Romero-Fresco, P. & Fresno, N. (2023). Accuracy of automatic and human live captions in English. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *22*, 114–133.

The accuracy of automatic and human live captions in English

Pablo Romero-Fresco (Corresponding author)
Universidade de Vigo, Galician Observatory for Media Accessibility (GALMA)
promero@uvigo.gal
https://orcid.org/0000-0003-2166-5792

Nazaret Fresno
The University of Texas Rio Grande Valley
nazaret.fresno@utrgv.edu
https://orcid.org/0000-0002-6702-159X

Abstract

Closed captions play a vital role in making live broadcasts accessible to many viewers. Traditionally, stenographers and respeakers have been in charge of their production, but this scenario is changing due to the steady improvements that automatic speech recognition has undergone in recent years. This technology is being used to create intralingual live captions without human assistance and broadcasters have begun to explore its use. As a result, human and automatic captions co-exist now on television and, while some research has focused on the accuracy of human live captions, comprehensive assessments of the accuracy and quality of automatic captions are still needed. This article airs this matter by presenting the main findings of the largest study conducted to date to explore the accuracy of automatic live captions. Through four case studies that included approximately 17,000 live captions analysed with the NER model from 2018 to 2022 in the United Kingdom, the United States, and Canada, this article tracks the recent developments with unedited automatic captions, compares their accuracy to that achieved by human beings, and concludes with a brief discussion of what the future of live captioning looks like for both human and automatic captions.

Keywords: live captioning, automatic speech recognition, accuracy, respeaking, NER model

1. Introduction

Technology has always played a key role in audiovisual translation (AVT) and media accessibility (MA), ever since the first subtitles were burnt onto silent films at the turn of the 20th century (Ivarsson & Carroll, 1998) and the first closed captions were produced for television programmes in the United States and the United Kingdom more than 40 years ago (Neves, 2005). Captioning, dubbing, and audio description rely on ever-evolving technology that, over the past several years, has had a tremendous impact on this area, as shown by the way in which workflows have changed recently with the introduction of cloud-based AVT and MA (Bolaños-García-Escribano et al., 2021). Until now, technology has mainly been used to aid human translators, but the latest developments in automatic speech recognition (ASR) and machine translation (MT) point to scenarios where the role of the human being is to revise the output produced by automatic software or where they may be replaced altogether. Needless to say, not all types of text lend themselves equally well to automatic translation and, in this sense, audiovisual media have often been regarded as a particularly challenging area. However, regarding the production of intralingual live captions, the steady improvement of ASR technology has led companies, event organisers, and broadcasters across different continents to resort to fully automatic captions, which now share the stage with live captions produced by stenographers or respeakers (Pérez Cernuda, 2022). As the different stakeholders consider what types of caption to use, it becomes essential to analyse the quality of the captions and, more specifically, their accuracy, to find out whether human and automatic captions can both provide viewers with the access they need.

This article presents the largest study available to date (at least to our knowledge) of the accuracy of English human and automatic live captions. It includes approximately 17,000 captions, that is, 798 minutes of live captions analysed using the NER model from 2018 to 2022 in the United Kingdom, the United States, and Canada. Of these 798 minutes, 388 are produced by human captioners (through respeaking and stenography) and 410 are produced by ASR. The four-year span analysed here has enabled us to track the recent development of automatic and human captioning and to draw conclusions about the current scenario and what may be expected in the coming years. Before moving on to the four case studies that comprise this article, the next section reviews the research carried out in this area and especially those studies that have focused on the quality of human and/or automatic live captioning.

2. Prior research on live captioning quality

Despite their importance in the provision of access to millions of viewers around the world, research on the quality of live captions is still very limited and almost exclusively focused on human captions. In the United Kingdom, and following complaints from the users, the official government regulator, Ofcom, decided to set up a nationwide assessment of live captioning in 2013. Inma Pedregosa and the first author of this article were tasked with analysing the accuracy, delay, speed, and reduction rate[i] of 78,000 human captions, mostly produced by respeakers for all terrestrial television channels in the United Kingdom between 2013 and 2015 (Romero-Fresco, 2016). Accuracy was measured using the NER model (Romero-Fresco & Martínez, 2015), which identifies the impact that omissions, misrecognitions, and other errors in the captions (as compared to the audio) have on viewers' comprehension. With the NER model, errors can be minor (if they do not have an impact on comprehension), standard (if they cause confusion or omit information that is needed to understand a message), or serious (if they introduce a new, credible, and incorrect meaning in the captions). In order to have an acceptable level of accuracy, captions must reach a 98% threshold.[ii] The results of the Ofcom study showed an overall accuracy rate of 98.4%, that is, the equivalent of a 6 in a 10-point scale.

In the United States, in 2003 and prior to the inception of the NER model, Jordan et al. (2003) assessed the closed captions delivered in local and national news broadcasts. They investigated their accuracy by deciding how closely the captions matched the audio track in the programmes and estimated that 100% of the national news included "clear" or "somewhat clear" captions. However, that percentage decreased to 68% for local news. A few years later, Apone et al. (2011) used the Weighted Word Error Rate (WWER), an updated version of the WER model that takes severity into account, to assess 20 news programmes. The researchers found that only 55% of them featured captions with a good level of accuracy and that 10% had low-quality captions that would not serve their intended audiences. Fortunately, the accuracy of live closed captioning has improved since then, at least for news programmes, as the most recent study conducted in the United States by Fresno (forthcoming) demonstrates. Using the NER model, she analysed 20 national news programmes broadcast between 2019 and 2020 and concluded that they featured a 98.8% accuracy rate on average, with 14/20 samples achieving accuracy levels ranging from "acceptable" to "excellent".

Fresno has conducted several studies on live captioning quality focusing on other genres and using the NER model. The first study explored the 2016 final US presidential debate and found that the average accuracy rate of the live closed captions delivered by six broadcasters was 98.8% (Fresno, 2019). The 2018 Super Bowl was the object of a similar quality analysis and it revealed that the closed captioning accompanying this widely followed event achieved a 99.4% accuracy rate (Fresno et al., 2021).

Several quality studies have also been carried out in languages other than English. For instance, Fresno (2021) focused on the national newscasts aired in Spanish in the United States. According to her study, these closed captions achieved a 97% accuracy rate, which did not reach the acceptable threshold under the NER model and which was substantially below the accuracy level found in the English newscasts broadcast in the United States (98.8%). Also for Spanish, the first tentative results of the QualiSpain project were released in 2019. This was the first quality analysis of live closed captioning conducted at a national level in Spain, and the findings gathered to date point at an accuracy rate of 98.9% for news programmes (Fresno et al., 2019).

More recently, Dutka (2022) used the NER model to analyse a corpus comprising 96 samples of live and semi-live captions (a total of 13,620 live captions) broadcast on Polish TV between 2021 and 2022. The average accuracy rate of the live captions in his corpus is 96.7% (1/10), which is significantly below the threshold of acceptable quality.

