Using AI Tools in Multimedia Localization Workflows: a Productivity Evaluation

Ashley Mondello[1], Romina Cini[1], Sahil Rasane[1], Alina Karakanta[2], Laura Casanellas[3],
[1]Language Scientific, Boston, MA, USA
[2]Leiden University Centre for Linguistics, Leiden University
[3]LCTM Solutions, Dublin, Ireland

Abstract

Multimedia localization workflows are inherently complex, and the demand for localized content continues to grow. This demand has attracted Language Service Providers (LSPs) to expand their activities into multimedia localization, offering subtitling and voice-over services. While a wide array of AI tools is available for these tasks, their value in increasing productivity in multimedia workflows for LSPs remains uncertain. This study evaluates the productivity, quality, cost, and time efficiency of three multimedia localization workflows, each incorporating varying levels of AI automation. Our findings indicate that workflows merely replacing human vendors with AI tools may result in quality degradation without justifying the productivity gains. In contrast, integrated workflows using specialized tools enhance productivity while maintaining quality, despite requiring additional training and adjustments to established practices.

1 Introduction

The demand to provide culturally and linguistically relevant content to global markets is at an all-time high. To remain competitive, businesses are pressured to produce broad-scale localized multimedia content faster and cheaper than ever before. As a result, Language Service Providers (LSPs) must find more efficient ways to provide multimedia localization services to meet these evolving client expectations. The evolution of artificial intelligence (AI) has introduced a plethora of tools designed to solve efficiency challenges for complex multimedia workflows. Existing research on AI tools in multimedia workflows has focused mainly on subtitling productivity, with studies investigating post-editing of machine-translated subtitles (Matusov et al., 2019; Koponen et al., 2020; Karakanta et al., 2022) or AI-enhanced subtitling workflows (Massidda and Sandrelli, 2023; Tardel, 2023). Research on AI-enhanced voice-over (VO) workflows is even scarcer, mainly focusing on quality assessment models (Spiteri Miggiani, 2024). In a recent survey, Mondello et al. (2024) evaluated several categories of multimedia AI tools for their suitability in LSP business operations. The categories evaluated were transcription, translation, subtitling, and VO, with tools ranging from modular task-specific applications, which proved to be most suitable for LSPs with low workloads, to fully integrated multimedia platforms, which demonstrated suitability for LSPs with high-volume workloads. However, the effectiveness of AI tools in enhancing productivity in real world multimedia workflows and the impact to end product quality have been largely unexplored.

Moreover, productivity gains must be weighed against the costs of leveraging AI. Incorporating AI in traditional workflows often requires additional computational power, specialized technical skills, training project managers and linguists in using new tools, and restructuring well-tested existing workflows. Thus, the questions for LSPs become: Are the productivity gains of leveraging AI worth the upfront cost and effort? Is the potential risk to end product quality worth the productivity gains?

In this paper, we address these questions by conducting a productivity study, comparing quality, time and cost gains in different AI localization workflows. This study focused on localizing two videos for subtitling and voice-over into Spanish-US and Simplified Chinese. To evaluate the gains and quality impact of AI tools on multimedia localization, we compared three different workflows: *i)* manual, where subtitling and VO were performed without the support of any AI tools, *ii)* cascaded, where the existing manual workflow was enhanced using automatic transcription, machine translation, and voice synthesis, and *iii)* integrated, where dedicated subtitling and VO platforms incorporating

Proceedings of Machine Translation Summit XX Volume 2, pages 1–7, June 23-27, 2025

AI were used to execute the workflow end-to-end. Our findings compare total time and cost, end product quality, and challenges associated with each workflow. Through the comparison of the traditional workflow against AI-augmented workflows for impact on quality, cost and time savings, our goal is to provide guidance to LSPs and other stakeholders on the implementation of AI automation in multimedia workflows.

2 Multimedia workflows and LSPs

Localizing multimedia content, such as videos, consists of projects focused on adapting audiovisual materials into different languages, in order to make them applicable and accessible to different linguistic and cultural audiences. These projects have traditionally been complex, time-consuming, and costly for LSPs, due to the fact that they require the involvement of a myriad of different specialized human resources to complete several different tasks, such as transcription, translation, subtitling, voice-over recording, and others. Some of the resources involved include: desktop publishing specialists, native-speaking and subject-matter expert linguists, video editors, subtitlers, voice-over artists, localization engineers, and Quality Assurance (QA) resources.

