Journal of
Audiovisual Translation

Using ASR Tools to Produce Automatic Subtitles for TV Broadcasting: A Cross-Linguistic Comparative Analysis

Elena Davitti ✉

University of Surrey, Centre for Translation Studies

Annalisa Sandrelli ✉

Università degli Studi Internazionali di Roma

Tomasz Korybski ✉

University of Surrey, Centre for Translation Studies

Yuan Zou ✉

University of Surrey, Centre for Translation Studies

Constantin Orasan ✉

University of Surrey, Centre for Translation Studies and Surrey Institute for People-Centred Artificial Intelligence

Sabine Braun ✉

University of Surrey, Centre for Translation Studies and Surrey Institute for People-Centred Artificial Intelligence

Citation: Davitti, E., Sandrelli, A., Korybski, T., Zou, Y., Orasan, C., & Braun, S. Using ASR Tools to Produce Automatic Subtitles for TV Broadcasting: A Cross-Linguistic Comparative Analysis. Journal of Audiovisual Translation, 7(2), 1–35. https://doi.org/10.47476/jat.v7i2.2024.305

Editor(s): N. Reviers, J. Neves & G. Vercauteren.

Received: November 23. 2023

Accepted: April 26, 2024

Published: December 19, 2024

Abstract

This paper explores the potential application of Automatic Speech Recognition (ASR) tools to produce intralingual subtitles in broadcasting. It is based on a study commissioned by an international broadcaster to compare the performance of two non-customised ASR tools in producing automatic subtitles for various genres of pre-recorded content in English and Italian. Our evaluation focused on the accuracy of the transcript and the readability of the subtitles. Accuracy was assessed by using an adaptation of the NER and NTR models (Romero-Fresco & Pérez, 2015, Romero-Fresco & Pöchhacker, 2017) to categorise ASR-generated errors by type (content- or form-related) and level of severity (minor, standard and serious). Readability was

✉ e.davitti@surrey.ac.uk, https://orcid.org/0000-0002-7156-9275
✉ annalisa.sandrelli@unint.eu, https://orcid.org/0000-0001-6010-4862
✉ t.korybski@surrey.ac.uk, https://orcid.org/0000-0003-2353-0816
✉ y.zou@surrey.ac.uk, https://orcid.org/0000-0001-8774-9978
✉ c.orasan@surrey.ac.uk, https://orcid.org/0000-0003-2067-8890
✉ s.braun@surrey.ac.uk, https://orcid.org/0000-0002-6187-3812

qualitatively analysed by examining text segmentation, namely line breaks and subtitle breaks. Our findings indicate that all the ASR outputs fell short of the 98% minimum accuracy threshold expected in the broadcasting industry, with notably better performance in English. Challenges in subtitle segmentation and timing were found to compound accuracy-related aspects. The discussion highlights the potential of ASR-generated subtitles to represent an intermediate step and identifies areas for improvement through ASR fine-tuning and human post-editing to achieve broadcast-ready subtitles.

Key words: Automatic Speech Recognition (ASR), subtitling, captioning, broadcasting, English, Italian, accuracy, readability, segmentation, post-editing, human effort.

Introduction

The technologisation process that has hit the audiovisual translation (AVT) sector in recent years includes the use of cloud platforms and a tentative integration of translation technologies (including Computer Assisted Translation, Machine Translation, and Automatic Speech Recognition) into workflows (Bolaños García-Escribano et al., 2021). The skyrocketing demand for media translation and accessibility services during the 2020 coronavirus pandemic sent language service providers (LSPs) looking for technological solutions in a quest for increased productivity.

The popularity of streaming services has given subtitling greater visibility even in countries traditionally associated with other AVT practices. Moreover, the availability of automatically generated subtitles on YouTube and platforms like Zoom has created an expectation for (intralingual and interlingual) subtitles to be present for any type of content, from live to pre-recorded materials, from broadcasting to corporate videos (Bolaños García-Escribano & Declercq, 2023). This has important repercussions on the organisation of work, the skills required of subtitlers and working conditions. It has also driven down translation rates and has led to many AVT translators leaving the market or considering a career change, resulting in a "talent crunch" in the sector (Bryant, 2021; Berman, 2022). Thus, full or partial automation of the subtitling process is seen by the industry as the only way to cope with mounting pressure.

The advancement of Automatic Speech Recognition (ASR) tools, which are now capable of transcribing spoken input and producing a time-coded output, thereby creating automatic subtitles, is of particular relevance (Díaz Cintas & Massidda, 2019). However, the reliability of these tools in industrial processes remains uncertain due to their substantial variations across languages and their optimal performance being contingent on specific conditions, namely clean audio, standard accents, absence of voice overlapping and minimal background noise. Human editing is still essential to correct transcription errors, check timing and verify text segmentation in subtitles. Moreover, "[i]t remains unclear whether a corrupt ASR transcript might still facilitate subtitling processes, and under which circumstances, or the role which the visual part of the video plays during computer-assisted transcription and translation" (Tardel, 2020, p. 81). Few studies are available on the actual effectiveness of such tools on different audiovisual genres.

This paper contributes to the debate, focusing on the potential use of ASR tools for intralingual subtitling. Two different ASR tools were trialled by an international broadcaster to produce automatic subtitles for pre-recorded content in English and in Italian, namely a British talk show and a US feature film dubbed into Italian. A study was commissioned to compare the performance of the two tools in each language and qualitatively evaluate the human effort needed for subsequent editing. Before outlining our methodology (§1.2), we will provide a short overview of recent contributions on the use of ASR tools in subtitling (§1.1).

1. Conceptual and Methodological Framework

The literature on the use of ASR tools in AVT, particularly subtitling, is rather scant, as the pace of research activities in this field cannot keep up with the pace of software development. An emerging body of research has focused on live subtitling and how AI-driven technologies like ASR and Machine Translation (MT) can be integrated into increasingly hybrid workflows[1]. Given the focus of this paper, the brief overview below highlights studies specifically focussing on the integration of ASR into the production of offline subtitles.

1.1. Review of Relevant Literature

The EU-funded CompASS project investigated the use of ASR to produce automatic intralingual subtitles for films, later automatically translated into English (Tardel, 2020). The project compared the work of 12 professional subtitlers and 13 translation students in transcription tasks under different conditions, i.e., working directly from the soundtrack, with an ASR transcript and with a human-produced script. Although the availability of an ASR transcript reduced the technical effort involved in transcription, it did not speed up task completion. Given relatively high word error rate (WER) scores in both English and German, the ASR transcripts were not deemed particularly useful. However, Tardel contends that "written assistance in AVT could be a contributor to decreased effort, if the quality of transcripts meets certain levels" (2020, p. 100).

Karakanta et al. (2022) conducted an experiment involving 22 professional subtitlers who post-edited automatically generated subtitles for short clips containing challenging elements, such as background noise, slang, overlapping speech and multi-speaker events. After the task, the participants completed a user-experience questionnaire. Although the results ranged from neutral to positive, most participants identified major speech recognition and spotting errors. Moreover, many noted critical issues, such as subtitles straddling across shot changes, inadequate timing, and poor text segmentation.

Tuominen et al. (2023) investigated the use of ASR and MT tools to translate Finnish news programmes into English. Potential users with little or no knowledge of Finnish watched some news clips before participating in a focus group and filling in a questionnaire. While the automatic subtitles were deemed helpful, many participants perceived their overall quality as inadequate, highlighting mistranslations, excessive speed, timing, and segmentation issues as key problems. As unscripted broadcasts and news programmes are characterised by a high speed of delivery, hesitations, interruptions, and other disfluencies, ASR may not be well-suited for this genre. This suggests that human editing remains essential for text condensation and improved readability.

[1] For example, see outcomes from the SMART (*Shaping Multilingual Access through Respeaking Technology*, ESRC, ES/T002530/1, 2020–2023, key findings and dissemination video) and MATRIC (*Machine Translation and Respeaking in Interlingual Communication,* E3, 2020–2022, key findings).

Finally, the *¡Sub!: Localisation Workflows that Work* project (2020–2021) and its follow-up *¡Sub!2* (2021–2022) compared three cloud subtitling workflows with varying degrees of automation (Sandrelli, 2024, and *forthcoming*; Massidda & Sandrelli, 2023). One workflow used only cloud-based subtitling tools, a "semi-automated" workflow added an ASR tool for automatic captioning, and a "fully automated" one included both ASR and MT. The source materials were two science documentaries, to be subtitled from English and Spanish into Italian. The participants worked in subtitling teams over a three-week period and were exposed to all three workflows, each team in a different order. The results indicate that, although ASR and MT reduced turnaround times to an extent, workflow duration decreased over time regardless of the technology employed, suggesting a higher importance of teamwork and familiarity with tools and procedures. Moreover, integrating ASR into the workflow became easier over time, with the teams experiencing progressively fewer errors. However, the teams using both ASR and MT tools in the first week produced a very high number of linguistic and technical errors, with marginal performance improvement over time. This suggests that integrating MT into workflows requires a longer adaptation period and is more challenging and time-consuming.