3. Our case studies

In this section we report on four case studies that we conducted between 2018 and 2022 in order to assess the accuracy of automatic and human live closed captions in English in the United Kingdom, the United States, and Canada.

3.1. Case study 1: Sky

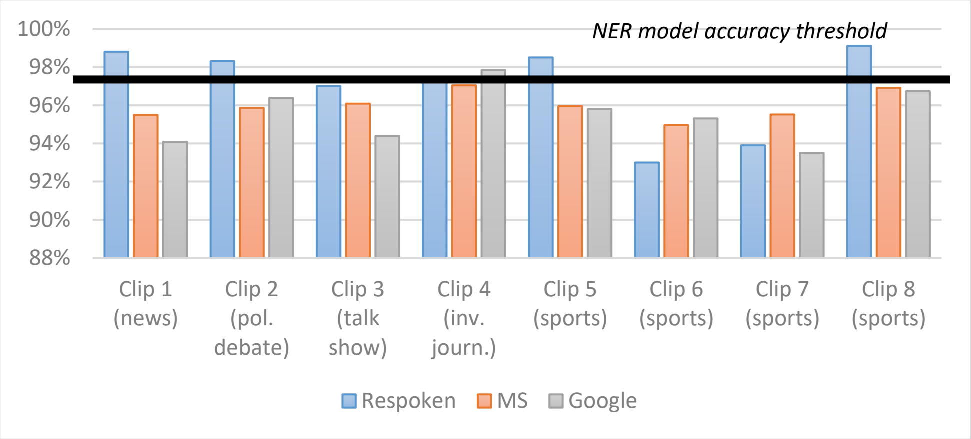
This project was conducted in the United Kingdom in 2018 and was aimed at studying the accuracy of the automatic closed captions produced by two speech-recognition engines, Microsoft and Google, compared to that achieved by respeakers. We worked with a total of 58 minutes of audiovisual content provided by Sky, which consisted of eight clips:
Clip 1 (5 minutes) was extracted from a news programme featuring three presenters.
Clip 2 (8 minutes) is an excerpt from a debate between several politicians.
Clip 3 (5 minutes) includes a segment from a talk show with four speakers.
Clip 4 (5 minutes) is part of an investigative journalism piece with one main narrator.
Clips 5–7 (10 minutes each) and Clip 8 (5 minutes) deal with sports.

For each clip, we had three captioning files, one produced by a respeaker, one created by the Microsoft speech-recognition engine and another generated by Google.[iii] None of them were broadcast.

The quantitative analysis that we performed for the respoken samples revealed an average accuracy rate of 97% (2.5/10), below the results reported in the Ofcom project (98.4%) and also below the acceptable accuracy threshold under the NER model (98%). These poor results were mainly due to two very weak sport samples that scored below 94%. The rest of the programmes were all between 97% and 99.1% accuracy, with four of them reaching acceptable accuracy levels and two of them being either “good” or “very good”.

Regarding the corpus of automatic captions, the average accuracy rate was 95.7% (0/10). All the individual samples in this study scored below 98% and were therefore considered “substandard”. Microsoft achieved an average accuracy rate of 96% (0/10), with all the samples ranging from around 95% to slightly above 97%. Google also performed poorly, with an average of 95.5%. However, it managed to produce captions that were very close to being acceptable for one programme (the investigative journalism piece, which attained 97.8%). The rest of the Google samples had accuracy rates from 93.5% to slightly over 96.7%. Figure 1 shows these results.

Figure 1
Accuracy rates in case study 1



As part of the study, we also looked at the reduction rate, which was 39.3% for the respoken captions and unexpectedly high for the automatic samples. The average reduction rate for the automatic captions was 6.6%, with the closed captions provided by Microsoft reaching 8.6% and those by Google maintaining the reduction rate at a lower 4.6%. In the case of human captions, the 39% reduction rate was due to editing strategies that the respeakers applied to keep track of the audio of the programme and therefore reduce delay. In the case of the automatic captions, however, the reduction was the by-product of glitches and, more often, misrecognitions. These were instances in which several words from the original message were either omitted entirely from the captions because the software was unable to recognise them or they were mistakenly transcribed in the captions using fewer words than the original (e.g., “there is the fence now” was captioned as “Darius defense now”).

Regarding errors, the respoken captions included a total of 587, that is, 8.5 errors/minute on average. Of those, 82% had to do with unsuccessful edition and 18% were recognition problems. This means that 8/10 mistakes involved relevant information being omitted from the captions, possibly in an attempt to cope with fast speech rates that exceeded 210 wpm in half of the samples. The automatic closed captions contained 3,407 errors, that is, as many as 24 errors/minute on average, with around 21 errors/minute reported for Microsoft and 28 errors/minute for Google. Interestingly, the two speech-recognition engines used in this study behaved somewhat differently when assessed by error severity: Microsoft delivered 65%, 33%, and 2% minor, standard, and serious errors, respectively, but these percentages changed to 80%, 19%, and 1% for Google. These findings indicate that Microsoft had fewer total and minor errors than Google, but the latter outperformed Microsoft in standard and serious errors. In other words, whereas the captions provided by Google contained more errors, they were less likely to cause comprehension problems for the viewers.

Most of the errors in the automatic captions were misrecognitions, which were more frequent in proper names and small words (e.g., prepositions or contractions) and after false starts, openings of programme sections, and changes of speaker. Low voices, background music, or noise and speaker accents that diverted from the language model used by the speech-recognition software were also problematic. In addition to misrecognitions, both engines accumulated many errors caused by missing or misplaced punctuation marks. Punctuation accounted for as many as 44% of the total errors in all the samples that we worked with, but these issues were more frequent in the captions created by Google (which was to be expected, given that three out of eight samples did not include any punctuation whatsoever). All the captioning files produced by Microsoft were punctuated, which substantially reduced the number of punctuation errors (448 for Microsoft compared to 801 for Google). Admittedly, the vast majority of the punctuation errors were minor and would not have interfered with the viewers' ability to comprehend the message. However, punctuation errors became problematic when they were not isolated. In other words, when several punctuation marks affecting the same cluster of sentences were missing or misplaced, making sense of the message became more difficult. This was perhaps more evident in the unpunctuated samples, where the lack of punctuation marks at times made it difficult to understand where each idea began and ended, hence disrupting the reading process.

