Caption quality:

International approaches to standards and measurement

SPONSORED BY

Contents

About Media Access Australia

Media Access Australia is Australia's only independent not-for-profit organisation devoted to increasing access to media for people with a disability.

We promote inclusion by providing information and expertise on the accessibility of mainstream technologies to government, industry, educators, consumer organisations and individuals.

We work as a catalyst for change across television, video, cinema, the arts, education, digital technology and online media, with a primary focus on people who are blind or vision impaired, or Deaf or hearing impaired.

Media Access Australia grew out of the Australian Caption Centre (ACC), which was founded in 1982. As the ACC we provided captioning services for all Australian television networks, as well as the captioning of live theatre, videos and DVDs. The captioning and other commercial operations of the ACC were sold to Red Bee Media in 2006.

About the author

Chris Mikul is Media Access Australia's Project Manager for television. He has been involved in access for over twenty years, and was formerly National Production Manager at Media Access Australia's predecessor, the Australian Caption Centre.

Published 25 March 2014

Media Access Australia
616-620 Harris Street
Ultimo NSW 2007
Australia
Email: chris.mikul@mediaaccess.org.au
Website: mediaaccess.org.au
Telephone: +61 2 9212 6242

This work can be cited as: Mikul, C. 2014, *Caption Quality: International approaches to standards and measurement,* Media Access Australia, Sydney.

Introduction to Media Access Australia's white paper series

By Alex Varley, Chief Executive Officer, Media Access Australia

This is the first in what will be an ongoing series of white papers produced by MAA covering important and topical issues in access to information and media through technology. We see the white paper series as integral to MAA fulfilling its role as a catalyst for change and advocate of practical approaches to these issues. Since our creation in 2006 we have carved out a reputation as an independent, credible source of information and advice on a range of media access issues. This has not been achieved in isolation, with MAA acting as a form of lofty think tank or research institute, but through engagement with and understanding of the media industry, suppliers, technology experts, governments and ultimately, the consumers who are trying to use their products and services. As such, we have always taken a real-world view to expanding access. The issues are often complex and interweave policy desires, market-driven solutions, rapidly evolving technologies, costs and revenue opportunities, and understanding the needs and behaviours of audiences. Put bluntly, there are no simple, magic bullet solutions to most access issues.

We are also careful to ensure that we draw from the wealth of experience and information in different parts of the world. The issues are fundamentally the same, but often the unique circumstances of a location will throw up interesting and different perspectives that we can all learn from. It is in this spirit that these white papers are created. They are not neat bags of evidence leading to an all-encompassing conclusion, but are designed to encourage discussion and debate and to gain a better understanding of the range of factors and nuances that make up an issue.

The white papers will be eclectic in their subject choice, reflecting the vast range of access topics that we deal with. The choice for our first paper was driven by the recent increase in interest in, and discussion of, caption quality worldwide. With regulators, providers and their clients demonstrating willingness to explore different approaches, we felt a review of these approaches was timely.

This paper provides a snapshot of caption quality standards which are currently in place around the world. Most of these standards have been formulated by regulators and are mandatory for broadcasters. The paper compares and contrasts these various standards, and looks at some of the academic work which has been done around captioning.

The paper also looks at how television captioning standards and techniques have evolved since captioning began in the 1970s, and evaluates some models which have been developed to

measure caption quality. The ultimate aim of the paper is to promote a discussion of caption quality and suggest practical ways in which it can be maximised for consumers.

Media Access Australia has a well-established range of relationships with access suppliers, governments, regulators, consumers and media organisations across the world and this is reflected in Red Bee Media's sponsorship of this paper. We thank them for their generous support.

Foreword

By Chris Howe, Managing Director, Red Bee Media Australia

The issue of caption quality standards and how to measure them has been a key area of focus and debate in our industry and rightly so. Captioning and other access services are a fundamentally important element in broadcast and content publishing services. Captioning provides access to critically important news and information services, and provides a vastly enriched experience for millions of consumers. It is also developing as an important tool for broadcasters and publishers to enhance the ability to search and discover their content on increasingly diverse publishing platforms.

The focus on the importance of access services by various consumer groups, lobby groups, broadcasters and government bodies has led to an increase in awareness and a big increase in the volume of captioning. This includes the proliferation of live captioning output and the inherent challenges this brings for captioning service providers. These changes have led to an increased focus on caption quality by both consumers and regulators to address the rightful expectations of the audience for accuracy and timeliness of information.

Red Bee Media Australia welcomes this discussion and is very pleased to support the important work of Media Access Australia. This white paper is a timely and substantial contribution to the international discussion about caption quality standards. Its recommendations provide valuable input to the evolution of a global framework for these standards.

Caption standards benefit consumers by promoting a basic level of quality and consistency in the presentation of captions. They benefit regulators by creating a benchmark for use against consumer complaints. The same benchmark benefits broadcasters by giving them a yardstick against which to measure suppliers.

Broadcasters and content producers increasingly address a global market of consumers. The creation of caption standards necessarily requires consistency of approach – and best practice – across worldwide markets.

Red Bee Media Australia is proactively investing in the tools, systems and applications, as well as the training and skills to ensure ongoing improvements in accuracy and reduced latency, delivering the highest quality captioning services and managed services.

Red Bee Media Australia commends the work done by Media Access Australia in informing this important debate, and looks forward to engaging with stakeholders both here and in our other markets in further discussion about this important issue.

