
VIDEO LOCALIZATION

Sherlock: 90% accuracy isn't enough

Why AI video localization still needs a human expert.

Table of contents

01	The AI Overwhelm of 2026	03
02	Introducing Sherlock	04
03	How Video Localization Works	05
04	The Localization Pipeline	06
05	Four AI Tools & Their Limits	07
06	The OST Case Study	09
07	A Unified Workflow, One Expert Driver	10
08	Work With Compass	12

01 Everyone has the tools. Nobody has the workflow.

In March 2026, directors within two large media companies complained that the abundance of new AI tools and models had overwhelmed them and their staff.

“We spend so much time evaluating new AI that we struggle to manage our everyday workflows and meet our content delivery schedules”.

Similarly, in May 2026, a Sr. Manager in a global eCommerce company complained that he had been reviewing AI models all morning and his head was spinning.

In the last month, client meetings with three large corporations revealed a common theme: each had invested in multiple AI tools (up to 5-6 models) but did not have the engineering staff to build the workflow integration of these tools. None of the AI models by themselves produced output that met company standards, so cumbersome manual steps were being used to complete the work of meeting requirements and schedules.

At Compass Languages we coined the term “Harnesses AI” to describe our approach to selecting appropriate AI tools and then building the infrastructure around multiple tools to deliver the performance envisioned. Two principles learned through our journey:



PRINCIPLE ONE

Every AI tool has limitations and must be supplemented by precise expert touchpoints.



PRINCIPLE TWO

Assuming that different AI models will integrate on their own sets teams up for disappointment.

02 Harnessed AI, applied to video

One of the clearest cases where we've applied this at Compass Languages is our video localization platform, Sherlock. To harness AI means giving a human expert fast, precise interfaces to adjust or correct AI output into a deliverable, quality product. For video, this is an iterative process — adjusting on-screen text and audio content against a precisely timed but changing background. AI gives the human operator speed and efficiency; the infrastructure around that expert is what turns AI speed into a finished video. In one recent case study of 1,000 videos, for example, only 20% could be handled by AI alone — the data behind that number is below.

Sherlock is an AI-powered localization system for video and images designed to match the quality of human-only processes. The evidence is in the workflow data below: a 90% baseline transcription accuracy, machine translation lifted from an 80-90% baseline to the zero-critical-error standard clients require, and an OST process validated across a 1,000-video case study. That combination — AI speed plus human quality control — is what lets Sherlock deliver quality localization of any video or image, at the turnaround times and cost that today's AI-driven market demands.

90%

**Transcription
baseline**

ACCURACY ACROSS
THOUSANDS OF FILES

97%

**Accuracy buyers
require**

VS. 80-90% RAW
MACHINE OUTPUT

20%

Handled by AI alone

OF A 1,000-VIDEO CASE
STUDY

Let's walk through how it actually works, step by step, with the accuracy data behind each stage.

03 Seven stages, one finished video

Localizing videos requires multiple steps:

- 1 Audio Content (in any language) transcribed into text.
- 2 Audio content translated into target language(s)
- 3 Voiceover of translated audio content via TTS platform into each target language
- 4 Video OST (on-screen text) extracted
- 5 Video OST translated into target languages
- 6 Translated OST inserted into each video frame including DTP adjustments and/or in-painting if necessary.
- 7 Sync re-packaged video — timing of audio/OST/background must be adjusted to compensate for new audio length.

THE ONE-BUTTON MYTH

Many suppliers of AI want to claim that their technology can produce a fully automated, one-touch-button, that performs all the above tasks without a human-in-the-loop. We understand the response by these suppliers. Since 2024, we have heard the demands from some of our own clients. **Why can't AI just convert my video content into any language with a push of a button?**

04 The end-to-end pipeline

A single source video splits into two parallel pipelines — a TTS (audio) pipeline and an OST (on-screen text) pipeline — each processed by automation and AI tools, then routed to linguists, engineers, and QA before both streams merge back into one localized final video. The schematic below maps that flow, end to end.