Given that punctuation accounted for almost half of the errors in these closed captions, we estimated the accuracy rate for each sample without factoring in punctuation and related capitalization errors (e.g., capital letters after a full stop). When punctuation was not considered, the accuracy rate increased but not sufficiently to reach acceptable levels under the NER model. Our automatic corpus went up from 95.7% to 97.3% on average, which was higher than the average accuracy rate achieved by respeakers in these samples. Furthermore, four programmes showed worse accuracy levels in the respoken captions than in the automatic samples. When punctuation was not factored in, the automatic captions over the 98% threshold went from zero to two samples. Clip 4, the journalistic segment, attained an accuracy rate of 98% with Microsoft and 98.5% with Google. These particularly good results were probably due to the fact that this programme was narrated by one speaker and most of the text was scripted, that is, the speech was well organised and the intonation and utterances were clear. These factors may have facilitated more efficient automatic recognition and punctuation as compared to the rest of the samples, where the speech was more spontaneous.

In summary, the live captioning accuracy achieved in 2018 by Microsoft and Google was insufficient to allow for viewers' comprehension. Most of the errors identified in the automatic captions were minor, but there were so many that making full sense of the programmes via the captioning was difficult at times. Punctuation was partly responsible for these poor results, but our study demonstrated that even if the punctuation were to be corrected, the accuracy remained below acceptable levels due to the misrecognitions present in the captions.

3.2. Case study 2: Vitac

This study was commissioned in 2020 by Vitac, a closed captioning company based in the United States. The main goal was to compare the accuracy of steno, respoken and automatic captions in an audiovisual corpus made up of three 10-minute samples.

Clip 1 was extracted from a local news programme with an anchor delivering the news.

Clip 2 was part of a morning talk show and included a cookery piece in which one host talked about some pastry and held a phone conversation with a cook.
Clip 3 was part of a videoconference with five participants.

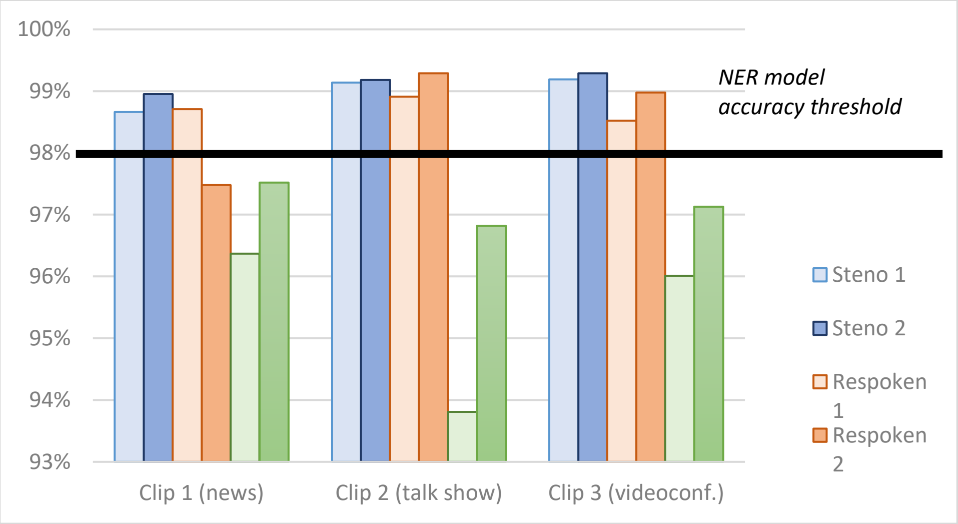
For each clip, we received six captioning files: two produced by stenographers, two created by respeakers and two generated by speech-recognition software (VoiceInteraction and Enco).

Beginning with steno captions, their accuracy reached 99.1% (7.5/10) on average and was "very good" according to the NER model. The news samples showed a slightly lower accuracy rate than those of the talk show and the videoconference (98.8%, 99.2%, and 99.2%, respectively). But all the captioning samples scored well above the 98% accuracy threshold.

Respoken captions achieved good results, although their average accuracy rate (98.6%) positioned them slightly below their steno counterparts. The respeakers also performed better with the talk show clip (99.1%) than with the videoconference (98.8%) and the news programme (98.1%).

As far as the automatic captions are concerned, their accuracy was poor: 96.3% (1/10) on average. In fact, none of the automatic samples reached acceptable accuracy levels, with those of the news programme scoring marginally better (96.9%) than the videoconference (96.6%) and the talk show segments (95.3%). Figure 2 shows the results obtained in our corpus.

Figure 2
Accuracy rates in case study 2



Regarding the reduction rate, stenographers applied an average 10.9%, respeakers decreased reduction to 8.1%, and automatic captions maintained a low 2.1%.

The closed captions produced using stenography showed the lowest number of errors (307 in total); 78% of those were minor, 18% standard, and 4% were serious. The respoken captions contained more mistakes (438) but a similar distribution of severity (76%, 21%, and 3%, respectively). On average, the automatic captions tripled the number of errors in human closed captions (1,295 in total), with 75% of minor errors, 24% of standard, and 1% of serious mistakes. The closed captions produced by steno captioners showed a total of 5 errors per minute, followed by those generated by respeakers and by the ASR engines, which included 7 and 22 errors per minute, respectively.

According to our findings, most of the samples captioned by human beings amply surpassed the 98% threshold and achieved “good” or “very good” accuracy levels for 11 out of the 12 clips analysed. There was only one respoken news sample that was below par (97.5%) due to unusually frequent editing problems. As mentioned before, both stenographers and respeakers performed better with the talk show than with the rest of the genres, which may be explained by the speech rate and density. For instance, when compared to the news programme, both clips had speech rates around 180 wpm, which led captioners to edit in order to keep pace with the programmes. However, the talk show included some reiterative dialogue and part of the information presented verbally was strongly reinforced through the images (e.g., the presenter described the texture of several pastries while she showed them on camera). This allowed the captioners to omit information that had been included in previous captions or that could be fully understood by looking at the images without affecting the viewers’ ability to comprehend the message. In the news sample, such redundancy was less common and editing became more problematic, since most omissions caused new information to be lost.

Human captions had more editing than recognition errors, with the latter being more common in respoken samples (38% of all the errors encountered compared to 18% in steno captions). This suggests that the interaction with their technical equipment was smooth for stenographers but somewhat more challenging for respeakers. The steno captions analysed here featured five errors per minute of programme on average, most of which were minor editing errors that involved a slight loss of meaning (e.g., adjuncts not making their way to the captions). The respoken captions included seven errors per minute on average, with four of those being editing errors and three being misrecognitions.

Contrary to the tendency identified in the steno and respoken samples, the automatic captions included fewer errors in the news sample, which may be due to the fact that it was a fully scripted programme with well-structured and -articulated speech. The scripted nature of this programme combined with its polished delivery seemed to enhance the performance of the speech-recognition engines, in particular by reducing the number of standard misrecognitions. Despite this, however, only Enco managed to produce closed captions for the news sample that were not too far from the 98% accuracy threshold (97.5%). This programme showed the fewest errors per minute (15 on average), but almost four times the number of mistakes in the best respeaking and steno captioning samples (four errors per minute). Furthermore, our corpus of automatic captions included many punctuation errors. As in case study 1, most of them were minor, but their abundance, together with the misrecognitions present in these captions, often caused the reading flow to be disrupted. Overall, the automatic closed captions contained 21.5 errors per minute, mostly in the form of punctuation errors and misrecognitions.