1. Executive summary

- Caption standards benefit consumers by promoting a basic level of quality, and ensuring consistency in the presentation of captions. They benefit regulators by providing benchmarks they can use when dealing with consumer complaints, inform broadcasters on the quality they should expect from their suppliers, and ensure that suppliers provide a product that will satisfy consumers.
- As levels of live captioning have risen in recent years, there has been an increasing focus on caption quality by consumers and regulators. The real time component of live captioning was initially performed by highly-trained stenocaptioners. In the last few years, however, stenocaptioning has been augmented, and in some European countries overtaken, by respeaking captioning, where captioners respeak dialogue and speech recognition software converts it to captions.
- Captioning systems offer a variety of ways that news can be captioned. The 'hybrid' method involves a combination of live captioning (stenocaptioning or respeaking) and captions pre-prepared from scripts and video. Live captions can also be converted to blocks of text for transmission.
- Binding national caption standards have been adopted in the UK, Canada, France, Spain and Australia. While these broadly cover the same areas, the details often differ. Public broadcasters (e.g. in Germany) may also have their own standards.
- There is considerable debate around the world about issues such as appropriate caption word speeds, accuracy rates and acceptable time lag for live captions.
- The British regulator Ofcom has been the most proactive about live captioning quality, launching a two-year initiative in 2013, with broadcasters required to provide accuracy measurements for sample live-captioned programs.
- A number of models have been developed to measure caption accuracy, the most recent being the NER Model, which has been specifically designed to evaluate respeaking captioning, and has now been adopted in several countries.

In conclusion, Media Access Australia makes the following recommendations.

1. Greater discussion across national boundaries and information sharing between regulators about their approaches will help standardise high quality.
2. That the move to towards practical experimentation (such as Ofcom's approaches) and field-testing of different approaches and evaluation models be encouraged.
3. That consumers are involved in decision-making about what works for them (e.g. the trade-off between time lag and block captions for live programs).
4. That the use of the hybrid method of captioning be encouraged for news and current affairs programs.

2. Introduction

In recent years, the quality of closed captions for the Deaf and hearing impaired (or hard of hearing) has become of an area of increasing concern for both caption users and industry regulators. Much of this concern can be traced to large increases in the volume of programs being captioned (driven

by consumer demand and quotas imposed on broadcasters by regulators), and in particular an increase in live captioning. This is generally performed by stenocaptioners or, increasingly, captioners using speech recognition software (known as 'respeakers'). For viewers, captions produced by stenocaptioning or respeaking have certain inherent drawbacks. There is an inevitable time lag between the audio and the captions appearing on screen, they are typically presented as 'scrolling' captions which appear one word at a time, they may be too fast to be easily read, and they are prone to errors. In some cases, the errors may be so bad that the captions are essentially useless. While viewers may dislike live captions, the fact remains that there are many programs, both genuinely live programs, and programs delivered very close to broadcast time, which can realistically only be captioned live.

In Europe in particular, a large number of academic papers have been published in the last few years which analyse various aspects of captioning, and caption quality has been a topic discussed at conferences such as the biennial Languages and the Media Conference in Berlin. A number of models have been put forward to measure caption quality, but this is an area fraught with difficulties. Not all errors are of the same magnitude. Some are negligible, while a single incorrect word can make an entire news story incomprehensible or, worse, change its meaning completely. A simple tally of the number of errors in a caption file may therefore provide little meaningful information about the quality of the captions.

One of the basic principles of captioning is that, as far as possible, all dialogue in a program should be captioned, but live captioners must routinely make decisions about what should be left out, and these deletions often serve to make the captions much more comprehensible for the viewer. The appropriate speed for captions, on pre-recorded and live programs, is another issue that has been debated since captioning began in the 1980s.

This paper provides a snapshot of some of the quality caption standards which are currently in place around the world, and evaluates the models which have been developed to measure quality. It also looks at the evolution of live captioning techniques, which have generally served to streamline production, but with the potential danger of reduced quality. Finally, it makes recommendations regarding approaches to standards in Australia and worldwide.

3. What are caption quality standards?

Caption standards may be written by caption suppliers, their clients, consumer representative groups or government bodies which regulate the television industry. These standards will usually, but not always, cover the following areas:

- Accuracy: Captions should reflect the dialogue and be correctly spelled and punctuated.
- Synchronisation: Captions should as far as possible be synchronised with dialogue (although there is generally an acknowledgement that this may not be possible for live programs).
- Word speed (also called presentation rate): For adult programs, this is generally set at around the 150 -180 word-per-minute mark, or verbatim.
- Colouring and positioning: Captions are often coloured and positioned to indicate speakers.
- Standards relating to the presentation of captions (number of lines, splitting of captions, breaks within captions, etc).
- Standards relating to the captioning of music, sound effects, etc.

- A stipulation that captions should not obscure important visual elements on the screen, e.g. faces or supers giving names.

Caption standards benefit consumers because they promote a basic level of quality, while also ensuring there is consistency in the way that captions are presented. The best captions are those which not only convey all the essential information on the audio track, but also indicate clearly who is speaking, including when people are off-screen. If the rules for identifying speakers, indicating sounds effects and so on are kept simple and used consistently, caption users will get to know them and spend less time looking at the captions (and perhaps puzzling over them) and more time watching the visuals and enjoying the program.

Caption standards benefit regulators because they provide benchmarks that can be used when evaluating complaints from the public about caption quality. For broadcasters, they set the level of quality they should expect and demand from their suppliers. And for suppliers, they are the key to creating a consistent product which will satisfy as many consumers as possible and allow competition on a level playing field.

