HENGE DESIGN

13 June 2023

The Manager
Content and Platform Projects Section
Australian Communications and Media Authority
Via upload to https://www.acma.gov.au

Re: Proposal to remake the *Broadcasting Services (Television Captioning) Standard 2013*

Introduction

Since the advent of captioning in the early 1980s, Australian viewers have enjoyed captions of a very high level of quality on free-to-air television broadcasts[1], including comprehensive real-time captioning as well as the use of block captions during news and current affairs programs. This high quality[2] is due in large part to the cooperative approaches of Australian broadcasters in providing access to internal production resources such as newsroom computer systems and scheduling applications, allowing pre-scripted sections of news bulletins to be transferred into the captioning application and stenographic and respeaking dictionaries to be prepared with key words relevant to a program's subject matter.

Many of the Australian captioning regulations pertaining to quantities of captioning were legislated between 2000-2006, with the *Broadcasting Services (Television Captioning) Standard* added in 2013 to address requirements for the quality of captioning. The terms of the *Standard* were derived from extensive consultation with Deaf organisations, hard-of-hearing groups, educators of Deaf and hard-of-hearing students, captioning suppliers and broadcasters, to determine consensus on what attributes yielded the most meaningful and achievable access to Australian television. Fundamentally, the *Standard* has been effective in that it:

- reflects consumer demand for captioning quality to be at a level which facilitates a transparent viewing experience;
- identifies the components of captioning upon which quality is contingent.

But the *Standard* is not flawless. Issues include:

- the complaints-based process and highly subjective nature of determining when the *Standard* has been breached (or alternatively, at its most granular level the *Standard* is arguably being breached daily by every broadcaster);
- the inclusion of aspects which are outside the control of broadcasters, such as the display attributes of various television receivers;

[1] In 2015, The UK's OFCOM published its Fourth Report of *Measuring live subtitling quality*, in which the average accuracy rate measured using the NER Model at 98.55 was "the highest of all four sampling rounds." This compares to samples of live captioning of broadcasts by Australia's Seven and Nine networks over the period October 2017 to September 2022 averaging 99.39 and 99.47 respectively.
[2] ***Use and Experience of captioning** – Consumer research to support the ACMA's Captioning Quality Standard review* May 2023: "Satisfaction with the quality of live captioning was high across all broadcasting services…"

- there is no defined hierarchy of the impact of breaches, for instance incorrect words vs grammatical formatting;
- questionable categorising of aspects under the headings Readability, Accuracy and Comprehensibility;
- interpreting "distinct program segment" to be, for instance, a 45 second item in a news bulletin is overly punitive, and it is questionable whether this is what the wording of the *Standard* actually dictates.[3]

This submission will elaborate on these issues and possible resolutions in the responses to the questions posed in the *Consultation paper*.

Question 1: If the Standard were to be remade as currently drafted, would it be appropriate for it to be accompanied by an accompanying commitment to:

> support industry to further examine the introduction of a metric measurement model in the future, particularly given the likely faster evolution in captioning related technology in the coming years?

> provide further guidance on the interpretation of key elements of the Standard?

Why or why not?

Answer 1: While it is not clear how the wording of a legislated instrument could commit to supporting industry to further examine something, it is appropriate for the ACMA to play a role in supporting industry to further examine the use of a metric measurement model to assess the quality of broadcast captioning, however such an examination must ensure that incorporating a metric model into the *Standard* does not throw the baby out with the bath water.

The use of a metric measurement model is a double-edged sword in relation to legislated standards. On the one hand, an effective metric measurement model brings objectivity and repeatability to quality assessments, providing broadcasters and consumers with confidence in the terms required to meet the legislated requirements. On the other hand, a metric measurement model must set a threshold below which a breach is triggered. This may result in an overall deterioration of captioning quality if the threshold becomes the target and an expenditure-wary broadcaster aims to produce captioning of a quality that is just above the breach threshold. The uncertainty around the nebulous nature of the current *Standard* may be contributing to broadcasters and captioning suppliers aiming high to ensure overall compliance.