Against this background, this study was commissioned by an international broadcaster to investigate the human input needed to edit ASR-generated subtitles for three languages (English, Italian and German) and three broadcast genres (talk show, film and TV series). This paper reports the cross-linguistic comparative analysis for two of the languages, English and Italian.

1.2. Data and Methodology

This section provides an overview of the characteristics of the English and Italian videos and ASR output files. Each clip was approximately 10 minutes long and represented one of the two chosen genres.

The ENGLISH clip is from an episode of the talk show *Last Week with John Oliver*. It is characterised by fast-to-very-fast speech delivered by the host in a British accent, featuring many proper names and (pop)cultural references, as well as snippets of TV news stories and interviews, adding to accent variety. The host delivers largely scripted material, presumably read out from the autocue. Prosody is driven by the script and by speed, and sometimes deviates from the natural prosody of spoken language. Background noises are limited to the opening of the show (music, applause) and laughter between the fragments of the narrative, background noises and overlaps in the video snippets used.

The ITALIAN clip is from the Italian dubbed version of *The War with Grandpa* (dir. Tim Hill, 2020), an American comedy. Although dubbed films tend to be characterised by fewer disfluencies than their original language counterparts, the clip presents some challenging features. Unlike the English material, the Italian material is entirely dialogic and presents fast dialogues, voice overlaps, background noise, and frequent scene changes involving various individuals, such as children in a school or people arguing outside a supermarket. Other challenging aspects are the high pitch of the children’s voices and a scene in which a child sings softly to herself.

The material for analysis included two output files (per recording and language) produced by the Broadstream and Amberscript ASR software solutions (see Table 1). They were text files with time codes, and there was no access to back-end information from the providers[2]. In this commissioned project, the ASR tools were implemented off-the-shelf with default settings and were not specifically trained on similar materials or customised for the purpose of subtitling.

Table 1

Materials

Language	Genre	Words per minute	Source material	Broadstream ASR output	Amberscript ASR output
EN > EN	Talk show	~190 wpm	1,768 words	1,748 words	1,658 words
IT > IT	Film	~130 wpm	1,081 words	1,073 words	984 words

Our analysis focused on two key dimensions of the subtitled output: **accuracy and readability**, which are not systematically accounted for in existing quality models.[3] In this paper, we refer to accuracy as the informativeness of the final product, i.e. offline subtitles, and the extent to which the text is devoid of misleading information. To measure ASR **accuracy**, we drew on two existing models for accuracy evaluation in live subtitling via respeaking, namely NER for intralingual subtitles (Romero-Fresco & Martínez, 2015) and NTR for interlingual subtitles (Romero-Fresco & Pöchhacker, 2017).

Figures 1 and 2

NER and NTR model formula

$$\text{Accuracy} = \frac{N - E - R}{N} \times 100 = \%$$

$$\text{Accuracy} = \frac{N - T - R}{N} \times 100 = \%$$

The NER model assesses the accuracy of live intralingual subtitles resulting from the interaction between human subtitlers and speech recognition technology. It accounts for both human edition

[2] To our knowledge, some providers use predominantly one engine and customise it depending on the recognition task or available resources, while others offer access to different engines, selecting the most suitable one for each task based on internal criteria. However, engine development and feature customisation tend to be treated as the company’s internal know-how, contributing to their competitive advantage.

[3] A step in this direction is the FAR model (Pedersen, 2017) which identifies functional equivalence, acceptability, and readability errors in *interlingual* subtitling and builds on the NER model (see Ludera et al., 2024 for a recent application to interlingual subtitling).

errors (E) and machine-generated recognition errors (R), deducted from the total number of respoken words (N). Each error is assigned a score reflecting its potential impact on viewers' comprehension. *Minor* errors (-0.25) do not make the subtitle harder to understand, while *standard* errors (-0.50) lead to confusion or loss of information, and *serious* errors (-1) convey false or misleading information. Additionally, the model recognises instances of correct editions (CE), where strategic changes are made without compromising clarity or accuracy. While CEs are not assigned numerical scores, they highlight the strengths of human editing. Several studies have validated the NER model in professional settings (e.g., Ofcom, 2015a, 2015b), and a minimum threshold of 98% has been set as the acceptable accuracy standard. In training, the NER model is a useful diagnostic tool to identify recurrent errors in performance.

The NTR model is an evolution of the NER model designed to evaluate live interlingual subtitles (Figure 2). Edition errors are replaced by Translation errors (T), covering content-related (Tc) and form-related (Tf) issues. Content-related errors include omissions, additions, and substitutions, while form-related errors involve grammatical correctness and style. The scoring system is similar to the NER model, with minor errors, major errors (same as standard) and critical errors (same as serious). While content errors span the full scale (minor, standard, serious), form errors are limited to minor and standard.

As this study analyses offline subtitles in a fully automated workflow, the two above models were merged and adapted. While all errors in this workflow are ASR-generated recognition (R) errors, it is important to distinguish different types of recognition errors. We based the categorisation on the NTR model. The analysis thus highlights the key challenges posed by misrecognition and facilitates the identification of editing requirements. Our proposed accuracy model is shown in Figure 3.

Figure 3

Adapted Formula for Accuracy Evaluation in the Study

$$\text{Accuracy} = \frac{N - R(Tc) - R(Tf)}{N} \times 100 = \%$$

N = total number of ASR-produced words in the target
R(Tc) = content-related recognition error (e.g. omission, addition, substitution)
R(Tf) = form-related recognition error (e.g. correctness, style)

Our approach considers ASR-generated subtitles as an intermediate step requiring further refinement for broadcast readiness. Although post-editing efforts cannot be quantified in this study (owing to lack of information on software deployment and procedures), it is possible to assess how close the ASR-generated subtitles came to the 98% accuracy threshold. The evaluation grid adapted by Davitti and Sandrelli (2020) from the Canadian NER score spreadsheet was used in the analysis. This tool has been used extensively in other projects, e.g., the ESRC-funded SMART project (*Shaping Multilingual Access through Respeaking Technology*, ES/T002530/1, 2020–2023), and ensured consistency across different evaluators, facilitating a comparison between ASR tools and enabling automated calculations.

However, the model in Fig. 3 does not account for more “technical” aspects of subtitling, like segmentation and timing. As Bolaños García-Escribano and Declercq (2023, p. 571) put it:

> The line breaking within subtitles and the segmentation across subtitles ought to be done according to syntactic and grammatical considerations rather than aesthetic rules, since the ultimate objective is to facilitate the reading and understanding of the message in the little time available.

Thus, a separate analysis of the ASR output was focused on issues that impact readability. Here “readability” refers to how the text is presented in the subtitles, namely whether the line breaks and subtitle breaks contribute to “ease of comprehension and coherence between individual subtitles” (Díaz Cintas & Remael, 2021, p. 142). The timecoded ASR files were examined on two levels: subtitle breaks (i.e., text distribution across different subtitles) and line breaks (i.e., text distribution within the same subtitle), as driven by the language models of the tools and by speaker delivery, pauses, emphasis, intonation, speaker changes, etc. The subtitles were examined to assess compliance with current subtitling standards and to identify all those instances requiring correction for improved readability.

In relation to line breaks, we highlighted all those cases where some words, according to the subtitling style guidelines provided by the broadcaster, should be moved to a different line within the same subtitle (e.g. when the first line ends with an auxiliary verb and the second line begins with a past participle or when a subject is split from its verb). In relation to subtitle breaks, we identified all those cases in which one or more words need to be moved to a different subtitle (e.g. when a subtitle ends with a conjunction introducing a subordinate clause in the following subtitle). While text segmentation does involve some subjective decisions, we identified all those instances where segmentation is unequivocally incorrect and requires human intervention to move text around and adjust timecodes, thus increasing overall post-editing time. In this sense, a subtitle may be deemed correct from the point of view of segmentation, even when it contains recognition errors.

In addition, although the main focus of our analysis was not on subtitle timings, the available time codes made it possible to check compliance with the specific guidelines provided by the broadcaster, based on the recommended rate of 160–180 words per minute. The recommended subtitle duration ranged between 1 and 6 seconds, and wherever possible, the minimum display time should be 2 seconds; whenever the subtitles are on screen for less than 1 second, it is recommended that the subtitle in question be merged with another subtitle, brought it earlier or left on the screen for longer.

Our findings are presented by language, starting with English (§2), and then followed by Italian (§3). The ensuing discussion (§4) adopts a comparative approach, juxtaposing the key issues identified in the performance of the two ASR tools in each language.

2. Analysis of the English Data

This section presents the main findings from the analysis of the English dataset, starting with accuracy (§2.1) followed by segmentation (§2.2). We identify prototypical errors and correct editions and offer qualitative illustrative examples highlighting key trends in each output.