In this study, neither VoiceInteraction nor Enco yielded acceptable closed captions and were clearly outperformed by the steno captioners and respeakers regardless of the genre and the sample characteristics. However, our analysis showed some progress in the accuracy of the automatic captions, especially for Enco, which managed to achieve accuracy rates above or close to 97% and fewer errors per minute than those reported for automatic captions in case study 1.

3.3. Case study 3: Enco

In this study, which was carried out in the United States in 2021, we compared the accuracy of human closed captions delivered on television to those provided by Enco, one of the speech-recognition engines that we had used in case study 2 one year previously. We worked with 30 minutes of programmes recorded from television using the Hauppage WinTV-HVR-1955 TV tuner. The televised closed captions were extracted with CC Extractor GUI 0.88. The automatic closed captions were provided by Enco's developer.

The materials for this study included a segment from a news programme (10 minutes), part of an American football game (10 minutes), and a piece from a talk show (10 minutes). Following Enco developer's instructions, we divided each recording into two five-minute clips.

Clip 1 depicted the evening newscast *PBS NewsHour* and consisted mainly of one anchor reading the daily news from a teleprompter.

Clip 2 was part of the same news programme and showed the same anchor interviewing President Joe Biden.

Clips 3 and 4 were extracted from the Super Bowl, with two announcers commenting on the game without overlapping. Throughout clips 3 and 4, the cheering crowd in the stadium could be heard in the background. While this was kept at a low volume and the commenters could be heard clearly over the background noise, the sports clips had the least clean audio in our corpus.

Clips 5 and 6 were part of NBC's morning talk show, *Today*. Clip 5 included three speakers commenting on popular events and Clip 6 incorporated a fourth presenter and a three-minute video with several people speaking, one at a time. Clips 5 and 6 combined parts in which the presenters followed structured scripts, with more spontaneous pieces consisting of relaxed conversations among the presenters. The speakers respected their talking turns most of the time.

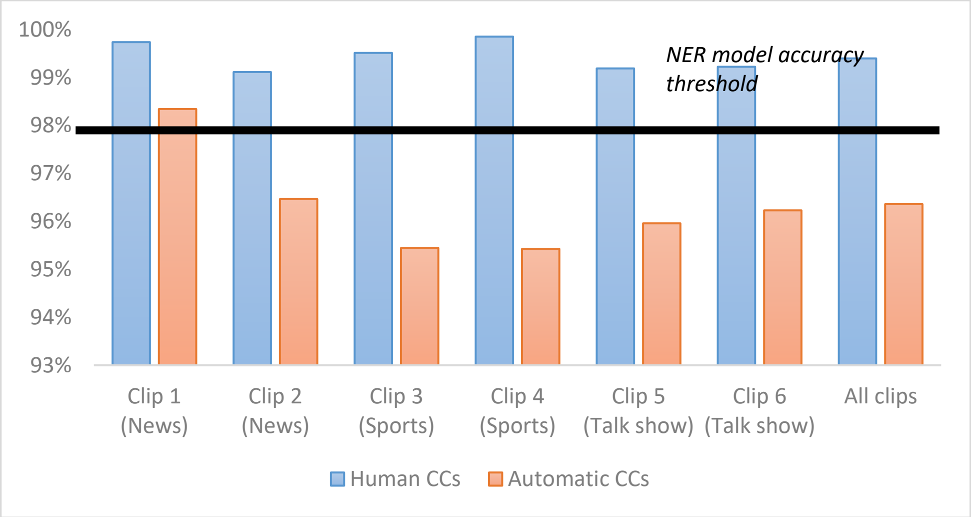
The human closed captions delivered on television boasted a 99.4% (8.5/10) average accuracy rate for the news samples, 99.6% (9/10) for the sports segments, and 99.2% (8/10) for the talk shows. All the human samples scored above 99.1% and, according to the NER model, the closed captions accompanying the news and talk show programmes would be "very good", while those from the football game would have an "excellent" accuracy.

In the case of the automatic captions, the overall accuracy rate for the entire corpus reached 96.3% (1.5/10). News programmes had the best results, with 97.4% (3.5/10), followed by talk shows and sports, which reached 96.1% (1/10) and 95.4% (0/10), respectively. One of the automatic samples (Clip 1) showed acceptable accuracy levels under the NER model (98.3% or

5.5/10), but the remaining five programmes scored somewhere between 95.40% and 96.5% (0–1/10). Figure 3 shows the accuracy rates in our case study 3 corpus.

Figure 3

Accuracy rates in case study 3



As far as the reduction rate is concerned, we found an overall 10.5% in our human closed captions and significant differences between genres (5.6% for news, 7.1% for sports, and 16.2% for talk shows). Regarding automatic captions, the reduction rate averaged 1.5%. Interestingly, the reduction also varied considerably across these samples, ranging from 7.9% in one of the talk shows to negative reduction rates in three samples – meaning that the captioned text was longer than the original message delivered by the speakers. In all cases, this was due to misrecognitions that added incorrect words to the captions (e.g., “for the first time since 1984” was captioned as “for the first time since 1980 for”).

The human closed captions analysed in this study showed a total of 89 errors. Of those, 65% were minor and would not have had a negative effect on the viewers’ comprehension; 33% were standard and would hamper the ability of the user to understand the message; and 2% were serious and would present the viewer with misleading information. Regarding genres, sports featured the fewest mistakes (19% of the total), followed by news (25%) and talk shows (56%). The automatic closed captions for the same clips included many more errors (606 errors in total). As for severity, 80% were minor, 17% were standard, and 3% were serious. The news segments contained the fewest errors (24% of the total), followed by sports (32%) and talk shows (44%).

Just as in the previous case studies, the human closed captions were more accurate than their automatic counterparts. Closed captioners kept their errors low, especially in the news and sports clips. Talk shows proved more challenging and included more incorrect editing, usually in the form of information omissions, possibly due to the faster pace of the speakers (213 wpm compared to 171 wpm for sports and 155 wpm for news). Overall, though, human captions showed an impressive three errors per minute on average.

When compared to automatic captions, human beings performed consistently better in three areas that proved especially problematic for Enco: punctuation, speaker identification, and misrecognitions. Enco performed better with the news samples: 60% of the errors encountered were punctuation and 11% were related to speaker identification. That is, either the software did not identify changes of speakers or it included chevrons to signal that a new person was talking when that was not the case. Most of the remaining errors were misrecognitions, especially of short words, which proved to be challenging in the three genres that we worked with. They were either transcribed for a mistaken word (e.g., “go” instead of “ago”), omitted (e.g., “is” instead of “he’s”) or incorrectly added to the closed captions (e.g., “and *in* the best example of that is” instead of “the best example of that is”).