Captioning-related technology may have profound influences on methods of producing captions, but the fundamental aspects of quality remain the same: do the captions effectively convey the information provided in the audio of the program to a human viewer? The most influential technology on the captioning landscape at the moment is Automatic Speech Recognition (ASR). While promising and providing excellent results in the right circumstances, from a live caption quality perspective ASR captioning generally performs below respeaking, which performs below stenocaptioning, and which in turn performs below hybrid live captioning. If the question refers to the use of developing technologies in the task of assessing captioning quality, until ASR has been proven to comprehensively out-perform hybrid, steno and respeaking, using ASR technology as a

[3] It is arguable that individual stories within a news bulletin are not "unrelated to other program segments" in that they are related by the fact that they are all *news*.

tool to assess the quality of live captioning would be akin to asking a casino to regulate the gambling industry.

The ACMA should provide further guidance on the interpretation of key elements of the *Standard*, especially around the areas of defining “distinct program segment” boundaries and articulating the hierarchy of the importance of the various terms of the standard in conveying information to the viewer. For the sake of practicability in real broadcast environments, this should be done in consultation with the broadcasting industry.

Question 2: Is the clarification that broadcasters have indicated they would like about elements of the Standard best achieved through informal guidance rather than redrafting the Standard? Why or why not?

Answer 2: Both are required. Terms in which the interpretation or the hierarchy of affect is unclear require clarification through informal guidance and consultation, while items which are outside the broadcaster’s control should be redrafted or removed from the *Standard*. Revising the classifications such that items are categorised according to their impact on conveying information (i.e. a hierarchy of affect) would also be beneficial.

The interpretation of “distinct program segment” is a case in point concerning clarity. It is something that is easy to define until you actually try to apply it to the real world of broadcasting. For instance, a magazine-style current affairs program often comprises a few distinct stories – almost mini-documentaries – within its confines. These may clearly be considered “distinct program segments.” However, how does one partition a news bulletin? The ACMA has advised one broadcaster that “..the sports section in a news bulletin is not a single segment but several distinct program segments relating to different sports.” Does this mean that a story on the Rabbitohs vs the Broncos and another story on the international Rugby League World Cup form a single program segment? Does a story on the Formula 1 and a story on MotoGP constitute a single program segment (motor sports)? What wording in the *Standard* results in the demarcation being drawn at the code level and not the competition level, or expanded the other way to draw the line at *Sports* in general, as distinct from *Weather* or *Finance*? And if the boundary is deemed to be at the code level, how does one deal with a magazine-style program specialising in AFL, but containing several distinct stories about that sport? With a news bulletin, is it fair to say that a world round-up comprising six 45-second items constitutes six separate programs, any of which may result in the broadcaster breaching the *Standard*? How does the wording of the *Standard* preclude the round-up being considered a distinct program segment?

7(b)(i) is a requirement that is outside of the broadcaster’s control. Captions are broadcast in Australia as OP47 data, which at its core is Teletext and is by default monospaced white or coloured text on a black background. Caption authoring tools used in Australia conform to this. However, the television receiver is the thing that decodes this OP47 caption data and determines how it is displayed on the viewer’s screen. While most receivers comply and display the captions as they were authored, some receivers have been known to:

- ignore the black background, such that for instance white captions disappear into a white part of the screen;
- ignore text colour codes thereby destroying speaker identification information;
- use proportionally-spaced fonts thereby destroying caption position;
- ignore leading blank spaces thereby destroying caption position.

The *Standard* applies to broadcasters, but you can't penalise a broadcaster for something that is outside of their control.

Question 3: If the ACMA did provide guidance to broadcasters about the application and interpretation of the Standard (in addition to information in published investigation reports), what would be an appropriate mechanism to achieve this? Please provide details.