2.1. English Subtitle Accuracy

This section presents accuracy scores, error types, frequency (number of occurrences) and degree of severity (score deductions) for the English subtitle files produced by the two ASR systems for the TV show "Last Week with John Oliver". Tables 2 and 3 provide the number of occurrences of each error type and the deductions (i.e., weighted occurrences) used to calculate the accuracy score. The error categories most responsible for affecting the final score are highlighted in red. The second part of the section shows the main error categories.

Table 2

Frequency by Error Type and Severity for Broadstream – English

Content	**Omissions**		**Additions**		**Substitutions**	
	Occurrences	Deductions	Occurrences	Deductions	Occurrences	Deductions
Minor	2	-0.5	1	-0.25	3	-0.75
Standard	4	-2	0	0	18	-9
Serious	0	0	0	0	10	-10
TOT	6	-2.5	1	-0.25	31	-19.75

Form	**Correctness**		**Style**		**Correct editions**	
	Occurrences	Deductions	Occurrences	Deductions	Occurrences	Deductions
Minor	54	-13.5	1	-0.25	N/A	N/A
Standard	10	-5	0	0		
TOT	64	-18.5	1	-0.25	4	0

	Occurrences	**Deductions**
Overall total	107	-41.25
Overall accuracy score: 97.97%		

Table 3

Frequency by Error Type and Severity for Amberscript – English

Content	**Omissions**		**Additions**		**Substitutions**	
	Occurrences	Deductions	Occurrences	Deductions	Occurrences	Deductions
Minor	16	-4	2	-0.5	9	-2.25
Standard	13	-6.5	1	-0.5	35	-17.5
Serious	1	-1	0	0	16	-16
TOT	30	-11.5	3	-1	60	-35.75

Form	**Correctness**		**Style**		**Correct editions**	
	Occurrences	Deductions	Occurrences	Deductions	Occurrences	Deductions
Minor	72	-18	2	-0.5	N/A	N/A
Standard	9	-4.5	0	0		
TOT	81	-22.5	2	-0.5	8	0

	Occurrences	**Deductions**
Overall total	184	-71.25
Overall accuracy score: 96.32%		

Broadstream’s output achieved an accuracy score of 97.97%, just below the 98% suggested by Romero-Fresco (2016) as the acceptable accuracy threshold. By contrast, with a score of 96.32%, Amberscript’s output falls considerably below the threshold, lagging 1.65% behind Broadstream. In both cases, recognition issues have resulted in similar error patterns (highlighted in red in Table 2). Content-related substitutions accounted for about half of all deduction points in both. Form-related correctness errors follow closely, with similar levels in both, and omissions, more frequent in the subtitles generated by Amberscript than Broadstream.

Substitution errors showed a notable disparity, with nearly twice as many in the Amberscript output than in the Broadstream file (60 vs 31). However, both ASR systems spanned the entire severity scale with comparable distribution patterns. In serious substitutions, the seemingly plausible output conveys inaccurate or misleading information: they account for approximately 30% of cases in the output of both ASR tools and generally affect nouns, adjectives, verbs, and conjunctions. Examples 1–4 present a selection of substitutions involving individual parts of speech such as nouns, proper names and adverbs (Examples 1–3), but also more elaborate transformations like in Example 4, where the verb form in the source (“I know”) is turned into a negative (“I don’t know”; Broadstream) or the pronoun “what” is changed into “why” (Amberscript).

Example 1

Source	they avoided having Camilla use (.) this crown during the coronation because it contains the **Koh-i-Noor** diamond
BROADSTREAM	[…] the **Cold War** diamond
AMBERSCRIPT	[…] the **Conor** diamond

Example 2

Source	yes the (.) coronation of the world's least likeable **orphan** is less than a week away now
BROADSTREAM	Yes, the coronation of the world's least likeable **author** is less than a week away now.

Example 3

Source	[…] to make sure that we're all (.) as **environmentally** sensible as we possibly can be
AMBESRCRIPT	[…] to make sure we're all **mentally** sensible as we possibly can be.

Example 4

Source	and **I know what** you're thinking
BROADSTREAM	I **don't know** what you're thinking.
AMBERSCRIPT	And I know **why** you're thinking

Standard substitutions accounted for over half of all cases (18 out of 31 in Broadstream and 35 out of 60 in Amberscript). Examples 5–6 show some key ASR "struggle" areas, i.e., articles and pronouns that are crucial for sentence coherence, and similar-sounding words which lead to contextually inappropriate substitutions that may confuse viewers.

Example 5

Source	I mean (.) I do kind of get where **he's** coming from **there**
BROADSTREAM	I mean, I do kind of get where he's coming from **that.**
AMBERSCRIPT	I mean I do kind of get where **it's** coming from there.

Example 6

Source	but I guess **a former** Brexit leader wouldn't know anything about that
BROADSTREAM	But I guess **a form of** Brexit leader wouldn't know anything about that.
AMBERSCRIPT	But I guess **a form a** Brexit leader wouldn't know anything about that.

In Example 7, the shift from the definite to the indefinite article does not hinder understanding, but it may create some confusion as to which coronation is being referred to.

Example 7

Source	**the** coronation is happening on Saturday
BROADSTREAM	**A** coronation is happening on Saturday.

Example 8 shows the misrecognition of similar-sounding words (“I know”/ “a lot”). Notably, a segmentation issue caused the misrecognised item to be incorporated within the previous subtitle. This layered combination of errors is prevalent in the dataset.

Example 8

Source	you don’t want him there (.) **I know** this is clearly a big moment in British history
BROADSTREAM	You don’t want him there **a lot.** This is clearly a big moment in British history,

The second most frequent category of errors in the output of both ASR systems is correctness errors, predominantly encompassing capitalisation (see Examples 9–10) and punctuation issues (e.g., adding unnecessary full stops, replacing commas with full stops, and missing quotation marks). The latter may arise from the unnatural prosody of the original, linked to the fast and scripted nature of the source (§1.2) and leading to ASR errors, with repercussions on meaning and segmentation – Examples 11–12). A recurring pattern was the presence of multiple problems within the same subtitle. However, most of such errors (54 out of 64 in Broadstream and 72 out of 81 in Amberscript) were classified as minor, as they would not produce major misunderstandings.

Example 9

Source	Buckingham Palace is releasing these three brand new images of the **soon-to-be-crowned King and Queen**
BROADSTREAM	Buckingham Palace is releasing these three brand new images of the **soon to be crowned king and queen.**

Example 10

Source	the **King and Queen Consort** have unveiled their coronation quiche
AMBERSCRIPT	The **king and queen-consort** have unveiled their coronation quiche.

Example 11

Source	that event was and this is true (.) five and a half hours long
BROADSTREAM	That event **was. And** this is **true. Fiv**e and a half hours long.

Example 12

Source	because I I I believe that the recipe for your particular secret sauce is is under the hood of what you **do well** not what you don't
AMBERSCRIPT	because I believe that the recipe for your particular secret sauce is is under the hood of what you **do. Well,** not what you don't.

By contrast, a similar proportion of errors (10 in Broadstream and 9 in Amberscript) were categorised as standard, as they could lead to significant shifts in the target output – for example, the transformation of a question mark into a full stop changes an interrogative sentence into a declarative one. Correctness errors are often combined with recognition errors, leading to unintelligible output and, consequently, increased editing time. In Example 13, segmentation goes wrong due to the insertion of a full stop in the wrong place after “in his autobiography” (which was the beginning of the following unit in the source), thus producing a change in meaning. The other correctness error in this example is minor (i.e., the lack of quotation marks in the fragment cited from the book). Still, it does mean that the subtitle is not broadcast-ready.

Example 13

Source	Elton John was famously close friends with Diana (.) in his autobiography he even called her “incredibly indiscreet, a real gossip”
AMBERSCRIPT	Elton John was famously close friends with **Diana in his autobiography**. **He even** called her **incredibly indiscreet, a real gossip**.

In Example 14, a question mark is turned into a full stop, thus changing a question into a declarative statement.

Example 14

Source	sorry to interrupt but your heart sank when you saw that news**?**
BROADSTREAM	Sorry to interrupt, but your heart sank when you saw that news.

More complex instances can be found in Examples 15–16, where multiple correctness errors are present at the same time. In Example 15, the host’s sarcastic remark about a quiche being “a pie’s weird camp friend” is misconstrued in the target output. Multiple recognition errors result in incorrect punctuation and substitutions, and the final output is nearly incomprehensible.

Example 15

Source	it’s pie’s weird camp friend and on (.) quiche’s best day it’s nothing
BROADSTREAM	It’s pies. Weird. Come, friends and on pieces. Best day. It’s nothing.

In Example 16, the ASR output contains many lines featuring misplaced or incorrect punctuation (missing comma after “still”, missing full stop after “TV”, redundant comma after “have made,”). Furthermore, capitalisation is unnecessarily applied in “Quiche” and lacking in “british”.