There is, however, a danger in making caption standards too restrictive. An absolutely rigid focus on a particular word speed, for example, may add to captioning production costs without providing that much benefit to the consumer. And applying standards which do not conform with those of other countries will restrict the ability to import and utilise caption files produced elsewhere, and again add to production costs. The unique colouring standards used in France (see section 5.4) could be said to be an example of this.

4. The evolution of live captioning techniques

Much of the current debate about caption standards is focused on issues related to live captioning, and the need for a practical, cost-effective way to measure caption quality that will satisfy the needs of regulators, caption suppliers and consumers.

Live captioning of news programs by stenocaptioners began in the US in the 1980s. It was first used in the UK in 1990, and in Australia in 1993. Stenocaptioners use a stenographic keyboard to create captions as a program goes to air. The captions appear on screen one word at a time, and there is an inevitable delay between the audio and the captions appearing of three seconds or more. The quality of stenocaptioning depends on a number of factors, including: the training and skills of the stenocaptioner, their familiarity with the material they are captioning, and the amount of preparation time they have been given prior to the broadcast. Stenocaptioners are able to load shortforms of names and phrases into their software's 'dictionary', enabling these to be written with two or three keystrokes.

Respeaking captioning was first used in the UK for sports programs on the BBC in 2001, and the Australian Caption Centre introduced it to Australian television in 2003. In this method, a captioner respeaks the dialogue of a program – while also adding punctuation and colouring to identify different speakers – into a microphone as it goes to air, and speech recognition software

such as Dragon or ViaVoice converts this to captions. As with stenocaptioning, the quality of the captions produced using respeaking depends on the skills of the respeaker, the quality of the software they are using, and the amount of preparation time they have.

To television viewers, stenocaptioning and respeaking captioning look much the same. With both methods, captions generally appear on screen one word at a time, and there is a similar time delay. The chief difference between them is that when a stenocaptioner's software fails to recognise a word, a string of random letters can appear on the screen, while speech recognition software will always attempt to create an actual word (even though it may be completely wrong).

Stenocaptioners typically train for a minimum of two years before they obtain the skills necessary to perform live captioning, and there has always been a shortage of them. Respeaking captioners – provided that they have an aptitude for the task – can be trained in a matter of months, and they are not as highly paid as stenocaptioners.

As caption levels have risen in the last few years in Australia and other countries, caption suppliers have increasingly had to turn to respeaking (with, typically, two respeakers working in tandem to caption a program, replacing one stenocaptioner). While stenocaptioning remains the most common method of live captioning in Australia, respeaking now accounts for most of the live captioning done in the UK, while news programs in France and Germany are exclusively captioned using respeaking.[1]

A number of software systems have also been developed which aid in the captioning of news bulletins and other live or near-live programs, including Screen Systems' WinCAPS and, more recently, Red Bee Media's Subito system, which is now in use in Australia. These systems can be used to perform what has been called the 'hybrid' method of news captioning. This involves captioners pre-preparing caption files as much as possible using whatever scripts and video elements are available prior to broadcast. The captions are keyed out as the program goes to air, either as scrolling captions (which appear onscreen one word at a time), or as block captions (which appear one or two lines at a time). The latter will look to viewers much the same as fully timed and edited captions on pre-recorded programs, although the timing may not be as precise. Only the unscripted elements of a news bulletin, such as a cross to a reporter in the field, are captioned live by a stenocaptioner or respeaker. It should be noted that the hybrid model works best when there is full cooperation between caption suppliers and broadcasters, with the latter sharing resources and allowing captioning to be integrated with their systems.

In the UK, Red Bee Media performs hybrid captioning for Channel 4, Sky and the BBC, as well as in France. In Australia, it was used on all networks prior to 2006, with the pre-prepared captions always sent as block captions. Currently, only the Seven network retains this hybrid model with all block captions. On the Nine network, captions prepared from scripts are sent as scrolling captions, while there is some preparation from scripts on SBS and the Ten network (generally sent as scrolling, but occasionally in blocks). The ABC has an all-live model using stenocaptioners only.

[1] Personal communication from David Padmore, Director of Access Services at Red Bee Media.

News captioning systems like WinCAPS can integrate the captioning software with the broadcaster's newsroom systems. Captioners are able to import scripts as they are written which are automatically converted into block captions and edited to ensure accuracy before they are transmitted.

Red Bee Media's recently developed Subito system also has the capability to take live captions, created by a stenocaptioner or respeaker, and convert them to block captions. The conversion to blocks also inevitably adds to the delay in the captions reaching the screen, but it may be that consumers will accept this in return for the greater readability of block captions. Red Bee Media Australia has told Media Access Australia that it will be testing this feature on Subito to determine its feasibility. Subito will also allow scripts to be automatically converted to block captions.

In general, the overall quality of live captioning is gradually improving, thanks partly to more efficient work practices, but particularly due to improvements in speech recognition and stenocaptioning technology. However, much work remains to be done in this area.[2]

5. Caption quality standards around the world

5.1 United Kingdom

5.1.1 Ofcom's Code on Television Access Services

Captioning requirements for broadcasters are set out in Ofcom's *Code on Television Access Services*[3]. As well as setting quotas for subtitling (the British term for captioning), signing and audio description, the code sets out recommendations for best practice in these services.

