that their rendition is not "polluted" by the patterns and habits of the other. Stiff, awkward, unidiomatic translations happen when the translator—and the machine—performs word-for-word conversion instead of holistic and artful translation, which necessitates an abstract extraction of the message, tone, and intent before conveying them naturally into another language.

The future of audiovisual translation is a "teamwork makes the dream work" scenario for human-machine interaction. In order to produce the sophisticated and creative localization critical to a delightful viewing experience, investment in translators of the highest caliber must be in tandem with machine-enabled productivity boost to keep tricky errors at bay. Future machine training might consider a targeted effort on collecting quality translation memories for frequently occurring slang terms, idiomatic expressions, and cultural references and provide multiple options for translators to select from based on the context.

3 Second Cornerstone: Scriptwriting Mindset

Though different from dubbing, which places a heavy emphasis on matching the mouth shapes and movements of the on-screen actors, a subtitle is nevertheless viewed as the replacement of the spoken line heard in the original language and requires a high level of character embodiment. To this end, the translated subtitles must match the audible utterances while conforming to the world-building and plot lines. Therefore, it is essential for subtitlers to have a scriptwriting mindset and understand that the translations are lines for the actors to say aloud as if they were acting in another language. This section focuses on the importance of creative recasting and subtitle flow in bringing subtitles to life.

3.1 Creative Recasting

Recasting is a natural process in translation, as different languages have different grammatical and syntactical construction rules. In subtitle translation, however, a different recasting is required to match the audio with the on-screen context and avoid potential spoilers. For instance, to translate "I'll let you play with your friends if you do your homework first" into Chinese, the order is switched to "if you do your homework first, I'll let you play with your friends," as it is customary to put the if-clause at the beginning of a sentence.

Let us imagine, however, this is a scene from a movie, where the child looks excited upon hearing the first half of the sentence at the prospect of a fun time with friends. Then, when the mother raises the condition of doing homework, he shows a slightly disappointed look. In this case, as demonstrated in the visual demonstration below, the usual recasting would fail as a subtitle because the translation delivers the information in a different order that's jarring for the storyline.

Conventional text translation strategy:
English: I'll let you play with your friends if you do your homework first.
Chinese: 如果你先把作业做了，我就让你跟小伙伴玩

Movie translation (incorrect):
If you do your homework first, [kid looks excited]
[Mom continues] I'll let you play with your friends. [kid looks disappointed]

A dilemma like this calls for a different translation approach: we need to rethink the source text so that we do not need to move the segments around in translation. In this case, a good solution is to circumvent the if-clause issue so that our translation reads, "I'll let you play with your friends, but you need to do your homework first." This way, the viewers can receive the information in the proper, harmonious order.

Movie translation (correct):
Subtitle 1: 你可以去跟小伙伴玩
Subtitle 2: 但是你要先把作业做了
I'll let you play with your friends, [kid looks excited]
[Mom continues] but you need to do your homework first. [kid looks disappointed]

Here is another example. A sentence such as "I want to live in a world where the grass is green, the roses are red, and the sky is blue with you" would be rendered into Chinese as "I want to with you live in a green-grass, red-rose, and blue-sky world." If this is a line from a movie, where images of green grass, red roses, and blue sky are shown on screen as the character speaks, and before "with you" is uttered, there is a deliberate pause before "with you," it intends to deliver a crescendo of romantic sentiments. However, the typical recasting would create a chaotic viewing experience where the subtitle does not match the on-screen sequence, and when the pivotal "with you" is spoken, the subtitle would show "world."

Conventional text translation strategy:
English: I want to live in a world where the grass is green, the roses are red, and the sky is blue with you.
Chinese: 我想和你生活在一个绿草如茵、玫瑰红艳、天空蔚蓝的世界里

Movie translation (incorrect):

I want to with you live in a [showing green grass] green-grass [showing red roses], red-rose [showing blue sky], blue-sky [impactful pause, eye contact] world.

Thus, human expertise is required to rewrite the sentence so the translation would read, “In my ideal world, there is green grass, red roses, blue sky, and you,” which unfolds the same way as the original.