The nature of these workflows poses further challenges for LSPs. The fact that each step requires specialized and highly-trained resources not only increases the operational cost and execution time, but it also requires dedicated, proficient, and meticulous planning and resource allocation. Meeting tight deadlines becomes challenging, especially when handling large volumes of content or multiple language pairs simultaneously. Additionally, quality control entails ensuring consistency and quality across all stages and, since each step involves human intervention, this can introduce variability in the output quality. Maintaining high standards requires rigorous QA processes, further adding to the time and cost. The challenge of sourcing specialized subtitlers and voice-over artists to cover the diverse range of languages required by LSPs serves as a key motivation for this article. Unlike dedicated multimedia providers or streaming services whose main revenue comes from multimedia projects, LSPs have distinct needs and workflows that may differ from those in the audiovisual industry. This distinction underpins our decision to test these workflows in this context.

3 Methodology

3.1 Data

This study involved subtitling and voice-over of two brief videos[1] (approximately 11 minutes in total) with two speakers (male, female). The videos are an interview between two doctors and contain specialised terminology, spontaneous speech, on-screen text and no background noise or music.

3.2 Workflows

We compare quality, time and cost savings in localizing the videos through three separate workflows: manual, cascaded, and integrated. The tools selected for the cascaded and integrated workflows are the ones found to be most efficient for LSPs and providing high quality for life science content based on Mondello et al. (2024).

Manual workflow The manual workflow is the workflow traditionally followed by LSPs for subtitling and VO of videos. For subtitling, we started with a transcription of the videos, followed by a transcription QA step, and then prepared the scripts to be uploaded to our CAT Tool. The benefit of a CAT tool is that linguists can leverage translation memories (TM), glossaries and other resources, necessary to support the translation process in specialized domains. We proceeded with human translation and editing, which were handled by two different linguistic resources. The translated and edited script was sent to a subtitler who formatted the subtitle lengths and lines and burned them to the video. We sent the subtitled video to a linguist, who performed a video QA and identified issues to be resolved by a second round of subtitle editing. Once these updates were applied by the subtitler, the linguistic QA resource reviewed the videos again to ensure they were properly implemented and the subtitled video was final.

For voice-over, the workflow is equally, if not more, time-consuming and rigorous as for subtitling. We began with transcription, timecoding, and transcription QA to produce the final original scripts, which were then prepared by a different resource for CAT Tool upload. Then, two separate but equally qualified linguists handled the translation and editing of the scripts. Once these steps were completed, we sent the translated and edited scripts to voice-over talents, broken into different

[1] https://www.youtube.com/watch?v=9xla1ZccFno
https://www.youtube.com/watch?v=ibw6-qKQMSY

segments which needed to be delivered as separate recordings. The recorded audio clips were sent to a linguist, who reviewed them for accuracy, appropriate pronunciation and intonation, and faithfulness to the script. The segments that needed updates were sent back to the voice-over talents, along with the description of the issues, who re-recorded them and provided updated audio clips. The final audio clips were sent to a video engineer, who applied them to the original video, making sure the audio and video were appropriately aligned. The video engineer delivered a video that was sent to a linguist to perform a final and comprehensive video QA. The findings from this step were sent back to the video engineer for implementation. Finally, a linguist reviewed the updated video to verify that all updates were properly applied and to confirm the video was final.

Cascaded workflow The cascaded workflow followed the manual workflow but replaced the manual steps of transcription, translation and voice synthesis with AI tools followed by post-editing and/or review. The advantage of this strategy lies in maintaining the familiar workflow and processes for project managers and linguists, with the sole modification being the introduction of AI tools.

In the cascaded workflow, the transcription was done using Amazon Transcribe, which offers transcription with timestamp prediction. This can be done through the graphic user interface and requires uploading and downloading various files. For MT, we evaluated Amazon Translate, ChatGPT (OpenAI, 2023) and Google Translate. The outputs were similar in quality but we used Amazon Machine Translation in XTM since that is the main CAT tool in terms of familiarity for the linguists and the project managers. Once the translation has been generated, the subtitling and VO workflows separate. For subtitling, the scripts were converted to subtitle format (*.srt*), using a python script and the srt library[2]. The subtitles were then burned onto the video using ffmpeg[3]. For VO, the translated scripts were used for synthetic voice generation. Synthetic voices were generated through Amazon Polly for Spanish and Google Text-to-Speech for Chinese. This choice was motivated by the lack of availability of Chinese voices in Amazon. Applying the synthetic voices obtained to the video is performed by a sound engineer as in the manual workflow.