A similar pattern was discerned in the sports programmes, where punctuation errors and speaker identification accounted for 52% and 10% of the errors, respectively. The proportion of misrecognitions increased for this genre because Enco struggled to recognise proper names correctly; this was problematic due to the frequency with which players were named by the commentators while they described the play-by-play action. While we expected the background noise in the sports clips to affect Enco’s performance, it did not seem to decrease recognition quality.

Finally, the talk shows displayed a different tendency, with a lower proportion of punctuation problems (28% of the total errors) and a higher proportion of speaker identification issues (26%) and misrecognitions. This was possibly due to the conversational nature of these samples, with shorter sentences, more turn-taking among speakers, and a faster delivery pace.

Again, the human closed captions included in this study were more accurate than those produced by Enco. In fact, they showed remarkably good results, which probably had to do with the fact that they accompanied some of the most well-known programmes in their slots. Therefore, the captions analysed here were possibly prepared by very experienced captioners who had trained their captioning software in advance – for instance, with glossaries of common terms used in those particular broadcasts. Enco fell significantly short of reaching similar results. However, it achieved acceptable accuracy levels above 98.3% for one news sample where errors were kept down to 8 per minute. While this performance was not at all consistent throughout the corpus, it illustrated that ASR engines have the potential to achieve acceptable accuracy rates, at least under very controlled conditions (for instance, with samples where the speaker reads the scripted content and delivers it with good diction).

3.4. Case study 4: Live TV captions in Canada

In 2015, the Canadian Radio-television and Telecommunications Commission (CRTC) started a consultation process to look at different methods with which to assess the quality of live captions in Canada (CRTC, 2015). This prompted the creation of the 2016 Working Group made up of broadcasters, captioning providers and user associations, which proposed a trial to adapt the NER model to the Canadian context and to test its validity in measuring the accuracy of live captioning on Canadian TV (Canadian Radio-Television and Telecommunications Commission, 2016). Between 2017 and 2018, the first author of this article trained 11 hearing and 10 deaf and hard-of-hearing evaluators to assess a series of live-captioned programmes using the Canadian version of the NER model. The objective of the trial was to ascertain whether the NER model could produce consistent results and if these results were in line with subjective impressions of caption quality. The study concluded that, provided that evaluators are trained in the use of the NER model, its results are reliable and useful (2016 Working Group, 2018) and they are aligned to the user needs as found by the Canadian Association of the Deaf in a user research project carried out that same year (CAD, 2018).

Following the result of this trial, on 30 August 2019 the CRTC issued its Broadcasting Regulatory Policy 2019-308 (CRTC, 2019) that required all broadcasters airing live captions to reach a 98% accuracy rate according to the Canadian NER model. Each month, every broadcaster must calculate the accuracy rate for two English-language programmes containing live content: one must be a news programme (if available) and the other must be from a genre that is representative of the live programme mix of the broadcaster. The captions must be assessed by NER-certified evaluators and by the end of each year broadcasters must provide the Commission with a report that includes their results and a description of their efforts to reach or maintain the required quality threshold.

Since the CRTC mandate was first enforced, nine broadcasters have submitted their assessments and reports for 2019–2020 and 2021–2022: Accessible Media Inc, Anthem Sports and Entertainment, Channel Zero, Crossroads Television System, Jim Pattison Broadcast Group, Salt and Light TV, Stingray Group Inc., Telelatino Network Inc. and The Miracle Channel Associations. The case study presented here includes the analysis of all these submissions, which make up a total of 18 reports and the assessment of 440 10-minute live caption samples (that is, 4,400 minutes or roughly 9,300 captions) selected by the broadcasters and reviewed by NER-certified evaluators.

The reports show that all the broadcasters are adopting measures to meet the CRTC mandate: for instance, hiring external NER-certified evaluators to assess their live captions or, as in the case of Crossroads Television System and Jim Pattison Broadcast Group, having their staff trained to become NER-certified evaluators. Interestingly, in a country such as Canada, where verbatim captioning has traditionally been favoured, the NER model has also been used to train captioners how to edit, that is, how to paraphrase or sum up the original audio when a verbatim transcript is difficult to achieve:

> When attempts at verbatim would make captions difficult to read or could be at risk of being cut off by commercial breaks after falling too far behind, Quay Media Services closed captioners use strategic editing, supported by NER theory, to make a more informed decision thereby reducing the risk of compromised context or meaning (AMI, 2021).

Other measures reported to improve the quality of human captions include providing captioners with custom dictionaries and lists of terms that are likely to be used on air and, in general, bringing the captioning teams closer to the creative teams of the programmes to be captioned, so that they are familiar with the content. As mentioned above, some broadcasters opted for fully automatic captions, often due to their reduced cost compared to human captions (Stingray Group Inc., 2021; TLN, 2021). Table 1 shows the average accuracy rate obtained by every broadcaster in 2019–2020 and 2020–2021 and the type of production method (human or automatic) they have opted for.

Table 1
Average accuracy rate by broadcaster in 2019–2020 and 2020–2021

Broadcaster	Human highest	Human lowest	Human average	Automatic highest	Automatic lowest	Automatic average	Samples
AMI 2019–2020	99.4	98.7	99.02				24
AMI 2020–2021	99.4	98.4	99.1				25
Anthem 2019–2020	98.8	96.6	97.9				15
Anthem 2020–2021	99.0	96.5	97.8				28
Channel Zero 2019–2020	99.7	96.98	99.0				24
Channel Zero 2020–2021	99.48	96.43	98.6				24
Crossroads 2019–2020	99.96	98.7	99.7				16
Crossroads 2020–2021	99.9	98.95	99.7				24
Jim Pattison 2019–2020				98.69	96.8	97.95	27
Jim Pattison 2020–2021				98.69	97.76	98.3	31
Salt and Light 2019–2020	99.81	97.86	98.7				9
Salt and Light 2020–2021	99.87	97.1	99.0				23
Stingray 2019–2020				98.7	96.9	97.4	48
Stingray 2020–2021				98.0	96.5	97.5	46
TLN 2020–2021				97.65	96.83	97.3	9
Miracle 2019–2020				98.76	97.0	97.7	35
Miracle 2020–2021				98.75	97.9	98.3	24

The results show that, overall, human captions are good according to the NER model (98.9% average accuracy rate). They are also more accurate than automatic captions, which feature more samples below 98% accuracy and fewer samples exceeding a 99% accuracy rate (60% of the automatic samples did not reach the minimum threshold and only 1% featured an excellent accuracy).[iv] This is in line with the findings obtained in the previous three case studies presented in this article, but some interesting developments may be noted. The average accuracy rate of human captions (98.9%) is the same as in case study 2, higher than in case study 1 (97%, which may be considered unusually low) and lower than in case study 3 (99.4%, unusually high). The lowest accuracy rates are 97.8%, slightly below the 98% threshold. The highest accuracy rates reach 99.7%, which is extremely high, especially considering that these are averages of 40 samples.