Example 16

Source	still it has taken up a truly stupid amount of time on British TV chefs have made the quiche, hosts have discussed it
AMBERSCRIPT	Still it has taken up a truly stupid amount of time on british TV Chefs have made, the Quiche, hosts have discussed it

In this monologic genre, the two ASR tools captured most of the content despite the speed. Broadstream only featured 6 instances of omissions, while Amberscript produced 30 omissions encompassing all levels of severity. Both tools omitted three full segments at the beginning of the show, probably due to background music/applause/noise. Furthermore, the particularly fast delivery of the source often led to word contraction and poor articulation, resulting in dropped content. Example 17 shows the drop of a modal verb ("might"), which produced a sentence that is grammatically correct but lacking a nuance. Example 18 shows how the omission of "heard" combined with punctuation issues (addition of a full stop) led to a distorted segment.

Example 17

Source	I know my dreams **might have contained** some pretty dicey racial imagery there but (.) would it help to know that at the end of it (.) I ejaculated
BROADSTREAM	I know my dreams **contain** some pretty dicey racial imagery there, but would it help to know that at the end of it, I ejaculated.

Example 18

Source	that is the face of a man listening to the best gossip about her husband that he has **ever heard** while still trying to appear sympathetic
AMBERSCRIPT	That is the face of a man listening to the best gossip about her husband that he has ever ***[no rendition]*. We're** still trying to appear sympathetic.

In addition to errors, some instances of positive editing were performed by the ASR tools. Specifically, the tools removed redundant or non-essential sentence openers resulting from the speaker's stuttering or hesitations in four lines, such as "and", "but", "well", etc. Examples 19–20 show how the output of an ASR system can include both successful and erroneous practices within a single idea unit and highlight that positive interventions in the dataset are limited to surface-level issues, with no deeper strategic reformulation moves.

Example 19

Source	**and** I'm not sure what lesson you're supposed to learn from it other than if you try to do drugs with Matthew McConaughey you will die
BROADSTREAM	I'm not sure what lesson you're supposed to learn from it, other than if you try to do drugs with Matthew McConaughey, you will die.

Example 20

Source	**I know** this is clearly a big moment in British history
AMBERSCRIPT	[no rendition] This is clearly a big moment in british history,

2.2. English Subtitle Segmentation

Segmentation in line with subtitling conventions remains a challenge for ASR tools, and Broadstream and Amberscript are no exception. Table 4 shows the proportion of subtitles with correct segmentation to those with incorrect segmentation in both outputs. The Broadstream file consists of 256 subtitles, of which only 68 are unaffected by any segmentation issues. The respective figures for the Amberscript file are 232 and 41. The remaining 188 subtitles for Broadstream and 191 subtitles for Amberscript included instances of bad line breaks and/or bad subtitle breaks, often accompanied by punctuation issues. Although Broadstream performed slightly better than Amberscript in this respect (+7%), considerable effort would be needed to turn both ASR outputs into broadcast-ready subtitles.

Table 4

Breakdown of Segmentation Issues Identified in Both ASR Solutions for English

ASR provider	Subtitles with correct segmentation / total subtitles	Percentage of subtitles with correct segmentation	Percentage of subtitles requiring editing
Broadstream EN	68 / 256	25%	75%
Amberscript EN	41 / 232	18%	82%

A pattern that was encountered in both ASR outputs is the production of shorter-than-needed lines in the absence of specific space constraints. The Amberscript output has more one-liners than the Broadstream one (45% vs 34%, respectively). Usually, they do not contain a full idea unit. Moreover, in many cases, the timing created by the tools is not in line with the broadcaster’s recommendation (i.e. there are many cases of full one-liners on the screen for less than one second – see §1.2). This suggests the subtitles do not stay on the screen long enough for viewers to read them. In several cases, this has a cascade effect on subsequent subtitles, with instances of two or three subtitles that should have been merged (in terms of content and timing) being kept separate. The Broadstream output contains more two-liners but is nevertheless affected by poor line breaks and subtitle breaks, often in the same subtitles. Amberscript also produced some three-liners, which account for approximately 4.7% of the overall number of subtitles. This is not standard practice on TV and goes against the broadcaster’s subtitling guidelines. The presence of three-liners may have been caused by the incorrect setting of the ASR engine, but it is impossible to tell with any certainty without access to the engine itself. Considerable editing effort would be needed to re-distribute the text more sensibly.

One of the key issues in both ASR outputs is that the subtitles are frequently segmented in a seemingly random manner, at non-natural breaks, often splitting verb phrases, noun phrases, prepositional phrases as well as subjects and their verbs. Our assumption is that both ASR tools were deployed off-the-shelf with default settings. The following examples are presented in a comparative manner, with the Broadstream output on the left and the Amberscript output on the right-hand side of each column, to show how the different ASR systems dealt with the same chunks. While it is difficult to quantify segmentation errors, these examples show how intertwined they are with accuracy errors. Punctuation has been found to magnify segmentation problems and add to the complexity of the editing task.

Example 21 shows the splitting of a noun phrase ("pretty dicey racial imagery") and a prepositional phrase ("at the end of it") both within the same subtitle and across subtitles, as well as other issues in relation to breaking up verbs ("help to know") and subject-verb units ("I ejaculated"). Line segmentation and subtitle re-arrangement are necessary for improved readability.

Example 21

Broadstream	**Amberscript**
19 00:01:18,040 → 00:01:20,600 I know my dreams contain some pretty dicey racial	15 00:01:16,120 → 00:01:18,880 Title: I know my dream.
20 00:01:20,720 → 00:01:22,920 imagery there, but would it help to know that at the	16 00:01:19,040 → 00:01:22,400 It contains some pretty dicey racial imagery there, but would it help to
21 00:01:23,040 → 00:01:24,840 end of it, I ejaculated.	17 00:01:22,480 → 00:01:24,800 know that at the end of it I ejaculated?

In Example 22, the Broadstream output, "we're going to start in the UK", could have been presented as one line, as it would not exceed the limit of 37 characters per line set by the broadcaster; the same goes for "recently said". Similarly, in the Amberscript subtitles, modifiers followed by nouns ("chief economist") and the auxiliary verb followed by a gerund ("be happening") should have been kept together. In terms of subtitle breaks, verb phrases ("we're all worse off"), subject and verb ("economy is") and the comparative adverbial phrase ("less than") should have been in the same subtitle.

Example 22

Broadstream	**Amberscript**
28 00:01:38,560 → 00:01:40,640 We're going to start in the UK, where the economy	22 00:01:38,600 → 00:01:41,560 We're going to start in the UK, where the economy is in turmoil,
29 00:01:40,760 → 00:01:42,640 is in turmoil and the Bank of England's chief	23 00:01:41,680 → 00:01:45,000 and the Bank of England's chief economist recently said that we're all worse
30 00:01:42,720 → 00:01:45,320 economist recently said that we're all worse off	24 00:01:45,120 → 00:01:48,400 off and we all have to take our share, which makes it less

31 00:01:45,400 → 00:01:47,920 and we all have to take our share, which makes it 32 00:01:48,000 → 00:01:50,600 a less than ideal time for this to be happening.	25 00:01:48,480 → 00:01:50,560 than ideal time for this to be happening.

Example 23 shows that poor segmentation is compounded by poor accuracy in both ASR outputs, resulting in unintelligible subtitles. Segments 148–149 in Broadstream and 134–135 in Amberscript should have read, "But this is what mustered emotion out of you?"

Example 23

Broadstream	**Amberscript**
146 00:06:02,560 → 00:06:04,760 Every day the earth gets one degree closer to death. 147 00:06:04,880 → 00:06:06,960 And it looks like Carrie's getting back together with Aiden. 148 00:06:07,040 → 00:06:08,080 But this 149 00:06:08,200 → 00:06:09,640 is what must **be the most** **out of you**.	131 00:06:02,600 → 00:06:04,760 every day, the earth gets one degree closer to death 132 00:06:04,880 → 00:06:05,320 and it looks like 133 00:06:05,400 → 00:06:07,000 Carry's getting back together with Agent. 134 00:06:07,080 → 00:06:08,840 But this is what must 135 00:06:08,920 → 00:06:09,920 **Elton? Also?**

Finally, Example 24 showcases another instance of poor segmentation by both tools, further compounded by poor accuracy – the segment should have read:

> and look (.) at the end of the day (.) the reason everyone's desperately talking about Charles wanting people to have bean quiche at his fake job party is that it is clearly more interesting than he is (.) which is a real problem for the monarchy.