Movie translation (correct):
Subtitle 1: 在我梦想的世界里
Subtitle 2: 有绿绿的青草
Subtitle 3: 红艳的玫瑰
Subtitle 4: 蔚蓝的天空
Subtitle 5: 还有你
In my ideal word, [showing green grass] there is green grass, [showing red roses] red roses, [showing blue sky] blue sky, [impactful pause, eye contact] and you.

The previous segment on the three-dimensional context of audiovisual translation demonstrates the challenge for the machine to produce accurate translations due to the lack of contextual input from the audio and visual elements. The above examples add another layer of difficulty by cautioning against switching subtitle orders arbitrarily, a common and innocuous practice for text-based translation that might create confusion in an audiovisual viewing experience. The deep understanding of the context and a high level of sensitivity to the nuances of language can be hard for machines to replicate.

3.2 Subtitle Flow

Another unique characteristic of subtitle translation is that a whole sentence can be broken into multiple subtitle instances due to audio’s delivery tempo (as seen in the example above) or shot changes. Different languages have different considerations for line breaking and how a sentence flows from one subtitle box to another. Focusing only on segment-to-segment equivalent horizontally (from source to target) cause incoherence in the vertical direction, which is how the content flows from the viewer’s perspective.

For example, in the Netflix show *Wednesday* (Season 1, Episode 6: “Quid Pro Woe,” 40:50), the original English line reads, “Now you know what’s at stake. Everything you vowed to protect, no less.” As this utterance is broken into two subtitles to accommodate the shot change, the Chinese subtitle translation reads:

Subtitle 1: 你知道这其中的利害关系
Subtitle 2: 你发誓要保护的一切

Subtitle 1 English back translation: You understand the situation of vital interests.
Subtitle 2 English back translation: Everything you vowed to protect.

Subtitle 1 English original: Now you know what’s at stake.
Subtitle 2 English original: Everything you vowed to protect, no less.

Subtitle 2 English original: Everything you vowed to protect, no less.

As seen above, the Chinese translation lacks coherence because "everything" is not a "situation," whereas in the English original, "everything" is precisely "what" is at stake. While the author cannot ascertain whether this incoherence is caused by machine error or human oversight, putting the whole sentence in versus putting in two segments does make a different in machine translation, too. Using DeepL as an example:

Full input: You understand what's at stake: everything you vowed to protect.
DeepL: 你明白什么是危险的：你发誓要保护的一切。
Back translation: You understand what's dangerous: everything you vowed to protect.

Input 1: You understand what's at stake.
DeepL: 你明白这其中的利害关系。
Back translation 1: You understand the situation of vital interests in this.

Input 2: Everything you vowed to protect.
DeepL: 你发誓要保护的一切。
Back translation 2: Everything you vowed to protect.

While the translation "what's dangerous," as opposed to "what's in danger," is not the most accurate, the example shows that segmentation affects machine translation output. Similarly, suppose the original English line is "I am embarrassed…for you." The machine can competently translate "I am embarrassed for you," but in this particular instance, the sentence needs to be broken into two parts to deliver the desired effect. Because the syntax would be different in Chinese, it needs human intervention to render it properly. The same applies to "I think the door is unlocked" versus "The door is unlocked…I think." The machine can render the first utterance correctly, but when "The door is unlocked" and "I think" are treated as two lines, the output will not establish the ideal subtitle flow.

Therefore, for better output, the machine needs to learn to move beyond segments to ensure coherence. It must be highlighted that this kind of segmented tunnel vision is not unique to machine translation. In the author's QC experience, linguists also make the mistake of focusing on the horizontal translation, seeing if their translation on the right matches the source text on the left, not paying attention to if the translation flows vertically from one subtitle event to the next. It is therefore advisable to "proofwatch"—as opposed to proofread—one's subtitle work before submission so that one can see if the translated lines match the audiovisual elements and flow naturally. It might be helpful for quality control purposes only to enable project submission after the finished work has been watched from beginning to end with the subtitles.