The main development, however, may be found in the accuracy of automatic captions. Whereas in the previous case studies, conducted between 2018 and 2021, the accuracy rate of automatic captions was still far from the NER threshold (95.7% in case study 1 and 96.3% in case studies 2 and 3, that is, a 0–0.5 in a 1–10 scale), the results from Canada between 2019 and 2021 show a significant improvement. The average accuracy does not drop below 96.5% in any of the automatic samples and, for the first time, broadcasters such as Jim Pattison Broadcast Group and The Miracle Channel Associations consistently reach accuracy rates above the required threshold in 2020–2021. Impressively, of the 55 samples analysed by these two broadcasters in 2020–2021, 52 reached 98%.

As explained in the report produced by Jim Pattison Broadcast Group (2020), the results obtained by their automatic captions in 2019–2020 were too variable and slightly below the required standards. This led the broadcaster to introduce the following improvements, which we summarise here, as they may be useful for other companies wishing to push their automatic captions above the NER threshold:

- meeting regularly with Enco, the manufacturer of the captioning software, to review the accuracy reports;
- introducing in the software, monthly, commonly used words and phrases spoken on air, especially locally relevant names;
- removing from the dictionary words that are no longer required;
- updating the workflow so that scripts from the newsroom can be loaded into the captioning software prior to broadcast;
- reviewing the quality of audio feeds;
- sharing best practices across teams to ensure that captioning errors which arise in one location can be avoided in other areas;
- placing particular focus on special programming whose characteristics can have an impact on automatic captioning quality, such as remote news where audio quality may be substandard or panel discussions that include cross talk and fast speech rates;
- training an in-house employee as a certified NER evaluator so that they can review additional programmes without depending on outside evaluators.

The experience of The Miracle Channel Associations is similar. In 2019–2020, their cloud-based version of the automatic captioning software EEG reached the 98% accuracy rate in 8/31 samples and 10/31 scored below 97.5%. For the 2020–2021 round, they decided to use the local server of the software (called Lexi Local) and to have a dedicated member of the staff compile glossaries with their phonetic spellings to improve the accuracy of the captions. This enabled them to reach the NER threshold consistently, with only 1/24 samples below par but pretty close to the minimum accuracy rate (97.92%).

Overall, Canadian broadcasters seem to have taken it upon themselves to improve their live captioning services to meet the new CRTC requirements. This improvement is reflected in the very high quality of the human captions and in the improvement of automatic captions, some of which are now finally reaching the NER threshold regularly with effective human help.

4. Final thoughts

The analysis presented in this article is, to our knowledge, the largest study on English human and automatic live captions available to date. It includes approximately 17,000 captions, that is, 798 minutes of live captions analysed with the NER model from 2018 to 2022: 388 minutes produced by human captioners using respeaking and stenography and 410 minutes produced by ASR.

Table 2

Average accuracy rates of all four case studies (human and automatic captions)

Case study	Human average (%)	Automatic average (%)	Year
Case Study 1 (Sky)	97 (2.5/10)	95.7 (0/10)	2018
Case Study 2 (Vitac)	98.8 (7/10)	96.3 (1/10)	2020
Case Study 3 (Enco)	99.4 (8.5/10)	96.3 (1.5/10)	2021
Case Study 4 (Canada)	98.9 (7/10)	-	2019–2021

As can be seen in Table 2, the quality of the human captions analysed here is overall very good and often excellent. The only exception is Case study 1, where respoken captions, with an average accuracy rate of 97%, do not manage to reach the NER threshold because of a series of unusually low-quality samples. This aside, the human captions in the other three case studies, all of them produced in North America, reach an average accuracy rate between 98.9% and 99.4% (7–8.5/10), higher than the 98.3% (6/10) obtained by the 78,000 captions produced by UK broadcasters between 2013 and 2015 for the Ofcom study (Romero-Fresco, 2016). It is difficult to pinpoint a single reason that can account for this higher level of accuracy. However, it is important to note that while almost all the captions in the Ofcom study were produced by respeaking, at least half of the North American captions analysed in this study were produced by stenographers. While more research comparing stenography to respeaking is still needed, steno captions would seem to yield higher accuracy rates than respoken captions (99.1% vs 98.6% in case study 2, that is, the difference between 6.5/10 and 7.5/10), mostly because the former feature fewer omissions than the latter. The downside, of course, is that steno captions often include an almost verbatim account of speech, which leads to very high presentation speeds that many viewers may find difficult to read (Romero-Fresco, 2011). This could be remedied if, as reported by some Canadian broadcasters (AMI, 2021), stenographers are trained to apply strategic editing that can reduce the captioning speed without compromising context or meaning. Respoken captions cannot normally be produced at the same speed as steno captions, which means that editing is an essential part of the job for respeakers. When this is done effectively, respoken captions (despite being perhaps slightly less accurate than steno captions) can strike a good balance between accuracy and readability, thus presenting a useful alternative to the fully verbatim and often very fast automatic captions.

As shown in Table 2, the accuracy of fully automatic captions has increased considerably over the course of this study, from 95.7% in case study 1 and 96.3% in case studies 2 and 3 (that is, a 0–0.5/10), conducted between 2018 and 2019, to some broadcasters consistently reaching the 98% threshold in case study 4, based on data obtained from Canada at the end of 2021. The steady improvement in automatic captions can also be seen in the poorest samples, the accuracy rates of which have also increased over time even though different engines were tested in our analysis. The lowest accuracy rate found in case study 1, conducted in 2018, was 93.5% (0/10). By 2021, in the second round of live captions analysed in Canada, the lowest accuracy rates obtained in the 31 samples analysed by Jim Pattison Broadcast Group and in the 24 samples analysed by The Miracle Channel Associations were 97.8% and 97.9%, respectively.

However, as noted by several Canadian broadcasters in their reports, if not aided, automatic captions still fall short of reaching the required accuracy threshold. Case study 3 illustrates this since it was also conducted in 2021 with one of the speech-recognition engines used by the Canadian broadcasters. In case study 3, the automatic captions were produced without aid of any kind and reached the 98% threshold only occasionally. Currently, automatic captions need help from a human operator or captioner who can tip the accuracy rate above the threshold by updating the glossaries and dictionaries in the software (adding commonly used words, eliminating others that are not relevant) and reviewing its accuracy and any technical aspect (i.e., quality of audio feed) that can improve its performance. In Canada, this has led a broadcaster such as The Miracle Channel Associations to attain an average accuracy rate of more than 98% consistently in 24 samples. These included long stretches of error-free captions that allow viewers to follow the content of the programme (one sample even reaches 98.7%, that is, 6.9/10).