Integrated workflow The integrated workflow substituted the manual workflow, by moving the entire localization process under a dedicated platform. This process not only integrates AI tools, but also transforms the workflow by automating some of the project management tasks, avoiding the need for file conversions, importing and exporting documents and sharing them per email. We selected Matesub[4] for subtitling and Speechify Studio[5] for voice-over.

For subtitling, we uploaded the videos to Matesub and ran automatic transcription. Then, we had a linguist conduct a transcription QA step directly on the tool and apply any necessary corrections. Then, the source language subtitles were automatically translated into the target languages and linguists conducted post-editing. During the post-editing step, the linguists were also tasked with conducting a subtitle QA, which focused on correcting any issues related to length, synchronization and reading speed, legibility, positioning, and appropriate line breaks, among other issues.

For voice-over, we uploaded the videos to Speechify and ran automatic transcription. A linguist conducted a transcription QA step, directly on the tool and updated the script as needed. Then, we applied machine translation to the script and selected the synthetic voices that would be used to create the audio in the target languages. A separate group of linguists was asked to perform two simultaneous tasks: post-editing and audio QA. The post-editing portion focused on reviewing the translations and making any necessary updates in order to correct any translation issues and ensure accuracy to the source material. The audio QA task involved playing the audio and performing *live* updates in the translation (such as reducing, incrementing, or eliminating pauses, condensing the text, paraphrasing sections or switching terminology choices whenever necessary) in order to aid the synthetic voice generation tool in producing the most appropriate audio renditions of the written script, in terms of pronunciation, timing, and intonation.

3.3 Evaluation criteria

The evaluation focused on productivity gains and final quality. For productivity, the criteria included

[2]https://pypi.org/project/srt/
[3]https://ffmpeg.org/
[4]https://matesub.com/
[5]https://speechify.com/

time (hs) and cost ($) savings, reported both per task and as total. For quality, an evaluation of the final videos of the three workflows was conducted by a separate set of four expert linguists (one per language per task). To obtain unbiased quality results, each linguist assessed all three videos using an error annotation scheme, without knowing which video corresponded to which workflow. For the subtitled videos, professional subtitlers annotated errors related to translation quality, length, reading speed, synchronization, line segmentation and visual aspects (font, color, positioning). For voice-over, translators with experience as voice artists were recruited. They annotated errors related to fluency of speech (natural, fluent pronunciation), pace (too fast, too slow), synchronization to the speaker, background noise, room echo or distortion or robotic sound (audio that sounds flat, or does not convey emotion).[6]

The evaluation followed a penalty system. Critical errors (-1) are errors that impact comprehension completely or render outputs that are offensive or inappropriate for the target locale. Major errors (-0.5) are highly visible, could potentially impact comprehension, produce a mismatch between the speaker on screen or their gender and audio/subtitles, or result in a subtitle not being comfortable to read, for example, due to high reading speed, excessive length, lack of synchronization of about one second, or segmentation on linguistic units. Minor errors (-0.25) are errors that would be noticed, e.g. unnatural or artificial, and could decrease stylistic quality or fluency, but do not impact comprehension, or result in non-conforming but still readable subtitles, for example, subtitles that are max 3 characters above the length/reading speed limit, that appear fractions of a second before or after the corresponding dialogue, or that split linguistic units without impacting readability.

Finally, we report qualitative findings related to the efficiency in integration and usability of the tools in each workflow based on the feedback from the parties involved in the workflows (project managers, engineers, linguists).

4 Results

4.1 Productivity

The productivity gains in terms of time cost savings for subtitling and VO are shown in Tables 1 and 2 respectively. Both time and cost savings were very similar for both language pairs, therefore we only report them once. We found significant time and cost savings between the manual workflow and the cascaded and integrated workflows. The cascaded workflow for subtitling needed 10 working hours instead of 22 and the VO workflow needed 13.5 hours instead of 27 per language, resulting in a 41% cost reduction for the subtitling workflow and a 73% cost reduction for the VO workflow compared to the manual workflow. Finally, the integrated workflow showed the biggest time and cost reductions. Both subtitling and VO integrated workflows needed 7 working hours per language to complete the project and showed a 71% cost reduction for subtitling and 86% for VO when compared to the manual workflows.