The code states that:

> *…pre-prepared block subtitles are the best approach to providing accurate, easily legible and well-synchronised subtitles and should be used for pre-recorded programmes… When scrolling subtitles need to be used, any scripted material should be used for advance preparation. In addition to achieving the highest possible levels of accuracy and synchronisation, live subtitles should flow continuously and smoothly.* (pp. 11-12)

On the subject of word speed, it states:

> *..the speed should not normally exceed 160 to 180 words per minutes for pre-recorded programmes. Although it may not be practicable to restrict the speed of subtitles for all live programmes, commissioning editors and producers should be aware that dialogue which would require subtitles faster than 200 wpm would be difficult for many viewers to follow…*

[2] For a discussion of these issues, see Red Bee Media's paper 'Subtitling on British Television: the Remaining Challenges', which can be downloaded from: http://www.redbeemedia.com/insights/subtitling-british-television-remaining-challenges

[3] http://stakeholders.ofcom.org.uk/binaries/broadcast/other-codes/ctas.pdf

> *Slower speed and more heavily edited subtitles are appropriate for young children, though care should be taken to ensure that these are accurate and grammatical, as children and parents use subtitles in developing literacy skills.* (p. 12)

Subtitles should be synchronised as much as possible.

> *In live programmes, the aim should be to keep the inevitable delay in subtitle presentation to the minimum (no more than 3 seconds) consistent with accurate presentation of what is being said.* (p. 12)

Other quality issues that the code covers include:

- Layout and line-breaks
- Description of all relevant non-speech information
- Different colours for different speakers
- Accuracy

5.1.2 Ofcom's live caption quality initiative

In May 2013, Ofcom issued a consultation paper[4] and requested submissions from interested parties regarding the quality of live subtitling and how to improve it. In doing this, Ofcom was fulfilling its obligations under Section 3 of the *Communications Act* which states that, in carrying out its duty to citizens and consumers, it should have regard to the needs of persons with disabilities. Ofcom identified the four key dimensions of live caption quality as latency (time lag), accuracy, intermittent subtitles and presentation (scrolling or block captions).

In October 2013, Ofcom released a statement, *Measuring the quality of live subtitling*[5], setting out its decisions. Broadcasters will be required, at six monthly intervals for the next two years, to measure the quality of subtitling on samples from three genres of programming: news, entertainment and chat shows. The dimensions of quality to be measured are:

- The average speed of captioning
- The average latency and range of latencies
- The number and type of errors

Ofcom is hoping to include the first round of measurements in its access services report due in spring 2014. Ofcom will wait until the measurement exercise is completed before deciding whether to set targets for speed, latency and accuracy.

Ofcom has chosen the NER Model (see section 6.3) as the one which broadcasters should use to measure caption quality. A team from Roehampton University, where the model was developed, will validate their measurements. Ofcom will also use the two-year period to evaluate the effectiveness of the NER Model.

Ofcom's positions on other issues raised in its consultation paper were as follows:

---

[4] http://stakeholders.ofcom.org.uk/binaries/consultations/subtitling/summary/subtitling.pdf
[5] http://stakeholders.ofcom.org.uk/binaries/consultations/subtitling/statement/qos-statement.pdf

- Ofcom noted that broadcasters were strongly opposed to introducing a delay in live transmissions of 30 seconds or less to reduce the time lag in live subtitles. It will continue to maintain a dialogue with broadcasters on this issue.
- Ofcom has asked broadcasters to report in January 2014 which late-delivered programs had to be subtitled live between July and December 2013. The information will be including in the first access services report due in spring 2014.
- In their submissions, viewers made it clear that block captions were superior to scrolling captions, but not at the expense of increased latency. Ofcom suggests there is still merit in broadcasters experimenting with turning live captions into block captions to see if the increase in latency can be minimised, and the results tested with viewers.

5.1.3 The BBC guidelines

The BBC's *Online Subtitling Editorial Guidelines V1.1*[6] were issued in 2009. They include detailed instruction for writing captions, with an emphasis on how the language should be reduced. It states that 'since people generally speak much faster than the text of their speech can be read, it is almost always necessary to edit the speech.' The word speed for pre-recorded programs 'should not normally exceed 140 words per minute', although a higher rate of 180 wpm is permitted 'in exceptional circumstances'.

The guidelines also set timings for the BBC's two children's channels, CBeebies (4-5 seconds for 1 line, 8-10 seconds for 2 lines) and CBBC (3.35 seconds for 1 line, 6-7 seconds for 2 lines.)

5.2 Canada

5.2.1 The CRTC's English-language and French-language caption quality standards

In 2007 the Canadian Radio-television and Communications Commission (CRTC) introduced a new policy that stated that all programs apart from promos and commercials had to be captioned, and instructed the Canadian Association of Broadcasters to establish caption working groups for the English-language and French-language markets. After draft policies were released for public comment in 2011, the *Quality standards for French-language closed captioning*[7] were adopted on 1 December 2011, and the *Quality standards for English-language closed captioning*[8] were adopted on 5 July 2012.

The English-language and French-language standards generally agree, although there are some differences. For live captioning, broadcasters must reach an accuracy rate of at least 95 per cent (in the English-language standard) or 85 per cent (in the French-language standard), averaged over the program.