4 Third Cornerstone: Effortless Reading

Subpar subtitle translation can not only lack accuracy or scriptwriting flair but also be exhausting to read. One study shows even the fact that whether subtitles are syntactically or non-syntactically segmented can affect the viewer's cognitive load [8]. To deliver a first-rate viewing experience, one must ensure concise subtitle translation. When the writing is efficient, clear, and devoid of unnecessary distractions, the viewer can have a better experience. Another factor that can add to the viewing effort is the movement of on-screen visuals accompanied by the subtitles. In this regard, shot change is an essential concept in the audiovisual field. This part will discuss concise writing and attention to shot changes as strategies for effortless reading.

4.1 Concise Writing

Every line of subtitles, whose display time is dictated by the audio length, comes with an expiration date in a matter of seconds, making concision a key element in audiovisual translation to reduce cognitive load. To this end, reading speed standards are established to set the parameters of how many characters should be in a subtitle, given the time constraint. Machine assistance comes in handy by automatically alerting linguists when the subtitle is too long for preset limits without the need for manual counting. However, reading speed should only be a preliminary baseline or reference, not a gold standard as a cut-off point.

The first reason is that languages expand or contract in relation to each other. Szarkowska and Geber-Morón [9] found that"faster subtitles with unreduced text were preferred in the case of English videos, and slower subtitles with text edited down in Hungarian videos." For English-into-Chinese translation, the text typically shrinks in character space, which is why many redundant and wordy Chinese translations that are overwhelming or unpleasant to read never trigger machine warnings. For example, in the Netflix show *Love Is Blind* (Season 2, Episode 14: "After the Altar: The Future Looks Bright," 19:11), a mother-in-law says to a young couple, "We said 'better or worse.' This is worse." This was intended to encourage them to support each other as they were going through a rough patch. The Chinese translation doesn't seem significantly longer than the English original in terms of character space:

English: We said "better or worse." This is worse.
Chinese: 说好无论好坏都要互相扶持 这就是不好的时候

However, upon scrutiny, the Chinese translation was a result of explanatory expansion because the translator unpacked the idea of this pithy utterance into: "You said you would support each other for better or worse, and this is one of the worse times." This is a fair translation that conveys the meaning without violating any reading speed limit. However, as the aforementioned three-dimensional nature of audiovisual translation suggests, the subtitle does not exist in a vacuum. When the viewer watches the show, it is the balance of what they are reading, watching, and listening that brings viewing satisfaction, yet

there is a perceptible gap in syllable density (eight for English versus 20 for Chinese) in this example, which creates disharmony.

An advisable strategy would be to rephrase to convey the intent and circumvent the inevitable expansion created by following the original wording. For example, this could be rendered as, "既然结了婚，就要共渡难关"(Since this is a marriage, you should get through this hard time together). The new rendition clearly states the intended message and cuts down the syllable count almost by half, creating a more unified reading experience. This demonstrates that when the target language gets more wiggle room for the characters-per-second limitation, it is all the more important for the translators not to think of the machine as the ultimate overlord to please—the final yardstick should always be the viewer experience. Linguists need to understand that passing a mechanical reading speed check does not mean the subtitle is good to go and that there should always be a conscious effort to tighten up and optimize the writing.

Moreover, a subtitle can be short and strenuous to read all at once if it is syntactically complex or contains unfamiliar phrasing. In languages like Chinese, where sentences are punctuated, but words are not spaced in between, characters can stick together and make a sentence difficult to parse upon first glance, necessitating rephrasing or more reading time. When the machine registers the reading speed, it is a simple calculation that relies only on the character amount and display time. Given the chasm between an objective measurement versus a subjective reading experience, the linguist must watch the audiovisual asset with their subtitle translation to gauge the reading experience instead of relying on a number.

Therefore, efficient writing and sound judgment are essential to concise subtitling. In future technological developments, it can be worthwhile to consider more sophisticated metrics for reading speed calculation so that it is not simply based on cold, hard numbers such as characters per second. Instead, it should incorporate nuanced and complex considerations of syntax and readability analysis from a viewer's perspective.

4.2 Shot Changes

An audiovisual file is a combination of various continuous shots spliced together. The "stitch," or the transition point, is a shot change. An eye movement tracking study found that "participants had significantly more gaze shifts between subtitles and image in the case of subtitles displayed over shot changes compared to those which did not cross any cuts" [10]. While it cannot always be achieved—an utterance can cross multiple shot changes quickly and could not be broken into mini segments—it is good practice to contain a subtitle event within a shot change so that the viewer's gaze is not disrupted.