Example 24

Broadstream	Amberscript
176 00:07:07,840 → 00:07:10,160 Look, at the end of the day, the reason everyone's 177 00:07:10,280 → 00:07:12,040 desperately talking about Charles wanting people 178 00:07:12,120 → 00:07:15,120 to **have been Keisha this** fake job party is that it is 179 00:07:15,200 → 00:07:18,360 clearly more interesting **than his**, which is a real 180 00:07:18,440 → 00:07:19,560 problem for the monarchy.	164 00:07:07,840 → 00:07:11,000 Look, at the end of the day, the reason everyone's desperately talking 165 00:07:11,080 → 00:07:14,360 about Charles wanting people to **have been shitty** fake job party 166 00:07:14,440 → 00:07:17,160 is that it is clearly more interesting than **his**, 167 00:07:17,240 → 00:07:19,520 which is a real problem for the monarchy.

To sum up, both outputs have a comparable percentage of subtitles requiring editing, as the ASR tools tended to distribute text within and across the subtitles in ways that do not follow natural linguistic breaks, thus potentially hampering reading. This is compounded by accuracy-related problems like substitutions and punctuation errors. The fast pace of speech in talk shows requires careful consideration when it comes to segmentation, as it is important to bear in mind the viewers' average reading speed. Additionally, humorous and sarcastic remarks, common in shows of this kind, may need to be flagged up with (!) or (?) or capitalised for emphasis to convey key characteristics of the genre.

3. Analysis of the Italian Data

This section presents our analysis of the automatic subtitles generated for the Italian material. The Italian video clip is entirely dialogic and there are frequent voice overlaps, as well as children's voices, which may be harder to transcribe for the ASR tools.

3.1. Italian Subtitle Accuracy

Our accuracy evaluation shows that both ASR systems achieved accuracy scores that are far below the suggested 98% threshold. There was, however, a considerable difference in performance. While

Amberscript performed better than Broadstream (93.55% vs 89.61%, respectively), in absolute terms, both results must be considered poor, especially when taken together with the numerous timing and text segmentation problems identified in both sets of subtitles (see §3.2).

A similar number of recognition problems was identified in the two files, 232 in the Broadstream output vs 220 in the Amberscript subtitles. The two tools differ in the distribution of error types and severity, as shown in Tables 5 and 6 (as for the English dataset, the error types predominantly impacting the overall score are emphasised in red).

Table 5

Frequency by Error Type and Severity for Broadstream – Italian

Content	**Omissions**		**Additions**		**Substitutions**	
	Occurrences	Deduction	Occurrences	Deduction	Occurrences	Deduction
Minor	34	-8.50	6	-1.50	6	-1.50
Standard	20	-10	2	-1	26	-13
Serious	18	-18	1	-1	43	-43
TOT	72	-36.50	9	-3.50	75	-57.50

Form	**Correctness**		**Style**		**Correct editions**	
	Occurrences	Deduction	Occurrences	Deduction	Occurrences	Deduction
Minor	39	-9.75	1	-0.25	N/A	N/A
Standard	36	-18	0	0		
TOT	75	-27.75	1	-0.25	0	0

	Occurrences	**Deduction**
Overall total	232	-125.50
Overall NER: 89.61%		

Table 6

Frequency by Error Type and Severity for Amberscript – Italian

Content	**Omissions**		**Additions**		**Substitutions**	
	Occurrences	Deduction	Occurrences	Deduction	Occurrences	Deduction
Minor	27	-6.75	43	-10.75	12	-3
Standard	9	-4.50	3	-1.50	19	-9.50
Serious	7	-7	0	0	17	-17
TOT	43	-18.25	46	-12.25	48	-29.50

Form	**Correctness**		**Style**		**Correct editions**	
	Occurrences	Deduction	Occurrences	Deduction	Occurrences	Deduction
Minor	46	-11.50	1	-0.25	N/A	N/A
Standard	32	-16	4	-2		
TOT	78	-27.50	5	-2.25	0	0

	Occurrences	**Deduction**
Overall total	220	-89.75
Overall NER: 93.55%		

The predominant error type in both sets of subtitles is substitutions. However, Broadstream performed much worse than Amberscript, not only in terms of overall deduction points in this category (almost double, i.e., -57.50 vs -29.50, respectively) but also in terms of error distribution. There were 43 serious substitutions in the Broadstream subtitles, which is almost the same as the total number of substitutions in the Amberscript version (48). Similarly, the impact of omissions in the Broadstream subtitles was twice as big than in the Amberscript output (-36.50 vs -18.25, respectively), once again with a much higher number of serious omissions (18 vs 7). It can be argued, therefore, that the Broadstream subtitles conveyed less content than the Amberscript ones.

As regards correctness errors, the subtitles produced by the two tools were affected almost equally (-27.75 vs -27.50 in the Broadstream and Amberscript outputs, respectively), mostly by punctuation issues (as in the English dataset). Finally, in the Amberscript output there were also 46 additions (-12.25 deduction points): almost all of them (43) were minor and involved the addition of the letter “f” at the beginning of the subtitles in question. This may have been caused by an incorrect setting in the ASR tool and would need to be fixed for broadcasting purposes, but it does not hinder comprehension.

In the Broadstream subtitles, almost 60% of substitution errors were of a serious nature, compared to about 35% in the Amberscript output. Example 25 is taken from a scene in which a customer is followed by a shop assistant outside the supermarket, and the two argue in the street. Fast overlapping dialogue and characters screaming at each other presented a challenge and neither version is intelligible. In the Broadstream subtitles, it is difficult to tell with any certainty which word(s) of the original were misrecognised as what in the Italian subtitles because the target version bears no resemblance to the original and is entirely nonsensical. In the Amberscript version, the imperative verb "si levi" ("get off") was misrecognised as "se avevi" ("if you had"), and "si allontani" (get away from me) was misrecognised as "a segno le mani" (the idiom "mettere a segno" means to score, but "mettere le mani" means to raise your hand to someone). Moreover, both tools failed to transcribe the shop assistant's line (character B), thus turning this dialogue into a monologue. Although in the NER model, the omission of a full independent idea unit is generally classified as a standard omission, in this case, it was considered a serious omission because the viewers see two characters talking but only get access to one character's words (the customer).

Example 25

Source
A: mi ha fatto male alla gamba (.) si levi!
B: insomma (.) signore!
A: si allontani! Sono un cittadino anziano! Se ne vada!

BT (backtranslation)
A: you hurt my leg. Get off!
B: really, Sir!
A: get away from me! I am a senior citizen! Go away!

Broadstream	Amberscript
Giornata sarà segnalata una segnalazione sull'uscita dell'anziano	poi. Mi ha fatto male alla gamba, se avevi messo a segno le mani. Sono un cittadino anziano se ne vada.
BT A: Day. will be B: // A: Reported a report on the elderly person's exit //	BT A: Then. He hurt my leg, if you had put A: your hands on the score. I am a senior citizen go away.

Standard substitutions (26 for Broadstream and 19 for Amberscript) are cases in which the misrecognised word results in a confusing subtitle. Examples 26 and 27 show that both tools struggled with short words such as interjections, articles, and pronouns. In Example 26, the interjection "oops" was misrecognised by Broadstream ASR as "Box". Presumably, the viewers would realise that this is a recognition error but would struggle to deduce the intended meaning.

Example 26

Source
eh!
oops! (.) che peccato! Eh!

BT
ah! woops! what a pity! Ah!

Broadstream

Box.

Che peccato.

BT
Box. What a pity.

In Example 27, the exclamation “OK” was misrecognised by Amberscript ASR as the pronoun “lei” (“her” or the politeness pronoun used to refer to “you”), whose reference is unclear in the context. Moreover, the verb “sapete” (“you can”, plural) was misrecognised as “sapeva” (“he/she could do”), causing further confusion.

Example 27

Source
oh! Ok
è il massimo che sapete fare?

BT
is that the best that you can do?

Amberscript

sotto. Oh, lei. È

il massimo che
sapeva fare. Oh no,

BT
Oh, her. Is it
the best that
she could do. Oh no,

In Example 28, both ASR systems struggled with the transcription of a robotic voice instructing a customer to scan his shopping items at an automatic till in the supermarket. Broadstream omitted the verb “scansionare” (“to scan”) every single time, while the other words (“l’articolo, prego”) were misrecognised in various ways (“piccolo, di quello preso”). Similarly, Amberscript transcribed the phrase “scansionare l’articolo, prego” as “suonare l’articolo, grazie dell’articolo”, and “Nazionale l’articolo”. These errors were all classified as standard because viewers can tell there is something wrong but can only retrieve the intended meaning via the images.

Example 28

Source
SCANSIONARE L'ARTICOLO PREGO
A: sto parlando con un robot
SCANSIONARE L'ARTICOLO PREGO
A: l'ho già fatto!
vediamo se questo coso qui

SCANSIONARE L'ARTICOLO PREGO
A: insopportabile.
SCANSIONARE L'ARTICOLO PREGO

BT
B: scan the item please
A: I am speaking to a robot
B: scan the item please
A: I've already done it!
Let's see if this thing here
B: scan the item please
A: unbearable.
B: scan the item please

Broadstream	Amberscript
Prego.	suonare l'articolo, prego.
Sto parlando con un robot	fSto parlando con un robot.
	-Grazie dell'articolo.
piccolo.	
	L'ho già fatto.
Ho già fatto.	Vediamo se questo coso.
Vediamo se questo caso più	fNazionale l'articolo.
	-Prego insopportabile.
insopportabile.	
	fNon so bene l'articolo prego.
Di quello preso.	
BT	BT
B: Please	B: play the item, please.
A: I'm talking to a robot.	A: fI'm talking to a robot.
B: small	B: thank you for the item
A: I have already done.	A: I have already done it.
Let's see if this case more	Let's see if this thing.
B: //	B: fnational the item.
A: Unbearable.	A: please unbearable.
B: than the one (that was) taken	B: fI don't know well the item please.