The method for calculating this in both standards is as follows:

---

[6] http://www.bbc.co.uk/guidelines/futuremedia/accessibility/subtitling_guides/online_sub_editorial_guidelines_vs1_1.pdf
[7] http://www.crtc.gc.ca/eng/archive/2011/2011-741.htm
[8] http://www.crtc.gc.ca/eng/archive/2012/2012-362.htm

$$\% \text{ of accuracy} = \frac{N - Sup - Sub - I}{N} \times 100$$

N: Number of words in audio
Sup: Number of supressed words (words in audio absent from captions)
Sub: Number of substituted words (words in audio replaced with other words in captions)
I: Number of inserted words (words not in audio but in captions)

For live programming, lag time between audio and captions must not exceed six seconds (in the English-language standard) or five seconds (in the French-language standard) averaged over the program.

The English-language standard states that:

> *Pop-on* [i.e. block] *captions are to be used for all new Canadian pre-recorded programming. Pre-recorded programs are those that have been delivered in their entirety – lacking only the closed captioning information – 96 hours before they are to be broadcast.*

Regarding the appropriate speed for captioning, both standards state that:

> *Captioning must be verbatim representations of the audio, regardless of the age of the target audience. Speech must only be edited as a last resort, when technical limitations or time and space restriction will not accommodate all of the spoken words at an appropriate presentation rate.*

However, the English language standard adds that

> *For the purpose of the standard, 'appropriate presentation rate' is defined as 120-130 words per minute for children's programming.*

This would seem to contradict the first statement.

The other areas covered in the standards are:

- For pre-recorded programming, captions must target an accuracy rate of 100 per cent.
- Broadcasters must calculate the accuracy rate for two programs containing live content each month, and provide information to the CRTC about their efforts to improve their accuracy rate every two years.
- For programs which are rebroadcast after initially airing live, caption errors must be corrected in Category 1 (News) and Category 2 (Reporting and Actualities programs) if the time between the original broadcast and rebroadcast is equal to at least two times the duration of the program. For all other programs, errors must be corrected if the program is rebroadcast more than 24 hours after the end of the original broadcast.
- Captions must be positioned to avoid covering action, visual elements or information required to understand the message, but if this is not possible, captions take precedence.
- Emergency alerts must be captioned with the exception of those issued by the National Alert Aggregation and Dissemination system.

- The French-language standard states that hyphens or chevrons must be used consistently to indicate that a different person is speaking, even when captions are positioned.

5.2.2. The CAB's caption standards handbook

In 2003, the Canadian Association of Broadcasters (CAB) released the first version of a handbook, *Closed Captioning Standards and Protocol for Canadian English Language Broadcasters*, with the latest version released in August 2012.[9] The latter includes all the CRTC's mandatory standards, along with other standards covering the finer details of captioning, including line spacing, structuring and shape of captions, the format for writing song lyrics, and so on. In this regard, it resembles the standards documents captioning suppliers typically put together for in-house use.

In the introduction, the CAB states that the handbook was initially produced because a lack of comprehensive standards meant that 'formal training of caption suppliers has not been possible,' so that 'captioning styles vary from supplier to supplier, sometimes frustrating the viewer'. The handbook:

> *…is designed to establish English language closed captioning standards acceptable to key stakeholders: the caption consumers, the caption creators, and the broadcasting industry, including private, public and educational television programming services. It is intended as the mandatory guide to Canadian English language closed captioning for television.* (p. 1)

The section on 'Guidelines for On-Line Real-Time Captions' states:

> *Real-time captioning is not well suited to dramas, movies, sitcoms, music videos or documentaries, and its use for these types of programming is discouraged.* (p. 23)

However, the section on 'Guidelines for Offline Roll-Up Captions' states:

> *Roll-up captions are to be used for all live programming such as news and sports and pre-recorded programming in instances where the amount of time between delivery and airing of the pre-recorded program is not sufficient to permit the production of pop-on captions (i.e. less than 96 hours).* (p. 21)

It should be noted that the intention in this section actually differs significantly from the CRTC's intention in its English-language standards. The CRTC's stipulation that any pre-recorded program delivered more than 96 hours before broadcast must have pop-on captions has become, in the CAB's handbook, a stipulation that any pre-recorded program delivered less than 96 hours before broadcast *must have roll-up captions*. If followed to the letter, this would lead to more programs having roll-up captions than necessary, resulting in a reduction in caption quality.

5.3 United States

In the US, captioning is required under the *Code of Federal Regulations*. This is regulated by the Federal Communications Commission (FCC) which draws on specific powers in Part 79.1 of the

[9] http://www.cab-acr.ca/english/social/captioning/cc_standards.pdf

code[10], and broader powers under the *Communications Act 1934*. None of this legislation makes any reference to caption quality.

Following a July 2004 petition filed by several advocacy groups including Telecommunications for the Deaf 'to establish additional enforcement mechanisms to better implement the captioning rules, and to establish captioning quality standards to ensure high quality and reliable closed captioning'[11], the FCC received 1,600 submissions in response.

On 20 February 2014, the FCC voted unanimously to approve new, comprehensive rules[12] to ensure that closed captioning on TV is of the highest possible quality.

The FCC's new rules will require that captions be:

- Accurate: Captions must match the spoken words in the dialogue and convey background noises and other sounds to the fullest extent possible.
- Synchronous: Captions must coincide with their corresponding spoken words and sounds to the greatest extent possible and must be displayed on the screen at a speed that can be read by viewers.
- Complete: Captions must run from the beginning to the end of the program to the fullest extent possible.
- Properly placed: Captions should not block other important visual content on the screen, overlap one another, or run off the edge of the video screen.