In this regard, automatic shot change detection is a helpful tool. However, false shot changes can sometimes be detected, such as when there is a sudden shift of lighting in a continuous shot, such as in a scene at a dance club with flashing disco balls. In these cases, human discernment is necessary to override machine detection.

During QC, the author frequently encounters instances where the machine indicates where the shot change is, but the linguist fails to split the subtitle event accordingly, even when it is grammatically and syntactically appropriate. For this issue, it would be helpful for future interfaces to detect shot changes and automatically create subtitle boxes around them upon determining that the spoken audio fits in the subtitle box as

a standalone unit grammatically. For example, consider a scenario where a woman is speaking on the stage, saying, "With advanced technology in renewable energy, we can create a better and more beautiful world," with a shot change to the audience's reaction during this utterance. It will be desirable to for the machine to detect whether the shot change falls on a comma (or a period in other cases)—namely a proper spot to split the subtitles—and automatically generate two subtitle boxes around the shot change.

5 Fourth Cornerstone: Equivalent Experience

In subtitle translation, we should strive not to subject our audience to a second-rate experience because they cannot understand the original language. Thus, it is essential to design subtitles in a way that provides an equivalent experience to what a native-speaking viewer would have without the subtitles as much as possible. This includes ensuring that subtitles reflect the rhythm, pace, and emotion of the spoken dialogue. It is also important to pay attention to details such as audio length, syllable density (as demonstrated in the "for better or worse" example), and parallelism or repetition in the dialogue. This section focuses on two aspects of equivalent experience. First, strategic subtitle segmentation ensures the viewers of our subtitle translation can watch the story as it unfolds rather than being spoiled significantly ahead of time. Second, an audiovisual product is the combination and culmination of the creative intent of all creators involved, which must be honored by the subtitle translator.

5.1 Subtitle Segmentation

If the audio utterance is water flowing down a faucet, then a subtitle box is a container to hold the water. In the author's QC experience, some subtitlers tend to focus on how much water one box can hold, only creating another subtitle event when the space runs out. This mindset is inconducive to creating an equivalent experience. Imagine a general giving a powerful speech, "Comrades, we must fight back with all our might." While delivering, he takes significant pauses between phrases. Putting out the whole sentence spoils the line for the audience because the emotion of the viewers is supposed to rise as the speech unfolds. Instead of seeing this as one sentence that can be fit into one subtitle box, the right approach is to watch the flow of the water.

Comrades, [long pause] we must fight back, [long pause] with all our might!

Subtitle 1: Comrades,

Subtitle 2: we must fight back,

Subtitle 3: with all our might!

Similarly, suppose someone is looking for Jason while shouting out his name. Then upon opening the door, something shocking is revealed, and the character says, "What are you doing?" The effect will be very different if the whole line is in one subtitle as "Jason! What are you doing?" One aspect of segmentation is facilitated by audio waveform detection, as we can clearly see how the water is flowing. For future technological developments, options can be explored about detecting dialogue audio and conducting syntax analysis so that the subtitle boxes can be automatically created according to the waveform.

Another consideration to time subtitle events even better is rethinking the way it is displayed. In the Chinese stand-up comedy competition show *Rock & Roast* (脱口秀大会),, a Karaoke-like subtitle display is used to avoid showing the punchline before the joke is completed. This innovative practice addresses a long-standing issue in subtitling, where one invariably reads ahead of the audio. Furthermore, imagine a movie scene where someone is speaking before an unexpected explosion or accident. This element of surprise, to the annoyance of the viewer, is usually foreshadowed by a dash in the subtitle. The ability to synch up a key and revealing subtitle with the audio is worth exploring as a new addition to the subtitling toolkit.

5.2 Creative Intent

From scriptwriting to direction and from camerawork to acting, filmmakers are intentional with their creative decisions. The translator needs to have an appreciative eye for these thoughtful intricacies and convey them to the viewers through deliberate decision-making. For example, if the character has a catchphrase, it needs to be rendered consistently for it to be memorable. If something is intentionally cryptic, the translation also needs to be vague. Timing decisions also need to be made by observing the camerawork, such as zoom-ins. Recreating humor is another aspect where human creativity shines, as the translator needs to reinvent wordplay while staying relevant to the plot line. Ultimately, human perception and intentionality are indispensable to successful subtitling, and modern subtitling tools must be placed in capable hands.