However, it is important to note that most of these captions, as analysed here, still feature some of the usual errors produced by automatic captioning software. These include missing or misplaced periods and chevrons, incorrect spellings of proper nouns (particularly names and places), misspellings of compound words and errors with tenses and agreements. Furthermore, since automatic captions are verbatim and follow the programme speech rate, they are often delivered at high or very high speeds (Romero-Fresco & Alonso-Bacigalupe, 2022), which renders them challenging or impossible to follow for many viewers. This means that, even when they reach the threshold, the experience they provide for the viewers is significantly worse than that of human captions.

Interestingly, as mentioned in Section 2, although the accuracy of automatic captions in other languages (especially minority/minoritised languages) is lagging behind that of English, the results obtained in Canada pave the way to a more positive result in the coming years. We now know that, with the right investment and, at least for now, the right human help, fully automatic captions can reach the required accuracy threshold. As a matter of fact, the recent introduction of a new generation of AI-powered ASR engines such as Ursa, by Speechmatics, and Whisper, by OpenAI (the developers of ChatGPT), is likely to trigger very significant changes in the area of live captioning. The first analyses that we have conducted so far (Romero-Fresco & Fresno, 2023) show that the automatic captions produced by Ursa and Whisper for speeches (one speaker, fast speech rate, clear enunciation, good sound) and interviews (two speakers, spontaneous speech, overlapping turns, good sound) in English and Spanish (monolingual, that is, audio and captions in the same language) yield a NER accuracy rate of between 99.1% and 99.9%. This is as good as (and, in some cases, better than) the accuracy rate normally obtained by human captioners. The main improvement seems to lie in a significantly enhanced recognition of punctuation marks and spontaneous speech. However, further analyses are needed to confirm whether, as it seems, this may be a step change in live captioning and how the new engines are going to cope with hitherto unresolved problems such as reduction and speed. Since speech rates in live programmes are not likely to decrease and automatic editing does not seem feasible for now (and has not so far been a priority for developers of automatic captioning), automatic captions are likely to exclude viewers who cannot keep up with fast captions. This is a major issue that should not be overlooked.

On a different note, the results presented in this article may also have some implications for interlingual live captioning, that is, the provision of live captions involving language transfer. A case in point is the live speech-to-text and MT tool recently launched by the European Parliament (EP). In order to make its debates accessible, the EP issued an invitation to tender on 6 August 2019 to acquire a licence for a tool that could "automatically transcribe and translate parliamentary multilingual debates in real time" (DGT, 2019, p. 3). After some initial tests (reviewed by a team of external consultants that includes the first author of this article), the tool was implemented in 2022 in 10 core languages (English, German, French, Italian, Polish, Spanish, Greek, Romanian, Dutch and Portuguese). Any speech delivered in any of these languages is transcribed with ASR and then machine translated into the other nine languages. The tool brings together two fully automatic processes with no human revision, which means that the potential errors caused by ASR are added to those made by the MT technology. As could be expected, the first tests show that the interlingual live captions produced by this tool are still far from reaching the required quality threshold.

However, the data presented in this article point to an interesting alternative solution to the accessibility of debates at the EP. As is widely known, the EP has a team of highly skilled simultaneous interpreters who provide live oral translations from and to the 24 official languages of the European Union. Since the accuracy of ASR is steadily improving, it may be useful to consider an alternative interlingual live captioning workflow made up of simultaneous interpreting plus ASR. For a speech delivered in Spanish, for instance, the oral translation into English provided by the simultaneous interpreters at the EP would be converted into live English captions by ASR software. The translation job would thus be done by (highly skilled) professionals. Crucially, since simultaneous interpreting often involves condensation (as opposed to literal or word-for-word translation), the resulting captions may be edited and thus presented at more readable speeds than those produced by the fully automatic tool that is currently being tested by the EP.

A few issues may need to be dealt with when using this workflow, though. They include the fact that interpreting at the EP may sometimes involve pivot languages – if, for instance, Bulgarian into Spanish interpreters are not available, it is necessary to have Bulgarian into

English and then English into Spanish translations. This adds an extra layer that is likely to result in captions with a great deal of delay. Here, especially when EP debates are streamed online, it may be necessary to introduce a signal delay (around 5–7 seconds) so that the video of the session can be re-synchronised with the captions. Taking this into account and in the light of the results presented in this article, the EP has agreed to test this workflow, which has the potential to increase the accuracy and readability of the captions produced at the EP and in any other setting where live language transfer is required.

To conclude, and returning to the main theme of this article, it seems that, given the right conditions and at least in English, both human and automatic captions can now possibly provide viewers with access to live programmes and events. This has led some companies to offer a two-tier approach to access: their customers can choose between the affordable automatic captions (which involve more errors, faster speeds, and less delay) and the more expensive human captions (which come with fewer errors, lower speeds, and more delay).

Software developers may now be expected to keep improving the accuracy of automatic captions in English and, especially, to apply these developments to other languages, which will help to provide access to programmes and events that are still not being captioned. Furthermore, these improvements may help alleviate the financial burden that closed captioning poses for some broadcasters by allowing the combined use of human and ASR captions. In this scenario, broadcasters could choose to have specific programmes captioned by human beings, who frequently deliver very good or excellent captions; alternatively, they could rely on less accurate (although still acceptable) ASR captions for secondary programming.

As for human captioners, despite the possible threat posed by automatic captioning, their jobs are still essential, because, for the time being at least, only they can provide high-quality access that is both accurate and readable. It is only they who can place the viewers at the centre, ensuring that no one is left behind in the provision of access to live content.

Acknowledgement

This research has been conducted with the support of the Spanish Agencia Estatal de Investigación within the framework of the project “The Quality of Live Subtitling (QuaLiSub): A regional, national and international study” (AEI/10.13039/501100011033).

Romero-Fresco, P. & Fresno, N. (2023). Accuracy of automatic and human live captions in English. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *22*, 114–133.

5. References

2016 Working Group: EBG NER trial. (2018). *Final report*.

AMI (Accessible Media Inc.). (2021). *Annual closed captioning reports 2020-2021 – AMI-tv (CRTC 2018-319)*. https://crtc.gc.ca/eng/BCASTING/ann_rep/annualrp.htm#ami

Apone, T., Botkin, B., Brooks, M., & Goldberg, L. (2011). Caption accuracy metrics project research into automated error ranking of real-time captions in live television news programs. *September*, 1–16.