Answer 3: Published investigation reports provide information on how the ACMA interprets the *Standard*, but the lag time between the offending broadcast and the report – on occasions more than a year – means that the broadcaster remains ignorant of what caused the breach and may in fact continue to breach while in that state of ignorance. For the published investigation reports to be considered an effective mode of interpretive information, this lag time must be considerably reduced.

A mechanism that should be considered is periodical random assessment of the quality of captioning with prompt feedback to the broadcasters on any quality issues.

Question 4: In the current legislative context, would a metric model be an appropriate alternative to the Standard's current approach to measuring the quality of captioning? If so, why and how?

Answer 4: As discussed in Answer 1 above, a potential problem with enshrining a metric model into the *Standard* is the risk of a 'race to the bottom'. Using the NER Model[4] as an example, it sets the threshold of acceptability at 98.00[5]. This score was arrived at after polling (with academic rigour) Deaf and hard-of-hearing viewers in the United Kingdom to determine at what point they found the captioning useless and unacceptable. However, with some stakeholders, this score of 98.00 has been identified as "the target."[6] 98.00 should not be the bullseye that we are attempting to hit, it should be considered the absolute bottom of the range. Given that Australian viewers have been enjoying live captioning which averages around 99.40[1] on free-to-air television, any introduction of a metric model into the regulations must ensure that the overall quality does not deteriorate.

Ultimately, it is up to Australian viewers to determine the threshold of quality that they consider acceptable.

Question 5: What benefits, if any, would a metric model have for viewers compared to the Standard's current approach to measuring the quality of captioning? Would viewers who are concerned about the quality of captioning be able to determine whether a captioning service has met a metric measure while they are watching a television program? If so, how?

[4] *Accuracy Rate in Live Subtitling: The NER Model* Romero-Fresco and Perez. https://link.springer.com/chapter/10.1057/9781137552891_3

[5] The NER score calculation does not divide the number of idea unit errors by the total number of idea units in the sample but instead divides the number of idea unit errors by the total number of words and punctuation, therefore it is not a true percentage. While a score of 98.00 sounds high, on average, a score of 98.00 equates to over 40% of the idea units being compromised.

[6] For instance the graphic in https://www.apptek.com/post/the-quality-of-live-captions-accessibility-series-part-5. By way of contrast, it is noted that some captioning suppliers set internal quality targets at NER 99.00 or even higher for certain live genres.

Answer 5: It is difficult to determine what benefit viewers would enjoy under a metric model compared to the current *Standard*, especially given the high quality of live captioning currently enjoyed by Australia viewers[2]. Broadcasters and the suppliers of captioning services, on the other hand, would benefit from the certainty that a metric model brings to clearly defining the threshold at which a breach occurs.

It is worth noting here that the NER Model concentrates very much on viewer comprehension rather than captioner effort. This is most evident in the rules around awarding a 1.0 'Major' error weighting described in Answer 6 below: a captioner who kicks back and completely fails to caption the spoken "*Malcolm Turnbull is **now** the new leader…"* will be pinged only 0.5, while the captioner whose valiant attempt yields the "lying" caption *"Malcolm Turnbull is **not** the new leader…"* is hit twice as hard with a 1.0.

Correctly assigning error weightings under the NER Model requires some understanding of grammatical constructions such as independent and dependent phrases, modifying words such as adverbs and adjectives and – in the absence of being able to hear the audio of the program – access to an accurate verbatim transcription of the program. It would therefore be difficult for a viewer to perform an NER assessment of a program while they were watching it on the fly.

Question 6: What, if any, metric model would be the most appropriate to assess the quality of captioning? Please explain how it would address the BSA requirements of readability, accuracy and comprehensibility.

Answer 6: The NER Model is by far the best metric model available today, largely because its strict rules for assigning error weightings minimise subjectivity. This means that different assessors, employing the Model correctly, will arrive at very similar scores when assessing the same material.