6 Conclusion

Audiovisual translation makes an intriguing arena to observe human-machine interaction because it combines the creative aspect of translation with the use of technological tools. This paper draws from the author's experience as a subtitler and QCer and discusses common issues and best practices through the lens of the four cornerstones of a pleasurable viewer experience. It documents the author's thought process and interaction with the subtitling interface in order to achieve accurate translation, scriptwriting mindset, effortless reading, and equivalent experience.

There are a lot of potential areas future technical advancements can explore in audiovisual translation. Continued research in these areas will lead to a more enjoyable subtitling experience for the audience. One possibility is to train the machine to understand both the source text and the situational context for the utterance. Machine translation can also be improved to take into account the prevalence of idiomatic expressions, slang terms, and cultural references in streaming and entertainment content. There is also potential in advanced rhetorical relationship and syntax analysis capabilities to ensure better subtitle segmentation and flow. In addition, there could be more advanced metrics than reading speed alone, such as readability scores and cognitive load assessment. Another innovative approach could be synching the punchline or key revelation with the audio, so that the subtitles are seamlessly integrated into the audiovisual experience.

Ultimately, the priority of audiovisual translation is to serve the story, and we need linguists who have the eye to appreciate creative intent and storytelling to craft a beautiful experience. The more advanced the subtitling machine becomes, the more essential it is to put the machine in the hands of top-notch linguists who can skillfully utilize machine assistance while practicing audience empathy and strategic judgment.

References

1. Díaz-Cintas, J., Massidda, S.: Technological advances in audiovisual translation. In: O'hagan, M. (ed.) The Routledge Handbook of Translation and Technology, pp. 255–270. Roudge, London (2019)
2. Chaume, F.: An overview of audiovisual translation: four methodological turns in a mature discipline. J. Audiovis. Transl. **1**(1), 40–63 (2018). https://doi.org/10.47476/jat.v1i1.43
3. Bywood, L., Georgakopoulou, P., Etchegoyhen, T.: Embracing the threat: Machine translation as a solution for subtitling. Perspectives **25**(3), 492–508 (2017). https://doi.org/10.1080/0907676X.2017.1291695
4. Okpor, M.: Machine translation approaches: Issues and challenges. Int. J. Comput. Sci. J. **11**(5), 2 (2014)
5. Shao, Y., Sennrich, R., Webber, B., Fancellu, F.: Evaluating machine translation performance on Chinese idioms with a blacklist method. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, pp. 31–38. European Language Resources Association, Miyazaki (2018)
6. Lin, B., Xu, F., Zhu, K., Hwang, S.: Mining cross-cultural differences and similarities in social media. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 709–719. Association for Computational Linguistics, Melbourne (2018). https://doi.org/10.18653/v1/P18-1066
7. Tekwa, K., Jiexiu, J.L.: Neural machine translation systems and Chinese *wuxia* movies: moving into uncharted territory. In: Jiao, D., Li, D., Meng, L., Peng, Y. (eds.) Understanding and Translating Chinese Martial Arts, pp. 71–89. New Frontiers in Translation Studies. Springer, Singapore (2023). https://doi.org/10.1007/978-981-19-8425-9_5
8. Gerber-Morón, O., Szarkowska, A., Woll, B.: The impact of text segmentation on subtitle reading. J. Eye Movem. Res. **11**(4), 2 (2018). https://doi.org/10.16910/11.4.2
9. Szarkowska, A., Geber-Morón, O.: Viewers can keep up with fast subtitles: Evidence from eye movements. PLoS ONE **13**(6), e0199331 (2018). https://doi.org/10.1371/journal.pone.0199331
10. Krejtz, I., Szarkowska, A., Krejtz, K.: The effects of shot changes on eye movements in subtitling. J. Eye Movem. Res. **6**(5), 3, 1–12 (2013). https://doi.org/10.16910/jemr.6.5.3