# Redefining Broadcast Accessibility

## WITH AI-POWERED CAPTION QUALITY CONTROL

BY SANA AFSER, SENIOR ENGINEERING MANAGER AT INTERRA SYSTEMS

The demand for high-quality captioning has become a cornerstone of accessibility, compliance, and viewer engagement. Business Research Insights projects that the global subtitle market will grow from $0.317 billion in 2024 to $0.76 billion by 2033, driven by increasing demand for multilingual content and accessibility across streaming platforms. OTT platforms now account for nearly half of this demand, according to Global Growth Insights, with major services implementing multilingual subtitles for the majority of their catalogues. At the same time, more than two-thirds of content creators are adopting captioning tools to expand their reach and improve viewer engagement.

As broadcasters and streaming platforms scale their content output, traditional methods of caption quality control (QC) — largely reliant on manual review — are proving increasingly inadequate. These approaches are resource-intensive and prone to error, especially when applied to large volumes of multilingual and visually complex content.

By automating intricate QC tasks and enhancing the precision of captioning workflows, AI enables media organisations to meet regulatory standards more efficiently while improving the overall viewer experience. This article explores how AI technologies, ranging


from speech recognition and computer vision to natural language processing and multimodal systems, are redefining the standards and practices of caption QC in the broadcast industry.

## THE CHALLENGE OF MANUAL CAPTION QC

Captioning is governed by stringent regulatory frameworks, most notably those set by the Federal Communications Commission (FCC), which require captions to be synchronised with audio, accurate in content, and complete in coverage. Beyond these core requirements, captions must also adhere to technical and aesthetic guidelines, including constraints on reading speed, display duration, segmentation at natural linguistic breaks, and layout considerations such as row and column limits. Captions must also be carefully placed to avoid obscuring important visual elements like speaker faces, on-screen text or graphical overlays.

Ensuring compliance with these multifaceted requirements through manual QC is a daunting task. Human reviewers must repeatedly listen to audio tracks, scrutinise video frames and verify linguistic accuracy across multiple languages. This process is resource-intensive and prone to error, particularly when scaled across diverse content types and delivery platforms. Moreover, frequent content edits, such as scene additions or deletions, can desynchronise captions, requiring repeated manual adjustments. As the volume and complexity of media content continue to grow, the limitations of manual QC have become a significant bottleneck in production workflows.

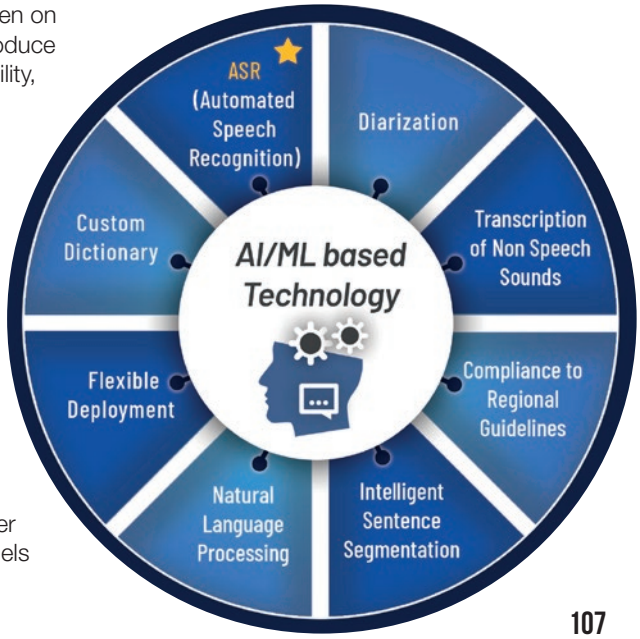
## AI TECHNOLOGIES DRIVING CAPTION QC INNOVATION

AI offers a suite of technologies that can automate and enhance nearly every aspect of the caption QC process. These tools not only reduce the burden on human reviewers but also introduce new levels of accuracy, scalability, and contextual awareness.

### 1. AUTOMATIC SPEECH RECOGNITION (ASR)

ASR systems use deep learning to transcribe spoken dialogue into text with precise timing. Unlike manual transcription, which can take hours, modern ASR engines operate faster than real time and adapt to accents, background noise, and speaker changes. Advanced ASR models leverage transformers and end-to-end neural networks to improve transcription quality. They also support speaker diarisation and punctuation correction, which are essential for readable captions. Multilingual ASR capabilities further enable global content delivery without the need for in-house language experts.

In multilingual workflows, ASR-based systems can even compare captions and audio using semantic similarity, such as cosine similarity, to ensure alignment across languages. This technique is especially useful when the audio and caption languages differ.




## 2. COMPUTER VISION (CV)

Caption placement is critical to viewer experience. Poorly positioned captions can obscure faces, scores or embedded text. CV models detect visual elements in each frame and flag overlap issues. These models can also identify lip movements and speaker locations, allowing captions to be placed near the speaking character. This not only improves readability but also enhances synchronisation. CV-driven placement algorithms ensure captions are optimally positioned without manual inspection.

Additionally, CV-based systems can detect scene changes, a process that plays a crucial role in refining caption timing. By analysing pixel values, edge transitions or motion vectors between frames, automated systems can identify visual transitions — such as shifts in location, time or character focus — and adjust caption in/out points accordingly to ensure captions appear and disappear in sync. This helps prevent captions from lingering across scene boundaries, which can confuse viewers and disrupt narrative flow. Traditional workflows often overlook these subtleties, but automated solutions ensure captions align with shot boundaries, enhancing coherence and viewer comprehension.