As regards omissions, it is worth pointing out that while the original Italian dialogue was 1,081 words long, Amberscript produced a shorter output than Broadstream (984 vs 1,073 words, respectively – see Table 1). Therefore, a higher number of omissions can be expected in the Broadstream output. Although most omissions in both versions were minor or standard, there were also 18 cases of serious omissions in the Broadstream output and 7 such cases in the Amberscript subs. A frequent case was the omission of entire lines of dialogue during fast exchanges, with the subtitles conveying the words of only one character: this is the mechanism we saw at work in Example 25. As was mentioned before, the omission of an entire idea unit is usually considered of standard severity, but when it occurs in dialogue, it can significantly impact the development of the exchange. For example, such omissions were often accompanied by text segmentation and punctuation issues, resulting in speech being attributed to the wrong character. In Example 29, Grandpa (character A) is discussing moving in with his daughter. There is a cut to a new scene in which his grandson is arguing with his parents, as he does not want to give up his bedroom. Here, Broadstream failed to recognise "per me", from the first line of dialogue uttered by A. Then it failed to detect the scene change and, as a result, the lines uttered by A and B were merged into one ("vorresti mettermi in soffitta"). The subtitles are confusing, as they read as if Grandpa is supposed to be moving into the attic, not the child. By contrast, in the Amberscript subtitles, the transcript is more accurate, with no omissions but the tool was unable to insert dashes in the correct place for speaker identification.

Example 29

Source
A: comunque tu non hai una camera per me (.) dove vorresti mettermi?
B: in soffitta? No (.) mai
C: avanti tesoro

BT
A: anyway, you don't have a spare room for me. Where would you put me?
B: in the attic? No, never!
C: come on, darling

Broadstream	Amberscript
Comunque tu non hai una camera dove	-Sì, può darsi. Comunque tu non
vorresti mettermi in soffitta?	hai una camera per me. Dove vorresti mettermi
No, mai.	fin soffitta? -No, mai. No, anzi, tesoro,
BT A: Anyway, you don't have a room where	BT A: Yes, maybe. Anyway, you don't
you would put me in the attic?	have a room for me. Where would you put me
No, never.	Up to the attic? B: -No, never. No, actually, darling,

The third largest category of errors in both sets of subtitles are correctness errors, almost all of them punctuation issues. Missing full stops or commas or the addition of unnecessary punctuation marks were frequent; these were considered minor errors. However, the transformation of statements into questions and vice versa was considered an error of higher severity because it changed the function of the sentence (Example 30).

Example 30

Source
A: a nessuno va peggio che a me
vedete quella ragazza laggiù?

B: oh sì

A: ogni volta che può mi tortura

B: aspetta (.) ma non è tua sorella?

BT
A: no one has it worse than me
do you see that girl over there?
B: oh yes,
A: whenever she can she tortures me
B: wait. but isn't that your sister?

Broadstream	Amberscript
a nessuno va peggio che a me. Vedete quella ragazza laggiù?	A nessuno va peggio che a me. Vedete quella
Sì.	fragazza. -Laggiù? Oh, sì, ogni volta.
Ogni volta che può mi tortura.	
	fChe può mi tortura.
Aspetta, ma non è tua sorella.	-Aspetta, ma non è tua sorella.
BT no one has it worse than me. Do you see that girl over there?	BT No one has it worse than me. You see that
Yes.	fgirl. -Over there? Oh, yes, whenever.
Whenever she can, she tortures me.	
	fshe can she tortures me
Wait, but that is not your sister.	-Wait, but that is not your sister.

This short overview of our accuracy analysis shows that the Italian automatic subtitles generated by both tools would require substantial human editing to reach broadcasting quality, with Amberscript performing marginally better.

3.2. Italian Subtitle Segmentation

As the film sequence was entirely made up of dialogue, poor segmentation had a bigger impact in comparison with the subtitling of monologic material, because it makes it difficult for viewers to understand who is saying what to whom. Moreover, recognition errors and poor segmentation often went hand in hand.

The Broadstream file had 226 subtitles, while the Amberscript file had only 140 (Table 7). This difference indicates that the two tools segmented their output differently. There is a higher number of two-liners in the Amberscript transcript, while most subtitles in the Broadstream file are made up of just one line. Although it is not necessarily wrong to use one-liners when subtitling dialogue it is frequent to use two-line subtitles, with one line per character. The Broadstream tool chunked the text far too much, with many one-line subtitles not containing a full unit of meaning and often being displayed on the screen for barely a second or less than a second, thus flouting the broadcaster's guidelines. In this version, text segmentation problems affect not just individual subtitles but often two or three consecutive ones, which should have been merged both logically (content-wise) and in terms of timing. Broadstream segmented only 36 subtitles correctly (about 16% of the total), while the remaining 190 subtitles included bad line breaks and/or bad subtitle breaks. Bad segmentation was often compounded by punctuation problems. Amberscript performed marginally better, with 22% of subtitles not requiring corrections.

Table 7

Breakdown of Segmentation Issues Identified in Both ASR Solutions for Italian

ASR provider	**Subtitles with correct segmentation / total subtitles**	**Percentage of subtitles with correct segmentation**	**Percentage of subtitles requiring editing**
Broadstream IT	36 / 226	16%	84%
Amberscript IT	31 / 140	22%	78%

Text segmentation is especially hard when dialogue involves more than two characters. In Example 31, a group of 4 children is discussing the beginning of the school year. Broadstream recognised that several voices were involved and gave one line to each character; however, some subtitles are very short and barely flicker on the screen for a second or less. Research suggests that some viewers might not be able to read so fast (e.g. Kruger et al., 2022). By contrast, Amberscript opted for a two-line subtitle followed by a one-line subtitle, with no clear indication of who says what. Moreover, the two-liner ends with an object pronoun ("li", meaning "them", as moustache is plural in Italian) and the verb ("vedo", i.e., see) is placed in the following subtitle: this breaks the viewers' reading flow.

Example 31

Source

```
A: a me andrà alla grande (.) parla per te (.) mi stanno crescendo i baffetti
B: non li vedo
C: no
D: eh (.) no
```

BT

A: it will be great for me, speak for yourself. I am growing a moustache
B: I can’t see it
C: no
D: ehm, no

Broadstream	Amberscript
00:01:44,210 → 00:01:46,100 A me andrà alla grande perché	0013 00:01:43:05 → 00:01:46:12 Questo è vero. A me andrà alla grande. Parla per te.
00:01:46,610 → 00:01:47,870 mi stanno facendo dibattiti,	0014 00:01:46:15 → 00:01:49:02 Mi stanno crescendo i baffetti. Non li
00:01:48,860 → 00:01:49,820 non li vedo.	0015 00:01:49:04 → 00:01:52:06 vedo. No. Beh, no.
00:01:49,920 → 00:01:51,730 No,	
00:01:51,830 → 00:01:52,600 no.	
BT It will be great for me because they are making me debates, I can’t see them, No, no.	BT This is true. For me it will be great. Speak for yourself. I am growing a moustache. I don’t them see. No. Well, no.

Our analysis indicates that in many cases, line and subtitle breaks were inserted without due consideration for either grammatical or prosodic boundaries. In Example 32, both tools did relatively well in terms of accuracy, and yet neither version is easy to read because of poor segmentation. In the Broadstream version, the first line of the first subtitle ends with a preposition (“con”), and the noun it is supposed to accompany (“idea”) appears in the second subtitle. Amberscript chopped the sentence in a different, but equally wrong, place: the possessive “la tua” (your) is split from the noun and adjective, which appear on the following line (“orribile idea”), and the verb “finiremo” is kept apart from the other verb (“per ucciderci”). There is also an unnecessary capitalisation of the conjunction “E” (and), while the rest of the sentence could have been accommodated on the same line (“per ucciderci e non funziona”).

Example 32

Source
A: con la tua orribile idea finiremmo per ucciderci
e non funziona
non ci sto
B: papà!
Mi manca (.) sai?

BT
A: with your horrible idea we would end up killing each other
and it doesn't work
I am not having it
B: dad!
I miss her, you know?