Video localization pipeline – TTS & OST streams

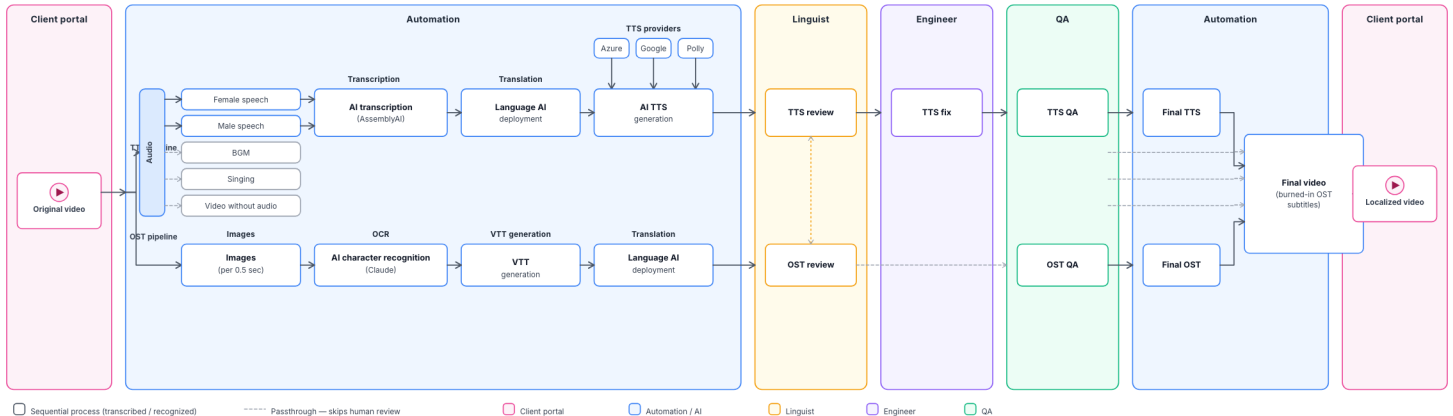


FIGURE 1 — ONE EXAMPLE OF A TYPICAL MULTI-MEDIA VIDEO LOCALIZATION WORKFLOW

05 Four AI tools, four ceilings

TOOL

1

Transcription

95% BEST CASE

90% OVERALL

Tools that take recorded words spoken within any audio or video file format and create a written text from it are surprisingly good — in quite a few languages. Depending upon the source language spoken, and the clarity of the speaker's voice and articulation of words, the accuracy can be as high as 95%. After performing AI transcription on thousands of files, the overall accuracy is probably in the 90% range. The limitations of AI transcription tools are often the same as those encountered by human transcribers but are amplified because human transcribers have methods to overcome the obstacles.

LIMITATIONS OF AI TRANSCRIPTION TOOLS

- ◆ Differentiating accents/dialects within audio content
- ◆ Acronyms, certain conjoined phrases
- ◆ Text generated is lacking correct punctuation — run-on sentences, lack of capitalization, etc., prohibiting a clean script for the TTS portion of the process

TOOL

2

Translation

80-90% RAW

97%+ REQUIRED

LLMs have exploded in number and capabilities. Built upon billions of words translated, machine translation engines currently have tremendous breadth in subject matter. For users requiring precision in a particular subject matter, or to adapt output to a client's particular style, the ability to initially train engines and then have them learn over time is impressive. Without a linguistic specialist performing post-editing, we are seeing machine translation hit an average accuracy in the 80-90% range. Just a few years ago this average was only 65-75%. The largest consumers of translation require 97% or better in measured accuracy - with no critical or major errors. Therefore, a human linguist is needed to correct translation errors to achieve customer expectations and protect the client's reputation in the marketplace.

LIMITATIONS OF AI TRANSLATION MODELS

- ◆ Adapting translation to customer-preferred terminology and style
- ◆ Literal translation vs. context-driven due to segmentation rules in the AI model
- ◆ Managing creative writing, idioms, slang

TOOL
3

Text-to-Speech

The most appealing and impactful AI tools we have worked with over the last decade are Text-to-Speech models. In the past, we used primarily (3) TTS software tools, but now there are new models available that provide more natural speech and some limited emotional appeal. More importantly, more voice choices per target language are entering the market.

LIMITATIONS OF AI TEXT-TO-SPEECH MODELS

- ◆ Overcoming the monotone nature of an AI-read script
- ◆ Annunciation problems (including standardization) with proper names and acronyms — and with words that intentionally stay in the source language, e.g. “coup d'état” should keep its French pronunciation rather than being anglicized
- ◆ Matching a tone of voice to the accent (or dialect) of a script — e.g., fr_CA vs. fr_FR

TOOL
4

On-Screen Text + In-painting

For videos, and to a lesser extent images, the most difficult frontier for automation to conquer is the extraction and replacement of text against backgrounds that range from simple to very complex. Many video files are in .mp4 or similar formats for the ease of file-sharing. On-screen Text (OST) is no longer an accessible layer within a video frame. The traditional method of coloring over all text (textbox overlay) with a blended color and then inserting new text within the colored area only works when the background is solid and static.