## 3. MACHINE TRANSLATION (MT)

Global distribution often requires subtitles in 30+ languages. Traditional MT systems struggle with context, leading to literal translations that miss the tone or intent of the original dialogue. AI-powered MT models, especially those integrated with LLMs, use contextual embeddings to produce fluent, culturally appropriate translations. New evaluation metrics assess translation quality beyond grammar, considering fluency and cultural relevance. Specialised MT models trained on subtitle datasets further improve dialogue-style translations, making multilingual QC more reliable and scalable.

Beyond timing and translation, AI also ensures metadata conformance — verifying that caption files meet standards for frame rate, timecodes, character sets and indexing. This prevents decoding errors and ensures compatibility across platforms. Additionally, AI-driven spell checkers can detect and correct typographical errors and homophones (e.g., “their” vs. “they’re”) using context-aware language models. These refinements enhance professionalism and viewer trust, especially in fast-paced or technical content where manual review may miss subtle mistakes.

AI systems can also verify that the correct language is paired with the correct audio track using Language Identification (LID), helping prevent distribution errors in multilingual environments. This ensures that viewers receive accurate subtitles in their intended language, improving accessibility.

## 4. NATURAL LANGUAGE PROCESSING (NLP)

Effective captions are not just accurate — they must be well-segmented and readable. NLP tools analyse sentence structure using dependency trees, which map syntactic relationships between words. This allows captions to be broken at natural clause boundaries, preserving the flow of dialogue. Dynamic programming algorithms optimise segmentation by balancing character limits, line breaks and proportional line lengths. This ensures captions are linguistically coherent and visually balanced, without requiring manual formatting.

For example, NLP-based systems can ensure that dependent words — such as negations, currency symbols, or compound names — are not split across lines. They can also maintain visual gaps between captions to signal speaker changes, improving viewer comprehension. (See Figure 1)

Caption segmentation becomes even more complex in multilingual environments. Languages like German, Japanese or Arabic present unique challenges — compound words, lack of spacing or different punctuation norms. AI-powered NLP systems overcome these barriers by analysing sentence structure and meaning, identifying optimal break points without requiring fluency in each language. This allows captions to maintain linguistic integrity and readability across diverse audiences, eliminating the need for costly language specialists and reducing segmentation errors.

## 5. MULTIMODAL AI

The most transformative advances come from multimodal AI systems that integrate audio, video and text. These models understand context across modalities, enabling more nuanced QC. For example, multimodal AI can detect non-verbal events like laughter, applause or gunshots and insert appropriate audio descriptions (e.g., “[Car honks loudly]”). It can also interpret facial expressions to adjust tone in captions (e.g., replacing “Okay.” with “Okay…” [hesitant tone]).

By combining ASR with lip-reading and scene-change detection, multimodal systems improve synchronisation and speaker identification, even in overlapping

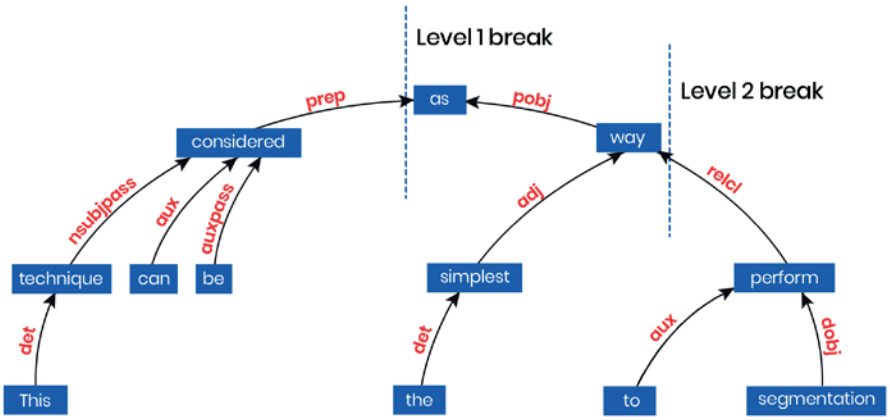



**Figure 1: Dependency tree illustrating the syntactic segmentation of the sentence “this technique can be considered as the simplest way to perform segmentation.”**

dialogue scenarios. This holistic approach ensures captions reflect not just what is said, but how and when it is said.

Multimodal AI also supports profanity filtering by using contextual analysis to distinguish between offensive and benign uses of similar words, ensuring that censorship is applied appropriately. This is especially important for family-friendly or educational content, where inappropriate language must be handled with precision.

## REAL-WORLD IMPACT AND FUTURE OUTLOOK

The integration of AI into caption QC workflows is already yielding significant benefits. Automation reduces the time and cost associated with manual review, while improving consistency and compliance. AI systems can operate at scale, handling vast libraries of content across multiple languages and formats. They also enhance accessibility by ensuring that captions are accurate, well-timed and contextually appropriate for diverse audiences.

Many modern captioning platforms now consolidate transcription, review and compliance checks into a unified interface. This streamlines collaboration between automated systems and human reviewers, reducing errors and accelerating delivery timelines. Moreover, the advent of on-device AI chips is enabling real-time processing without reliance on cloud infrastructure. This not only improves performance but also addresses concerns around data privacy and security — an increasingly important consideration in media production environments.

Looking ahead, the capabilities of AI in captioning are poised to expand even further. Future systems may offer adaptive captioning that responds to individual viewer preferences, or real-time sentiment-aware subtitles that reflect the emotional tone of a scene. As AI models continue to evolve, they will unlock new possibilities for making media content more inclusive, engaging and expressive.

**www.interrasystems.com**

**Future systems may offer adaptive captioning that responds to individual viewer preferences, or real-time sentiment-aware subtitles that reflect the emotional tone of a scene.**