Captions for pre-recorded programs will be expected to comply fully with the above rules, but the FCC recognises that they will be more difficult to achieve on live and near-live programs.

The FCC's order also includes best practice guidelines for video programmers and caption suppliers. These include video programmers providing caption suppliers with advance scripts, proper names and song lyrics for live programs. Caption suppliers will need to ensure proper training and supervision of their captioners, and that their systems are functional to prevent interruptions to the caption service.

The FCC also put on notice that it will be looking at the question of who is responsible for caption quality, and whether quantitative standards are also necessary.

Other captioning quality standards exist in the US, the most prominent of which are the Described and Captioned Media Program's *Captioning Key: Guidelines and Preferred Techniques*[13] (DCMP Captioning Key) and *Captioning FAQ*[14] by WGBH's Media Access Group. Some of the standards in these documents are discussed in section 7 of this paper.

5.4 France

The standards for captions on French television are set out in the *Charte relative à la qualité du sous-titrage à destination des personnes sourdes ou malentendantes* ('Charter relating to the

[10] A copy of Part 79.1 is available from http://www.fcc.gov/cgb/dro/captioning_regs.html
[11] http://hraunfoss.fcc.gov/edocs_public/attachmatch/DA-05-2974A1.pdf
[12] http://www.fcc.gov/document/fcc-moves-upgrade-tv-closed-captioning-quality
[13] A copy of the document can be downloaded from http://www.dcmp.org/captioningkey/
[14] http://main.wgbh.org/wgbh/pages/mag/services/captioning/faq/sugg-styles-conv-faq.html

quality of subtitling addressed to the deaf or hard of hearing')[15] which was released by France's audio-visual regulator, the Conseil supérieur de l'audiovisuel (CSA), in December 2011.

The document contains some unusual colour standards which are not used anywhere else.

- White: for speakers visible on the screen
- Yellow: for speakers not visible on the screen
- Red: for sound effects
- Magenta: for music and lyrics
- Cyan: for the thoughts of a character, a narrator and voiceovers
- Green: for people speaking a foreign language

For offline captioning, word speed is defined in terms of number of characters (12 characters for a second, 20 characters for two seconds, 36 characters for three seconds and 60 characters for 4 seconds), with a tolerance of 20 per cent. This works out to be about 144 – 180 wpm.

For live captioning, the time lag should be less than 10 seconds.

In a forthcoming paper, 'France's National Quality Standard for Subtitling for the Deaf and Hard-of Hearing: an Evaluation', Tia Muller notes that the unusual colour standards were developed in the 1980s by the National Institute of the Young Deaf of Paris and some Deaf people. Because colours are used in this way, they cannot be used to identify different speakers as they are in many other countries. Instead, speakers are identified by caption positioning or by dashes to indicate new speakers. In a survey of Deaf and hearing impaired television users[16], 45 per cent of the survey participants said they had experienced difficulties identifying off-screen characters, while 41 per cent said they did not know the caption colour code by heart. In her paper, Muller suggests that the colour scheme be simplified and name tags be used for character identification, as is common in the UK, Australia and other countries.

Muller also notes that the relatively long delay of up to 10 seconds in live captioning is to allow for to the complexities of the French language, and the expectation that spelling should be perfect. (In reality, this can cause the time delay to stretch to 20 seconds.) As her survey participants identified time delay as a more serious problem than language errors, this 'could be seen to support the need to reconsider the current approach to live captioning in France'.

David Padmore of Red Bee Media, which produces captions in France, has told Media Access Australia that the unusual colouring standards add at least 10 per cent to the cost of captioning.

5.5 Spain

[15] A copy of the document can be downloaded from http://www.csa.fr/Espace-juridique/Chartes/Charte-relative-a-la-qualite-du-sous-titrage-a-destination-des-personnes-sourdes-ou-malentendantes-Decembre-2011

[16] The full results of this survey have not yet been published.

Spanish captioning standards are set out in the document *Subtitulado para personas sordas y personas con discapacidad auditiva* ('Subtitling for the deaf and hearing impaired')[17] issued by AENOR (the Spanish Association of Standardisation and Certification). This states that the maximum speed for captions for pre-recorded programs should be 15 characters per second. The average length of Spanish words is approximately 5.2 characters (slightly longer than English words at 5 characters) so this works out to be about 173 wpm.

The document also sets standards for identifying speakers, sound effects, colouring and positioning, etc, which are similar to UK and Australian standards.

5.6 Germany

Germany does not have caption standards which apply to all channels, but the public broadcasters ZDF and ARD both have comprehensive, voluntary standards documents[18]. ZDF's word speed for adult programs is 15 characters per second, and ARD's is 12 characters per second. The average word length in German 6.26 characters, so this works out to be 147 wpm and 115 wpm respectively.

For children's programs, the standard for ZDF is 12 characters per second (115 wpm) and for ARD 8 characters per second (77 wpm).

Other standards cover subjects including colouring, positioning, identifying speakers, music, sound effects and fine details of punctuation.

5.7 Australia

Captioning began in Australia in 1982, with the Australian Caption Centre (ACC) providing captioning for the Nine and Ten networks, the ABC and SBS, while the Seven Network did its own captioning in-house (later the ACC would take over Seven's captioning as well). The ACC developed a comprehensive set of standards, drawing on existing standards in the US and UK. At that time, captions tended to be tailored more to the profoundly deaf, so it was common for language to be reduced and simplified. The ACC initially captioned adult programs at 120 wpm, and children's programs at 90 wpm or 60 wpm. In the early 1990s, after a consultation process with consumer representative groups which included showing Deaf and hearing impaired TV viewers excerpts of programs captioned at different speeds, the maximum speed for adult programs was increased to 180 wpm, and for children's programs to 120 wpm.