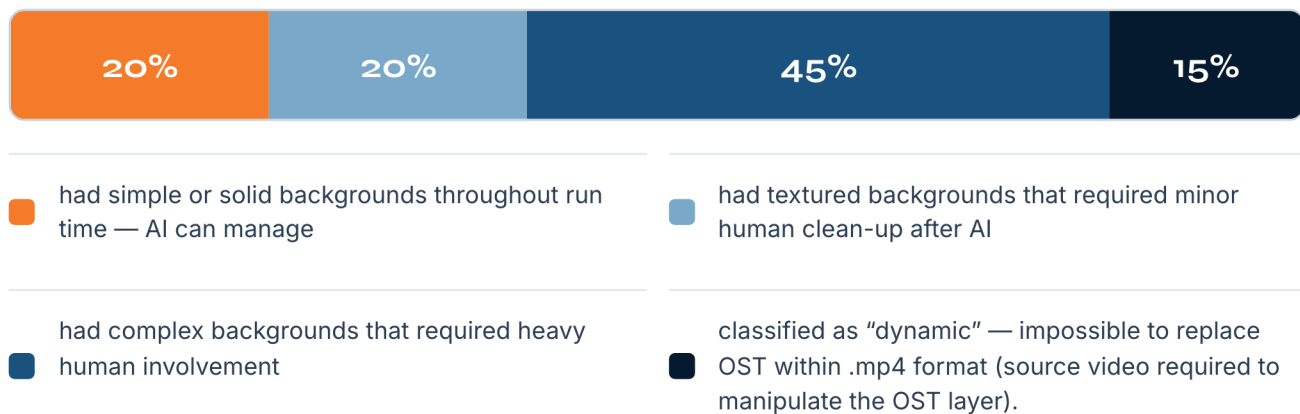
LIMITATIONS OF AI TOOLS FOR OST

- ◆ Tools cannot detect and match backgrounds that are textured, dynamic, or photographic in source.
- ◆ Automation can only assist in reviewing 30 frames/second and performing a copy/paste function of OST when the OST and background are static.
- ◆ Generative AI tools tend to leave artifacts that must be cleaned up

06 OST case study: 1,000 videos

In a large annotation project performed in 2025, a client gave us 1,000 videos (all .mp4 format) and asked us to perform either AI or human in-painting to localize the on-screen text of the short advertising videos. The single most difficult requirement was the condition of “no artifact”. This means a flawless localized video, no blemishes. Old OST completely removed and new OST added with no blemishes in the background presentation.

From an assortment of 1,000 advertising videos (20 to 60 seconds in length):



conclusions

(1) only 20% of videos managed by AI tools alone; (2) when human in-painting was required (for complex videos) the execution time by specialists averaged over 12 hours per video; (3) AI technology not far enough along to manage the majority of videos.

07 Sherlock: a unified workflow with an expert driver

A pure AI-produced localized video does not meet the quality requirements of most global companies. Yet, these same companies want the responsiveness (time-to-market) advantages of AI localization tools not to mention the budget advantages that come with automated solutions.

Sherlock is a video localization platform that can deliver localized advertising videos (< 60 seconds) within 48 hours, at a price that starts at \$75, in any of 25 languages, and is a product quality-controlled by a human expert. This harnessed AI operation addresses all of the AI tool deficiencies identified above.

48h

Turnaround

FOR VIDEOS UNDER 60
SECONDS

\$75

Starting price

PER LOCALIZED VIDEO

25

Languages

ANY TARGET MARKET

THE DRIVER / OPERATOR

The "driver/operator" of the platform is a professional linguist/engineer who uses the power of AI to perform multiple quality edits to assure that the localized video meets all client requirements. This eliminates one of the greatest problems in global advertising — a sloppy localization effort that deters the buyer's interest and potentially impacts the reputation of the brand in the marketplace.

Good news for the experts

Sherlock has been an important piece of good news for professional linguists and localization engineers. Generative AI models have depleted the work of many localization professionals. With Sherlock, they are trained how to adjust any one of four AI tools — and usually all of them in an iterative process — to produce a quality product they are proud to approve. Where before these professionals contributed to only one section of the localization process, now they control the edits of all segments in an efficient and user-friendly platform.

It is the win-win scenario that hopefully keeps global companies **expanding their reach and impact with confidence.**



Whether the product is a marketing video, a training video, or a product information video, Sherlock delivers dozens of videos each day while operating in 24/7 environments for our largest clients.



FOR THE CLIENT

AI speed and budget, with the zero-critical-error quality that protects the brand in every market.



FOR THE EXPERT

Control of every segment of the process — a product the linguist/engineer is proud to approve.

— YOUR NEXT MOVE

See Sherlock deliver at AI speed without giving up quality

See how Sherlock delivers localized video at AI speed without giving up quality. Schedule a meeting to discuss your use case.

Visit compasslanguages.com to learn more



compasslanguages.com