Because of these strict rules, the *Consultation paper*'s statement "..an NER assessment includes some subjective assessments, particularly in attributing the weighting to be allocated to specific errors," is somewhat misleading. The NER rules are:

> A 0.0 'Correct Edition' weighting is assigned when spoken filler words such as *you know*, *really*, *I think* are purposely edited from the captions, or when a phrase is reworded (often with fewer words) but conveys the same meaning.
>
> A 0.25 'Minor' weighting is assigned to Recognition errors such as clear homophone misspellings e.g. *there*, *their*, *they're*, or to capitalisation or punctuation errors which lead to confusion, or to Edition errors such as the absence of an incidental modifying adjective, adverb or phrase (*"The assailant was hiding in the bedroom under the bed"* captioned as *"The assailant was hiding under the bed."*). A Minor error may disrupt the viewer but does not destroy meaning.
>
> A 0.5 'Standard' weighting is assigned to Edition errors where information is missing which would otherwise contribute to the viewer's understanding (*"There were road closures in Dunoon today,"* spoken but missing from the captions) or Recognition errors where a word is mistranslated such that it is clearly incorrect (*"..this famous pianist…"* represented in the captions as *"..this famous penis…"*). A 0.5 Recognition error is usually identified by the viewer as being nonsensical and therefore erroneous.

> A 1.0 ‘Major’ weighting is assigned to errors where the captions may appear to be correct to the viewer, but are actually providing misinformation, such as an incorrect number (*1,000* vs. *1,000,000*) or the mistranslation of *now (“Malcolm Turnbull is not the new leader…”* instead of *“Malcolm Turnbull is now the new leader...”*).

The NER Model clearly aims to measure how effective the captions are in conveying the information payload of the audio of a television program, and these weightings form a highly effective method of scoring the following items in the *Standard*:

> ***7 Readability***
> *(a)(ii) whether standard punctuation of printed English has been used in the captions to convey the way speech is delivered;*
>
> ***8 Accuracy***
> *(b)(i) whether spoken content has been captioned;*
> *(b)(iii) where it is not possible for the captions of spoken content to be verbatim, whether the captions reflect the actual meaning of the spoken content;*
> *(b)(vi) whether sound effects and/or music, material to understanding the program and not observable from the visual action, have been accurately described;*
>
> ***9 Comprehensibility***
> *(b)(i) whether the captions clearly identify and distinguish individual speakers, including off-screen and off-camera voices;*[7]
> *(b)(v) whether the words used in the captions have been spelt correctly;*
> *(b)(vi) where a word is not spelt correctly, whether the spelling provided nevertheless conveys the meaning of the actual word.*

The NER Model does not address aspects of captioning such as:

- synchronicity of the captions with the speech;
- the duration of caption display;
- positioning on the screen including avoiding onscreen graphics, lips, etc;
- breaking captions with grammatic sensibility and sympathy to shot and scene changes;
- number of rows;
- caption word rate vs speech word rate.

It is common practice, however, for NER assessors to provide commentary on these aspects of the program that they are assessing in addition to the numerical NER score.

Because the NER Model is such an effective method for assessing the conveyance of audio information in captioning, it should always be one of the tools in a quality assessor’s arsenal, even if it is not enshrined in regulations. Rather than engaging in protracted subjective deliberations over whether a phrase in the captions is effectively conveying meaning or not, the application of the weighting rules is quite a straightforward process and the assessment of a complete television program, including writing the report, can be done within one day.

[7] It is common practice for an NER assessor to log missing double chevrons or missing colour changes which signify a new speaker as an error.

The use of the NER Model internally could significantly improve the ACMA's turnaround time in reporting on breaches of the Standard.[8]

Question 7: Metric models used or considered overseas do not include details about the latency or synchronicity of captioning (although these are addressed in other elements of the legislative framework). Should these issues also be addressed by a standard dealing with the quality of captions?

Answer 7: Latency and synchronicity have significant impact on a viewer's ability to derive meaningful information from the captions, so it is imperative that they continue to be included in any quality standard. Errors in captioning often have a cumulative effect, so it is valid for the *Standard* to recognise issues that may be considered minor when compared to critical errors such as incorrect words or missing information.