While the cascaded workflow rendered quite considerable cost and time savings when compared to the manual workflow, we found that it was significantly more labor-intensive and complex than the integrated workflow. This was mostly due to the fact that the AI-assisted steps included in the cascade workflow had to be handled by a dedicated resource (engineer), since the selected tools needed a high level of technology expertise and were too complex for the project management and linguistic teams to be trained on during a feasible timeline. For this reason, even though the cascaded workflow showed considerable benefits, it may not be the most time- and cost-effective workflow, especially when considering its final quality results, which are explained in detail in the next section.

4.2 Quality

The quality assessment scores for the three workflows are shown in Table 3. In general, the manual workflow has the highest scores, closely followed by the integrated workflow, except for the Spanish subtitling where the integrated workflow remarkably resulted in an error-free output.

Comparing the scores among the workflows, for subtitling into Chinese, most minor errors in the manual workflow are related to synchronization and line segmentation, while in the cascaded and integrated workflows to positioning. In Spanish, the manual workflow showed a few stylistic issues, such as formality and acronyms. The cascaded workflow demonstrated severe quality issues, as shown by the negative score (-0.25). While the translation was of sufficient quality, the technical aspects showed several major synchronization

[6]The scorecards can be found at: https://tinyurl.com/3y2c6cby

	Manual		Cascaded		Integrated	
Task	Step	hs	Step	hs	Step	hs
Transcription	Transcription Transcription QA	3 1	Auto Transcription Transcription QA	0 2	Auto Transcription Transcription QA	0 2
Translation	Translation Editing	8 2	Machine Translation Post-editing Editing	0 3 2	Machine Translation Post-editing & Subtitle QA	0 3
Subtitling	Subtitle engineering Video QA	7 1	Subtitle engineering 1 Video QA Subtitle engineering 2	1 1 1	Final QA	2
Total		22		10		7
Cost reduction				41%		71%

Table 1: Productivity gains for subtitling in terms of time (hs) for each task and in total, as well as cost reduction (in percentage) of the total workflow.

	Manual		Cascaded		Integrated	
Task	Step	hs	Step	hs	Step	hs
Transcription	Transcription Transcription QA	3 1	Auto Transcription Transcription QA	0 2	Auto Transcription Transcription QA	0 2
Translation	Translation Editing	8 2	Machine Translation Post-editing Editing	0 3 3	Machine Translation Post-editing & VO QA	0 3
voice-over	VO Recording 1 Audio QA 1 VO Recording 2 Audio QA 2 Video Engineering 1 Video QA 1 Video Engineering 2 Video QA 2	4 1 1 0.5 4 1 1 0.5	Voice generation Engineering 1 Video QA Engineering 2 Video QA	2 1 1 0.5 1	Final QA	2
Total		27		13.5		7
Cost reduction				73%		86%

Table 2: Productivity gains for VO in terms of time (hs) for each task and in total, as well as cost reduction (in percentage) of the total workflow.

	En→Zh		En→Es	
	Sub	VO	Sub	VO
Manual	9.67	9.5	9.38	8.375
Cascaded	9.58	9.25	-0.25	2.625
Integrated	9.58	9.375	10.00	7.375

Table 3: Quality assessment of the final videos in the three workflows based on the error annotation. 10 equals to an error-free output.

and line break issues, as well as overlapping text. Specifically, "since most subtitles appear in one long line instead of two, the viewer must direct their eyes from end to end of the screen to read it". The integrated workflow was assessed as error-free, with the evaluator reporting that the translation quality is the best of all three conditions and having correct terminology, great grammar and syntax and good readability. Specifically for the technical aspects, the subtitles were found "centered and distributed in two lines, concise yet accurate, readable in full within the time they remain on screen and in synchrony with the sound. Font, colour and position are appropriate at all times, making sure that they never get on top of other on-screen text or important visual information".