Amendments to the *Broadcasting Services Act* passed in 1999 made captioning compulsory for all stations broadcasting digitally from 1 January 2001. Prior to the passage of the amendments, the then Department of Communications, Information Technology and the Arts sought submissions regarding captioning provisions, including whether there should be 'a standard in relation to the quality and accuracy of closed captioning'. It was ultimately decided not to include such a standard in the amended act.

---

[17] The document can be purchased from AENOR's website: http://www.aenor.es/aenor/actualidad/actualidad/noticias.asp?campo=4&codigo=23835&tipon=2

[18] These were provided to Media Access Australia by David Padmore, Director of Access Services at Red Bee Media. See also the European Commission's *Study on Accessing and Promoting E-Accessibility* (page 97), which can be downloaded from: http://ec.europa.eu/digital-agenda/en/news/study-assessing-and-promoting-e-accessibility

5.7.1 Deafness Forum's Caption Quality Code of Practice

In 2004, the Deafness Forum of Australia launched a campaign to improve the quality of captions on TV. As part of this campaign it put together a Caption Quality Code of Practice[19] outlining basic captioning standards.

It includes sections on general grammar and presentation, timing and editing (180 wpm is the appropriate speed for adult programs), colouring, positioning, sound effects and children's programs (appropriate speed is 120 wpm, although 90 wpm or 60 wpm may be appropriate for programs aimed at younger children).

Deafness Forum's code was endorsed by consumer representative groups and access suppliers, but not by the TV networks. However, it came to be considered the de facto basic standard for captioning on Australian TV.

5.7.2 The ACMA's captioning standard

In December 2010, following an investigation into access to electronic media for the hearing and vision impaired, the Australian Government released its *Media access review final report*[20]. Its recommendations included:

> *That the Australian Communications and Media Authority hosts captioning quality workshops, via a captioning committee, to develop criteria that the ACMA can use when assessing the quality of captions.*

In 2012, amendments to the *Broadcasting Services Act* gave the ACMA the power to determine standards that relate to captioning services by commercial television licensees, national broadcasters and subscription television licensees. For the purposes of this, quality includes

a) *readability; and*
b) *comprehensibility; and*
c) *accuracy.*

A series of workshops were subsequently hosted by the ACMA, attended by representatives of the free-to-air networks and subscription television, caption suppliers, the Deafness Forum of Australia, Deaf Australia, the Australian Communications Consumer Action Network (ACCAN) and Media Access Australia.

Some of the consumer advocates argued for metrics to be included in the standards (specifically a target of 98 per cent accuracy for live captions, and a target of three seconds for the maximum time lag for live captions). Consensus could not be reached on an appropriate model for measuring errors, however, while the broadcasters and caption suppliers argued that it was inappropriate to set a maximum time lag as this was influenced by too many factors outside a captioner's control. A

[19] The document, and a Deafness Forum position statement, can be found at: http://www.deafnessforum.org.au/pdf/Position%20Statements/Captioning%20Quality%20V2.pdf
[20] http://www.abc.net.au/mediawatch/transcripts/1105_bcd.pdf

representative from one of the free-to-air networks conducted tests during the negotiations which indicated that the average time lag on live captions on Australian TV was five seconds.

Media Access Australia proposed that a statement be included to the effect that pre-prepared block captions are inherently superior for the consumer to live captions, and programs (and parts of programs) should have pre-prepared captions whenever possible.

There was also some discussion about whether it was appropriate to nominate a cut-off time between the delivery of a program and its broadcast, below which it was acceptable for it to be live captioned. The caption suppliers argued that this was not feasible, as whether it is possible to pre-prepare captions depends not just on the program itself, but the other programs which need to be captioned at the time.

In the end, the ACMA decided not to include metrics, or a statement that captions be pre-prepared whenever possible, in the *Broadcasting Services (Television Captioning) Standard 2013*[21], which was released on 5 June 2013. Instead, the standard details the factors it will consider in each of the three criteria of readability, comprehensibility and accuracy when dealing with caption complaints from the public.

6. Models for measuring caption quality

6.1 The WER Model

The WER model is the most basic model for measuring the accuracy of a transcription of speech, and was originally designed for the evaluation of automatic speech recognition systems. One common formulation of it is:

$$\text{Accuracy} = \frac{N - D - S - I}{N} \times 100$$

N = Number of words in the audio
D = Deletion (a word in the audio is omitted)
S = Substitution (a word in the audio replaced by an incorrect one)
I = Insertion (a word not in the audio is added)

There are a number of problems which arise when this model is used to measure the quality of captions. The most obvious is that it does not take into account the seriousness of errors. A slight misspelling of a word, or the substitution of one homophone for another (e.g. son/sun) as is common in live captioning, may have very little impact on the comprehensibility of a caption, but the inclusion of an incorrect name or number can change the meaning completely.

The issue of deletions is also problematic. Word speed is a significant factor in the comprehensibility of captions, so captioners routinely omit unnecessary repetitions, irrelevant figures of speech, asides and so on. Rather than being errors, these deletions often increase comprehensibility.