Broadstream	Amberscript
00:04:56,440 → 00:04:57,960 Dopo tre giorni con la tua orribile	0066 00:04:56:10 → 00:04:58:22 Dopo tre giorni con la tua orribile idea finiremo
00:04:58,060 → 00:04:59,510 idea finiremo per ucciderci.	0067 00:04:58:24 → 00:05:00:14 per ucciderci E non funziona. Non ci sto
00:04:59,610 → 00:05:00,150 E non funziona.	0068 00:05:00:16 → 00:05:08:06 fpapà. -Mi manca,
00:05:00,250 → 00:05:01,270 Non ci sto papà.	
BT After three days with your horrible idea we would end up killing each other. And it doesn't work. I am not having it dad.	BT After three days with your horrible idea we will end up killing each other And it doesn't work. I am not having fdad. -I miss her,

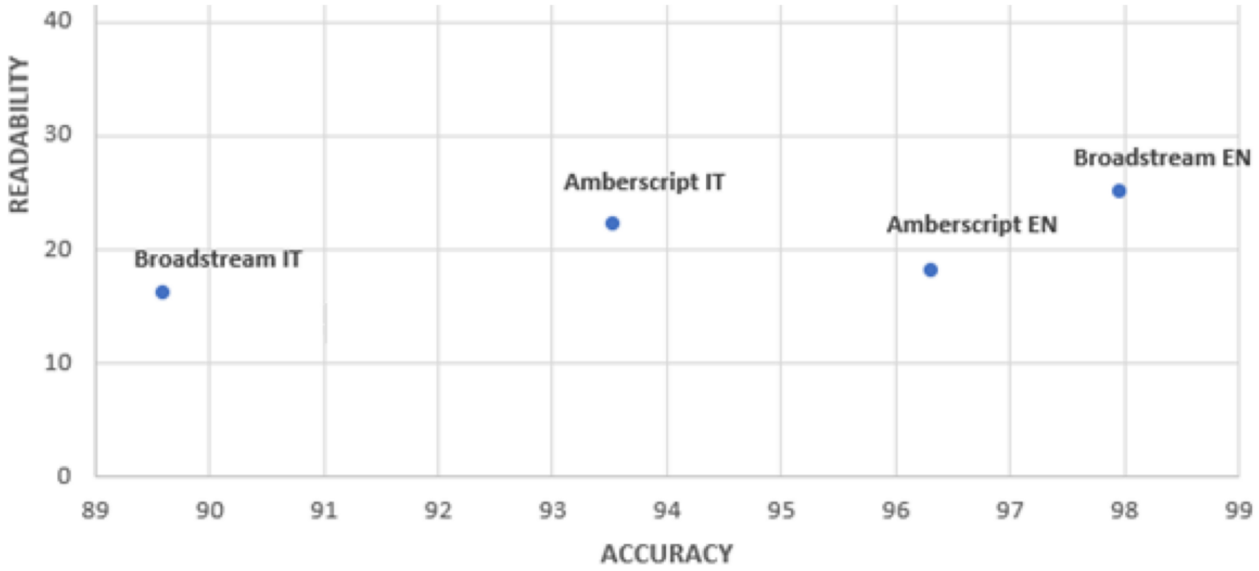
4. Discussion

This study was commissioned by an international broadcaster and focuses on evaluating the human effort needed to post-edit ASR-generated subtitles for two specific genres (i.e., talk show and feature film) in two languages, English and Italian, exploring to what extent off-the-shelf ASR systems could be used as a basis for subtitle creation, thus potentially reducing the effort associated with traditional human-made subtitles. The brief specifically asked for qualitative insights into prototypical problems in fully machine-generated subtitles.

Given the nature of the data provided, it was impossible to quantify the required editing effort. We only had access to the final output and could not replicate the in-house procedures and professional practices typically used for editing nor compare them to a fully human-driven workflow. Nonetheless, our analysis indicates that the bulk of human effort needed to refine the ASR outputs would be spent on addressing accuracy and segmentation issues. While several additional factors would need to be considered for a more accurate estimate of editing effort (e.g., language, audiovisual genre and related linguistic features, post-editing experience, familiarity with the software, knowledge of the subject matter, workflow organisation), accuracy and readability may serve as robust indicators for gauging baseline human effort. Figure 4 provides a visual representation of the performance of each ASR system, with the best performances in the top right quadrant.

Figure 4

ASR System Performance by Language



In relation to accuracy, the ASR tools achieved better results in English than in Italian. It is impossible to explain this with any certainty: it could be software-related (i.e., better language model for English), genre-related (talk show vs film), speech-related (monologic vs dialogic), or related to other factors. Based on the assumption that a greater distance from the 98% benchmark implies a higher degree of human editing, our findings suggest that lighter post-editing may be sufficient on the English subtitle files, particularly those generated by Broadstream. The recognised text seems to offer a good foundation to work on, especially considering the high speed of the original talk show, which would have been very time-consuming for manual transcription. This also suggests that in offline subtitling for pre-recorded content, the acceptable usability threshold may be lower than 98%, as long as the subsequent editing effort and time (which also hinge on other factors such as readability issues) do not outweigh the potential benefits of employing ASR. Determining the accuracy ranges corresponding to specific estimated effort levels would require an in-depth process investigation, ideally via an experimental study with a control group, accounting for different workflows and ASR configurations (e.g., customised vs non-customised). It is important to reiterate that our analysis has pinpointed specific challenges for non-customised ASR, as requested by the brief.

Although the impact of error categories may vary across different languages and ASR systems, both content- and form-related errors may be expected in the outputs, especially substitutions, correctness errors, and omissions. Error severity is also key, as it influences the level of human editing required. A relatively high accuracy score, combined with minor correctness issues like punctuation errors and minor substitutions/omissions, indicates that light editing should suffice. Conversely, detecting and rectifying multiple standard or serious errors (in a seemingly plausible output) requires constant source cross-checks and additional editing time. Most of the omissions identified in the files are minor, except for the English Broadstream output, with more instances of standard omissions. A similar pattern emerges in substitutions, with standard substitutions prevalent across the two evaluated ASR systems and languages; an exception is the Italian Broadstream file, where substitutions are mostly serious. Correctness errors are largely minor in both languages and ASR systems.

Finally, editing requirements have been found to be related to the broadcast genre. Unsurprisingly, monologic programmes (even with fast speech, multiple voices, and accents) seem to require less editing than the ASR outputs for films, which pose additional challenges related to identifying and signalling speaker changes. In our dataset, speaker changes were either poorly implemented or wholly absent. Marking a speaker change (aka speaker diarisation, i.e. segmentation according to who is speaking) is a very complex task that is not currently implemented by many commercial tools, despite representing an active research area. Additional source features found in both datasets and associated with increased editing requirements include background noise, overlaps, code-switching (i.e., alternating between two or more languages within a single conversation) and loanwords. Genre type also drives the density of proper names, which has emerged as a recurring problem. For instance, in our English dataset, the name "Zooey Zephyr" was transcribed by both ASR systems as "Zoe", which is incorrect but plausible: detecting and correcting such cases increases effort and time. Therefore, ASR engine customisation for specific audiovisual genres (or even within a genre, e.g. for specific programmes) needs to be considered to improve accuracy, even though not all names or entities can be expected to always be recognised.

As regards readability, the percentage of subtitles with incorrect segmentation (i.e., segmentation that does not follow the broadcaster's guidelines) gives an indication of the editing that would be required (see Tables 2 and 3). A higher percentage of subtitles requiring intervention corresponds to an increased level of expected human effort. Our findings show that line and subtitle break adjustments are required in 75% of cases in the English dataset and over 80% of cases in Italian. Segmentation issues appear less genre-dependent and are significantly exacerbated by accuracy problems, particularly punctuation errors. Customising ASR engines for segmentation issues (which, as previously highlighted, can be driven by the language model, speaker delivery, speaker changes, emphases, etc) remains challenging. Some improvements could be achieved through specific tailoring. For instance, tweaking the software to improve line breaks using post-processing rules that merge two short segments may help to produce subtitles aligned with display conventions. Nevertheless, segmentation remains closely linked to punctuation, variability in the source audio and contextual meaning, thus requiring a flexible approach to prevent issues from propagating into subsequent sets of subtitles. The potential impact of such customisation on reducing editing effort would require exploration in a dedicated, comparative study.

Although not subject to systematic analysis in this study, a final dimension that has emerged is reading speed. ASR engines typically aim for "verbatim" transcription of every sound in the soundtrack. While some tools can edit out disfluencies, stuttering, false starts, and hesitations, the overall word count, especially when transcribing fast speakers, may exceed the limits specified in subtitling guidelines. This would mean choosing between very fast subtitles (thus compromising readability – e.g. Szarkowska et al., 2024; Kruger et al., 2022) and edited subtitles (sometimes bordering on summarisation). Striking the right balance between these two extremes is challenging and requires human intervention, as both options may compromise accessibility. In the future, reception studies will be essential to identify audience needs and preferences.