For VO, in Chinese the manual workflow has the highest scores with only a few minor synchronization errors and cases where the voices sound unnatural. The cascaded workflow obtained lower scores, mainly due to synchronization and fluency issues. The evaluator reported that "the synchronization issue exists, but a bigger problem is that

both male and female voices sound quite robotic, making me believe that they were read by AI instead of humans". For the integrated workflow, a few minor synchronization issues were spotted. In Spanish, the output of the manual workflow was found fluent, with some minor synchronization and overmodulation issues in some of the sections. As in subtitling, the cascaded workflow scored low due to several major and minor fluency issues, with voice sometimes sounding robotic and distorted. The audio "sounds like reading a list of non-related sentences with no natural intonation, chopped at random points that do not follow the original syntax". The scores for the integrated workflow are higher. A few synchronization issues were reported, for example lip movements at the end of sentences. "Male VO has good fluency, pace and intonation in most sections and is easy on the ears. Female VO is more robotic sounding with exaggerated intonation, particularly in questions or exclamations".

We found that the integrated workflow performed remarkably well, especially for subtitling. Additionally, when considering how extensive the time and cost savings were for this workflow, our assessment is that this can be an extremely beneficial option for clients who need fast and cost-effective localization services for multimedia assets of this nature. The subtitled videos were found to be of very high quality by our linguistic reviewers and, while there were a few existing issues in the VO final videos, none of them were related to comprehension, ambiguity, or readability.

5 Recommendations

The goal of this experiment was to identify potential strategies of making the process of localization of multimedia products leaner and more cost effective. We think we have achieved that. Here are some recommendations to LSPs who want to test AI for such workflows:

- If you are going to apply AI in one task only, you might want to choose a standalone technology, rather than a platform.
- It is important to test the quality of the output in order to assess the human effort that will be required afterwards.
- Check the format of the output, as some formats are more user friendly than others: can you work with it directly?
- Make sure the languages required are fully covered by the provider, as there is variability in that regard.
- Visualize the workflow and add quality checks after AI.
- Bear in mind that most subtitling/VO AI tools do not have basic functionalities such as spell and QA checks, glossary or TM support.
- Decide who within your team is going to be the owner when it comes to applying the technology: will it be a developer, a technically competent project manager?
- If you are going to use integrated platforms, you will need to train your team; you might want to add that to your cost.

6 Conclusion

Our productivity and quality analysis showed that AI technologies can be used successfully in the localization of multimedia products. Amongst all the tasks analysed (transcription, translation, subtitle generation and voice-over), the one that is still lacking finesse and human quality is artificial voice generation. Having said that, there are a large range of voice generation providers that were not tested during this exercise. A key observation from this experiment is that most AI tools, especially those offering AI dubbing/VO, are not designed with post-editing in mind, as they lack fundamental functionalities commonly found in CAT tools. At the end of the day, companies need to strike a balance between quality of the end AI product, cost, learning curve and experience. The human element is still important in the form of post-editing (the transcribed source and the translation) and QA (subtitles and voice-over). The integrated workflow, with the use of platforms designed for the specific tasks, is the real winner in terms of quality and productivity, especially for subtitle generation. But it implies a steep learning curve, as language workers need to learn how to work in an alien environment. One of the clear conclusions of this experiment is that there is a need for training language providers workforce on the use of AI technologies; not only on the physical use of the various interfaces, but on the fundamentals of AI. By doing that, production teams will understand the possibilities of AI on their day to day tasks.

References

Alina Karakanta, Luisa Bentivogli, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2022. Post-editing in automatic subtitling: A subtitlers' perspective. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 261–270, Ghent, Belgium. European Association for Machine Translation.

Maarit Koponen, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020. MT for subtitling: User evaluation of post-editing productivity. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 115–124, Lisboa, Portugal. European Association for Machine Translation.

Serenella Massidda and Annalisa Sandrelli. 2023. ¡ sub! localisation workflows (th) at work. *Translation and Translanguaging in Multilingual Contexts*, 9(3):298–315.

Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou. 2019. Customizing neural machine translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy. Association for Computational Linguistics.

Ashley Mondello, Sahil Rasane, Alina Karakanta, and Laura Casanellas. 2024. Leveraging AI technologies for enhanced multimedia localization. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 145–151, Chicago, USA. Association for Machine Translation in the Americas.

Giselle Spiteri Miggiani. 2024. Quality assessment tools for studio and ai-generated dubs and voice-overs. *Parallèles*, 2.

Anke Tardel. 2023. A proposed workflow model for researching production processes in subtitling. *TransKom*, 16(1):140–173.