The *Standard* currently provides no information on the hierarchy of its requirements. Instead, the requirements are categorised under the headings of *Readability*, *Accuracy* and *Comprehensibility*. In addition to some questionable allocations[9], it is difficult to see how these categories actually help viewers, broadcasters or caption suppliers in the pursuit of quality – even though the requirements listed within them are valuable.

To assist broadcasters in setting service levels with their caption supplier, and to assist caption suppliers in devising training and QC programs, it would be more useful to categorise the requirements according to their importance in effectively conveying the information payload of a television program's audio in the form of captioning. As a suggestion:

Critical

8(b)(i) whether the spoken content has been captioned;
8(b)(iii) where it is not possible for the captions of spoken content to be verbatim, whether the captions reflect the actual meaning of the spoken content;
8(b)(vi) whether sound effects and/or music, material to understanding the program and not observable from the visual action, have been accurately described;
9(b)(ii) whether the captions are displayed for a sufficient length of time to allow the viewer to read them and follow the action of the program;
9(b)(v) whether the words in the captions have been spelt correctly;

Required

7(b)(iii) whether standard punctuation of printed English has been used in the captions to convey the way speech is delivered;
7(b)(iv) whether the captions are positioned so as to avoid obscuring other on-screen text, any part of a speaker's face including the mouth and any other important visuals where possible;

[8] By way of example *ACMA Investigation Report BI-580* was published 14 months after the offending broadcast.

[9] 7(b)(iii) stipulating correct punctuation is arguably more of an "accuracy" or "comprehensibility" issue than it is "readability"; 8(b)(v) deals with describing the manner of a speaker's delivery, which seems better placed under "comprehensibility" than "accuracy"; 9(b)(ii)-(iv) and (viii)-(ix) deal with the timing and duration of the display of a caption, which seems better placed under "readability"; 9(b)(v) deals with correct spelling which seems better placed under "accuracy".

8(b)(v) whether the manner and tone of voice of speakers has been conveyed, where practical and material;
9(b)(i) whether the captions clearly identify and distinguish individual speakers, including off-screen and off-camera voices;
9(b)(vi) where a word is not spelt correctly, whether the spelling provided nevertheless conveys the meaning of the actual word;

Recommended

7(b)(ii) whether the caption lines end at natural linguistic breaks and reflect the natural flow and punctuation of a sentence, so each caption forms an understandable segment;
7(b)(v) whether the captions are not more than three lines in length;
8(b)(ii) whether the captions of spoken content are verbatim;
8(b)(iv) where the intended target audience of a program is children and the captions are not verbatim, the extent to which the captions take into account the intended audience;
9(b)(iii) and (iv) the extent to which the appearance/disappearance of the caption coincides with the onset/end of the speech of the corresponding speaker, sound effect or music;
9(b)(vii) whether explanatory captions are provided for long speechless pauses in the program;
9(b)(viii) and (ix) the extent to which a caption overruns a shot or scene change or to which the appearance or disappearance of the caption, as the case may be, coincides with the relevant shot or scene change.

It is worth noting in the above suggestion that the 'Recommended' section contains several items which are inherent to pre-recorded captions but often lacking in live captions. Such categorising would be a neat resolution to the at-first-glance paradoxical sections 130ZZA(2A) and (2B) of the *Broadcasting Services Act*.

Section 7(b)(i) is not included above as the broadcaster of closed caption data has no control over "whether colour and font is used in the captions in a way that makes them legible."[10] This is instead under the control of the television receiver. Section 7(b)(i) should be removed from the *Standard*.

Question 8: How should compliance with a metric model be measured and monitored?