5. Conclusions

This study has provided methodological insights and empirical evidence regarding human editing in automated intralingual subtitling. Our results underscore the importance of a nuanced and balanced evaluation of the impact of ASR on human effort. The ASR outputs analysed here do not fully meet broadcast-ready standards, at least for the audiovisual genres, languages, and ASR solutions commissioned for evaluation in this study. While acknowledging the potential productivity gains offered by ASR technology, research must adopt a multidimensional approach to understand the synergies between human intervention and AI-driven processes. Further research is needed to explore the threshold beyond which the human effort required to rectify ASR shortcomings outweighs the potential gains of automation. This prompts an exploration of how these technological developments redefine the concept of human agency within fully automated workflows. While our findings indicate that the combination of automatic captioning via ASR tools and human editing may be a viable option, possibly preferable to fully human-driven workflows, they also identify areas where additional specific editing skills are needed.

Overall, while the concept of editing is still not clearly defined (Bolaños García-Escribano & Declercq, 2023), our study confirms that different types of editing are indispensable before and after using ASR tools. Human agents can make informed decisions about the potential suitability of ASR tools based on information about the genre and characteristics of the source material. ASR optimisation tasks that can be carried out, e.g., software customisation through dedicated vocabulary lists and proper names, as well as segmentation-related adjustments, are linked to this. These can build on solid source content research and pre-existing knowledge of the key strengths and shortcomings of ASR. When human experts are asked to edit ASR output in AVT, this includes a dual-level intervention, namely "compliance with the linguistic and technical demands of the project" (Bolaños García-Escribano & Declercq, 2023, p. 568). Our empirical study has provided evidence in relation to both, with insights into the accuracy and segmentation-related issues that might emerge. In some cases, corrections of recognition errors and radical re-arrangements of subtitle segmentation may outweigh the benefits of using ASR in the first place, as substantial skills and expertise are required to turn ASR output into usable subtitles.

As a very dynamic field subject to continuous optimisation, progress in ASR is likely to involve the further refinement of algorithms for speech recognition and expansion of training corpora for ASR engines. Additionally, the development of Natural Language Understanding and Large Language Models are deemed to enhance resilience against accents, ambient noise disturbance and technical language. While ASR technology finetuned for subtitles is likely to lead to more streamlined editing processes, human refinement of ASR output remains necessary for the time being. This points to the importance of tailored training and upskilling to enable subtitlers to integrate these tools into their workflows more effectively. As was also shown in Sandrelli (2024), software developers seem to expect subtitlers to have all the required editing skills and expertise when presented with automation outputs. In line with previous research (Tardel, 2020; Karakanta et al., 2022), our study confirms that while ASR may support language professionals in their technical effort (i.e., transcription and time-coding), it may hinder them in their temporal effort, owing to the time needed to correct spotting problems, to check against the source and edit complex segments with multiple errors. Therefore, there is no “optimal” workflow or degree of automation and no “one-size-fits-all” solution, as the use of any assistive technology needs to be assessed against the specific characteristics of its purpose. Subtitlers need to become the ideal “augmented translator that puts the human front and centre and uses technology to enhance their capabilities” (AVTE, 2021), as was recently advocated by the European Federation of Audiovisual Translators.

With current concerns around the potential threat posed by the rise of AI-powered technologies to AVT and, more generally, to language-related professions, the only way forward to counteract these concerns is research providing empirical evidence of what humans are required to do in highly technologised AVT (and other translation) environments, how agency is shaped and, ultimately, what can be done to ensure productivity and efficiency without compromising quality.

Acknowledgements

This study was supported with resources from an Expanding Excellence in England (E3) grant by Research England (2019–2024). We would like to add a special thank you to Matilde Carbutto and Željko Radić who complemented the core team in analysing and reviewing the dataset for methodological rigour.

Authors contributions

Although the commissioned research was carried out jointly by the authors, specific contributions to this paper are: Davitti Data curation, Conceptualisation, Formal analysis, Writing – original draft (1.2, 2.2, 3.1, 4, 5); Sandrelli Data curation, Conceptualisation, Formal analysis, Writing – original draft (sections introduction, 1.1, 2.1, 3.2, 5); Korybski Data curation, Formal analysis, Review & editing; Zou English data evaluation, Review & Editing; Orasan Validation, Review & Editing; Braun Resources, Validation, Review & Editing.

References

AVTE. (2021). *Machine translation manifesto*. Audiovisual Translators Europe. https://avteurope.eu/wp-content/uploads/2022/10/Machine-Translation-Manifesto_ENG.pdf

Berman, J. (2022). *CWMF: Talent crunch is creating a localization challenge*. M+E Daily April 06. https://www.mesaonline.org/2022/04/06/cwmf-talent-crunch-is-creating-a-localisation-challenge/

Bolaños García-Escribano, A., & Declercq, C. (2023). Editing in audiovisual translation (subtitling). In A. Bolaños García-Escribano, & C. Declercq (Eds.), *Routledge encyclopedia of translation technology* (pp. 565–581). Routledge.

Bolaños García-Escribano, A., Díaz Cintas, J., & Massidda, S. (2021). Subtitlers on the cloud: The use of professional web-based systems in subtitling practice and training. *Revista Tradumàtica. Tecnologies de la Traducció*, 19, 1–21. https://doi.org/10.5565/rev/tradumatica.276

Bryant, M. (2021, November, 14). Where have all the translators gone? *The Guardian*,. https://www.theguardian.com/tv-and-radio/2021/nov/14/where-have-all-the-translators-gone

Davitti, E., & Sandrelli, A. (2020) Embracing the complexity: A pilot study on interlingual respeaking. *Journal of Audiovisual Translation, 3*(2), 103–139. https://doi.org/10.47476/jat.v3i2.2020.135

Díaz Cintas, J., & Massidda, S. (2019). Technological advances in audiovisual translation. In M. O'Hagan (Ed.), *The Routledge handbook of translation and technology* (pp. 255–270). Routledge.

Díaz Cintas, J., & Remael, A. (2021). *Subtitling: Concepts and practices*. Routledge.

Karakanta, A., Bentivogli, L., Cettolo, M., Negri, M., & Turchi, M. (2022). Post-editing in automatic subtitling: A subtitlers' perspective. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, 261–270. EAMT.

Kruger, J.-L., Wisniewska, N., & Liao, S. (2022). Why subtitle speed matters: Evidence from word skipping and rereading. *Applied Psycholinguistics, 43*(1), 211–236. https://doi:10.1017/S0142716421000503

Ludera, E., Szarkowska, A., & Orrego-Carmona, D. (2024). Expertise in interlingual subtitling: Applying the FAR model to study the quality of subtitles created by professional and trainee subtitlers. *Translation & Interpreting, 16*(1), 55–75. https://doi.org/10.12807/ti.116201.2024.a04

Massidda, S., & Sandrelli, A. (2023). ¡Sub! localisation workflows (th)at work. *Translation and Translanguaging in Multilingual Contexts, 9*(3), 298–315. https://doi.org/10.1075/ttmc.00115.mas

Pedersen, J. (2017). The FAR model: Assessing quality in interlingual subtitling. *JosTrans, 28*, 210–229.

Romero-Fresco, P., & Pérez, J. M. (2015). Accuracy rate in live subtitling – the NER model. In J. DíazCintas, & R. Baños Piñero, (Eds.) Audiovisual translation in a global context: Mapping an ever-changing landscape (pp. 28–50). Palgrave.

Romero-Fresco, P., & Pöchhacker, F. (2017). Quality assessment in interlingual live subtitling: The NTR model. *Linguistica Antverpiensia*, 16, 149–167. https://doi.org/10.52034/lanstts.v16i0.438

Sandrelli, A. (2024). Integrating ASR and MT tools into cloud subtitling workflows: The ¡Sub! and ¡Sub!2 projects". In Y. Peng, D. Huang, & D. Li, (Eds.), New advances in translation and interpreting technology, springer new frontiers in translation studies (pp. 79–97), Springer Nature. https://doi.org/10.1007/978-981-97-2958-6

Sandrelli, A. (forthcoming). The technologisation of AVT: An experiment on cloud subtitling and implications for training. *Textus*.

Szarkowska, A., Ragni, V., Orrego-Carmona, D., Black, S., Szkriba, S., Kruger, J-L., Krejtz, K., & Silva, B. (2024). The impact of video and subtitle speed on subtitle reading: An eye-tracking replication study. *Journal of Audiovisual Translation, 7*(1), 1–23. https://doi.org/10.47476/jat.v7i1.2024.283

Tardel, A. (2020). Effort in semi-automatized subtitling processes: Speech recognition and experience during transcription. *Journal of Audiovisual Translation, 3*(1), 79–102. https://doi.org/10.47476/jat.v3i2.2020.131

Tuominen, T., Koponen, M., Vitikainen, K. Sulubacak, Y., & Tiedemann, J. (2023). Exploring the gaps in linguistic accessibility of media: The potential of automated subtitling as a solution. *JosTrans, 39*, 77–98.