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. *Tradumàtica: Tecnologies de La Traducció*, *19*, 1–21. https://doi.org/10.5565/rev/tradumatica.276

CAD (Canadian Association of the Deaf). (2018). *Understanding User Responses to Live Closed Captioning in Canada*. http://www.livecaptioningcanada.ca/assets/User_Responses_Survey_Key_Findings_FINAL.pdf

Canadian Radio-Television and Telecommunications Commission (CRTC). (2015). *Broadcasting notice of consultation CRTC 2015-325*. https://crtc.gc.ca/eng/archive/2015/2015-325.htm

Canadian Radio-Television and Telecommunications Commission (CRTC). (2016). *Broadcasting Regulatory Policy CRTC 2016-435*. https://crtc.gc.ca/eng/archive/2019/2019-308-1.pdf

Canadian Radio-Television and Telecommunications Commission (CRTC). (2019). *Broadcasting Regulatory Policy CRTC 2019-308*. https://crtc.gc.ca/eng/archive/2019/2019-308.htm

DGT (Directorate-General for Translation). (2019). *Live Speech to Text and Machine Translation Tool for 24 Languages – Innovation Partnership – Specifications*. https://etendering.ted.europa.eu/cft/cft-display.html?cftId=5249

Dutka, Ł. (2022). *Live subtitling with respeaking in Polish: Technology, user expectations and quality assessment* [Unpublished doctoral dissertation]. University of Warsaw.

Fresno, N. (n.d.). Live captioning accuracy in English-language newscasts in the United States. *Universal Access in the Information Society*.

Fresno, N. (2019). Of bad hombres and nasty women: The quality of the live closed captioning in the 2016 US final presidential debate. *Perspectives: Studies in Translation Theory and Practice*, *27*(3), 350–366. https://doi.org/10.1080/0907676X.2018.1526960

Fresno, N. (2021). Live captioning accuracy in Spanish-language newscasts in the United States. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12769 LNCS*, 255–266. https://doi.org/10.1007/978-3-030-78095-1_19

Fresno, N., Romero-Fresco, P., & Rico-Vázquez, M. (2019, June 17-19). *The quality of live subtitling on Spanish television* [Conference presentation]. Media for All 8 Conference, Stockholm University, Sweden.

Fresno, N., Sepielak, K., & Krawczyk, M. (2021). Football for all: The quality of the live closed captioning in the Super Bowl LII. *Universal Access in the Information Society*, *20*(4), 729–740. https://doi.org/10.1007/s10209-020-00734-7

Ivarsson, J., & Carroll, M. (1998). *Subtitling*. TransEdit.

Jensema, C., McCann, R., & Ramsey, S. (1996). Closed-captioned television presentation speed and vocabulary. *American Annals of the Deaf*, *141*(4), 284–292. https://doi.org/10.1353/aad.2012.0377

Jim Pattison Broadcast Group. (2020). *Cover letter and report to CRTC*. https://crtc.gc.ca/eng/BCASTING/ann_rep/annualrp.htm#jim

Jordan, A. B., Albright, A., Branner, A., & Sullivan, J. (2003). *The state of closed captioning services in the United States: An assessment of quality, availability, and use*. The Annenberg Public Policy Center of the University of Pennsylvania. Report to the National Captioning Institute Foundation. https://dcmp.org/learn/static-assets/nadh136.pdf

Neves, J. (2005). *Audiovisual translation: Subtitling for the deaf and hard-of-hearing* [Unpublished doctoral dissertation]. University of Surrey-Roehampton.

Pérez Cernuda, C. (2022). *Subtitulado automático bilingüe: Una solución no tan sencilla*. Panorama Audiovisual.Com. https://www.panoramaaudiovisual.com/2022/10/25/subtitulado-automatico-bilingue-la-idea-es-sencilla-la-solucion-no-tanto/

Romero-Fresco, P. (2009). More haste less speed: Edited versus verbatim respoken subtitles. *Vigo International Journal of Applied Linguistics*, *6*, 109–133. https://revistas.uvigo.es/index.php/vial/article/view/33

Romero-Fresco, P. (2011). *Subtitling through speech recognition: Respeaking*. St. Jerome.

Romero-Fresco, P. (2016). Accessing communication: The quality of live subtitles in the UK. *Language and Communication*, *49*, 56–69. https://doi.org/10.1016/j.langcom.2016.06.001

Romero-Fresco, P., & Martínez, J. (2015). Accuracy rate in live subtitling: The NER model. In J. Díaz-Cintas & R. Baños-Piñero (Eds.), *Audiovisual translation in a global context: Mapping an ever-changing landscape* (pp. 28–50). Palgrave MacMillan. https://doi.org/10.1057/9781137552891_3

Romero-Fresco, P., & Alonso-Bacigalupe, L. (2022). An empirical analysis on the efficiency of five interlingual live subtitling workflows. *Xlinguae*, *2*, 3–13. https://doi.org/10.18355/XL.2022.15.02.01

Romero-Fresco, P., & Fresno, N. (2023, July 5-7). *AI and live captioning: Comparing the quality of automatic and human live captions in English* [Conference presentation]. Media for All 10 Conference, University of Antwerp, Belgium.

Stingray Group Inc. (2021). *Cover letter and report to CRTC*. https://crtc.gc.ca/eng/BCASTING/ann_rep/annualrp.htm#stingray

TLN (Telelatino Network Inc.). (2021). *Cover letter and report to CRTC*. https://crtc.gc.ca/eng/BCASTING/ann_rep/annualrp.htm#tln

[i] Also known as edition rate (Romero-Fresco & Martínez, 2015), the reduction rate accounts for the extent to which captions are or not a verbatim rendition of the speech in a given programme. Whereas live subtitles in the United States tend to be almost verbatim (Jensema et al., 1996), in Europe they vary from the near-verbatim UK subtitles to the more heavily edited subtitles in Spain or Switzerland (Romero-Fresco, 2009).

[ii] The NER model is based on WER (Word Error Rate), a tool that is often used to assess the accuracy of ASR. WER distinguishes between different types of error in the captions compared to the audio (substitutions, deletions and insertions) but it does not contemplate different degrees of severity or how these errors have an impact on the viewers' comprehension. The NER model classifies captions as "excellent" (accuracy rate above 99.5%), "very good" (99–99.49%), "good" (98.5–98.99%), "acceptable" (98–98.49%) and "substandard" (below 98%).

[iii] Sky wished to test two versions of the Google engine, one that delivered unpunctuated captions and another which included punctuation marks. Some of the captioning files that we received included punctuation while others did not. For two particular clips, we were provided with both the punctuated and the unpunctuated captions. For these two clips, we report on the punctuated version only because this one was more aligned to the captions created by Microsoft and also by the respeakers.

[iv] While we gathered information on the average accuracy rate per sample, we could not access the detailed NER assessment for all the automatic samples in our corpus of automatic closed captions. Therefore, we could not estimate the average accuracy rate for the automatic samples in case study 4.