Answer 8: The NER Model has proven to be the most effective metric method available for assessing the quality of live captioning. The challenge would be to settle on the threshold score at which a breach occurs. It is worth noting on this point that the *Broadcasting Services Act* "..does not authorise the ACMA to determine that a lower quality… of captioning service is acceptable for a kind of program or program material."[11] Consequently, unless this section of the *Act* is changed to allow different quality levels for live and pre-recorded programs[12], the breach threshold would have to apply equally to live captioning and to pre-recorded captioning. If set too low, it would effectively be saying that it is now alright for pre-recorded captioning to have as many errors as live captioning

[10] Teletext was initially employed in analog television broadcasts and the ease-of-viewing hierarchy of coloured text was then considered to be (from best to least): white, yellow, cyan, green. This distinction is no longer as critical with contemporary digital receivers.
[11] *Broadcasting Services Act 1992* section 130ZZA(2B)
[12] It should also be noted that some genres of live programming are also more difficult to caption than others, e.g. an election debate vs a news bulletin, or live sports vs current affairs.

does. If set too high, it would represent a goal for live captioning that is impossible to reach with current practices.

The use of the NER Model would be of great benefit when it comes to monitoring the quality of captioning, especially around the critical requirements of whether spoken content has been captioned and whether the captions reflect the actual meaning of the spoken content[13]. This is because the clear error weighting rules create an objective, level playing field for all broadcasters and facilitate efficient processing of assessments. It is worth noting here that with most programs of less than one hour duration – after first viewing the program in its entirety to make a rough assessment of the consistency of the caption quality – assessing a 10-minute sample from that program will often yield a score that is indicative of the quality of the entire program, precluding the need to perform a comprehensive assessment of the entire program.

A monitoring program could include a periodic random sampling of live programs subjected to NER assessment as well as anecdotal comments on degrees of latency, synchronicity, positioning, duration, grammatical breaks and reading rates.

Section 130ZZA(2B) of the *Act* notwithstanding, it is inevitable that some errors will occur sometimes with live captioning. Rather than penalising a broadcaster with a breach as the result of a single complaint, it would be better to augment the complaints process with regular random assessments (and timely feedback to the broadcaster) and for the ACMA to levy a breach only when a broadcaster is shown to be recalcitrant, rather than having unintentionally enacting a sporadic error.

Question 9: What arrangements would need to be in place to provide confidence in the results of a trial of a metric model?

Answer 9: Engagement by a range of Deaf and hard-of-hearing viewers in a comprehensive trial is imperative. The trial would need to show examples of multiple genres (sports, news, current affairs, talk shows, live entertainment) each captioned to differing quality levels (this is what a 99.50 looks like, this is what a 99.00 looks like, this is a 98.50, a 98.00, a 97.50…) to determine the acceptability threshold for the Australian viewer environment.

Any trial would also need engagement by the broadcasting industry, especially around punitive issues. Currently there is a degree of subjectivity that the ACMA can bring to assessing a viewer's complaint. Would this change in a metric assessment environment? For instance would penalties be on a strictly-enforced scale, inversely proportional to the NER score of the assessed program?

A trial should be overseen with statistical and academic rigour to ensure that the results are honestly reflective of the broad base of captioning users. Perhaps stating the obvious, it would also be necessary to involve experienced practitioners of the metric model that is being trialled in assessing the sample programs.

[13] *Broadcasting Services (Television Captioning) Standard 2013* sections 8(b)(i) and (iii)

In Conclusion

Given the high quality of live captioning presently provided on Australian free-to-air television, it is safe to say that the *Standard* in its current form does nothing to harm the quality of captioning. The contents of its requirements are a comprehensive compendium of what makes for good captioning, however the categorising of these requirements could be improved. The complaints-based approach also needs attention in that a channel with more viewers is more likely to suffer complaints than a channel with less viewers, resulting in, for instance, a subscription broadcaster of live baseball getting away with significantly lower quality captioning than a free-to-air broadcaster of the NRL.

The review of the *Standard* is a welcomed initiative, and the ACMA is encouraged to ensure that any changes to it are done in the interest of continual improvement.

Robert Scott
robert@hengedesign.com