[21] http://www.comlaw.gov.au/Details/F2013L00918

While the WER model has been used to evaluate caption quality, and is the model that has been adopted by the CRTC in Canada (see section 5.2), its limitations in this area have led to other models being developed.

6.2 NCAM's Caption Accuracy Metrics Project

In 2010, the Carl and Ruth Shapiro Family National Center for Accessible Media (NCAM), which is based at the non-commercial education channel WGBH in Boston, USA, embarked on a project to quantify errors in stenocaptions. The project had two aims:

- *develop an industry standard approach to measuring caption quality, and*
- *use language-processing tools to create an automated caption accuracy assessment tool for real-time captions on live news programming.*[22]

NCAM began with an online survey of viewers of captioned television news programs. 352 people took part in the survey, during which they were shown examples of caption errors and asked how these affected their ability to understand the program.

After reviewing many examples of caption texts, and drawing on the National Court Reporters Association's error grading system, NCAM divided caption errors into 17 basic types.

Table 1 NCAM's caption error types

	Caption error	**Example (caption/actual)**
1	Substitute singular/plural	man/men
2	Substitute wrong tense	run/ran
3	Substitute pronoun (nominal) for name	this man/Proper Name
4	Substitute punctuation	period instead of question mark
5	Split compound word/contraction	foot note/footnote did not/didn't
6	Two words from one (one wrong)	might yes/mighty
7	Duplicate word or insertion	criticism criticism
8	Word order	would I/I would
9	Correction by steno	disznits–dissidents
10	Dropped word – 1 or 2	"you know"
11	Dropped word(s) – 3+	"figure out what the best options are going to be"
12	Nearly same sound but wrong word(s)/homophone	sail/sale or work/werk
13	Substitute wrong word	blogger/hunger
14	Phonetic similarities/not a valid word	Human milating/humiliating
15	Random letters (gibberish)	Igbavboa
16	Word boundary error (also "stacking error")	paying backpack Stan/paying back

[22] Documents relating to the project, including the survey report and final report, can be downloaded from http://ncam.wgbh.org/invent_build/analog/caption-accuracy-metrics

	Caption error	Example (caption/actual)
		Pakistan
17	Transmission errors/garbling	GM sto/GM stock

The following table lists the seven error types which over 50 per cent of the survey respondents defined as a “severe error” which greatly impacted or destroyed their understanding of a sentence.

Table 2 NCAM caption error types with the worst ratings

Error type	Description	% of respondents
17	Transmission error, dropped letters	84%
16	Word boundary error	65%
15	Garbled syllable, not words	65%
14	Phonetic similarities (not words)	59%
11	Dropped words (significant, 3+)	59%
13	Substitute wrong word	55%
3	Substitute pronoun for proper name	53%

Using the survey results, NCAM developed a system for weighing errors in a caption transcript according to their severity. The basic calculation for the Weighted Word Error Rate (WWER) is:

$$WWER = \sum_{t=1}^{ErrorTypes} severity_t * errors_t \Big/ N$$

Figure 1 NCAM's Weighted Word Error Rate equation

NCAM worked with Nuance Communications to develop an automated caption accuracy assessment tool, CC Evaluator™, based on this method of error assessment. This tool could be used to compare an accurate or ‘clean’ verbatim transcript of a program with its corresponding caption file, and produce a Weighted Word Error Rate. But NCAM wanted to find out whether the tool could be used with a transcript created using automatic speech recognition (ASR) and still produce a meaningful result, thus achieving a completely automated system.

NCAM notes the apparent difficulties in using an ASR transcript like this. In most cases, such transcripts are actually less accurate than corresponding caption files would be.

> *The question becomes: how can you rate the quality of captions against an ASR transcript that might actually be worse that the captions themselves? The answer is that, by carefully analysing where errors occur in each transcript, we can statistically account for much of this difference.*

To test this theory, CC Evaluator was used to evaluate sample news captions against both a clean transcript (to produce a Weighted Word Error Rate), and an ASR transcript (to produce an Estimated Word Error Rate). Comparison of the two sets of data showed that they generally correlated, although the Estimated Word Error Rate was slightly overestimated for programs with very low error rates, and underestimated for programs with very high error rates.

NCAM notes that, ‘as more program samples are accumulated by the evaluation engine, better error estimates will be generated, particularly when applied to specific programs. The use of the ASR tools will also improve as the software is exposed to larger data sets.’

While there is still work to do to perfect NCAM’s system, it demonstrates that a completely automated system for evaluating the quality of live captions can produce meaningful results.

6.3 The NER Model

The NER Model[23] for measuring the accuracy of live captioning produced using the respeaking method was developed by Pablo Romero-Fresco of the University of Roehampton, and respeaking consultant Juan Martínez. Their goal in creating it was to have a model which would be functional and easy to apply, and ‘include the basic principles of word error rate calculations in speech recognition theory, which have been tested and validated for a long time’.

The model is based on the following formula:

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

N The number of words in the respoken text, including punctuation marks (which the respeaker needs to read out), identification of speakers, etc. (Note that this differs from the WER model, where N stands for the number of words in the original audio.)
E: ‘Edition’ errors, introduced by the respeaker. (The most common of these is that the respeaker omits information due to the speed of the dialogue).
R: ‘Recognition’ errors, which are caused by mispronunciation or the software not recognising a word.

To use the NER Model, a verbatim transcript of a program must be created, which the user then compares to the caption file, noting differences between the two (including dialogue which has not been captioned). The user marks each error as ‘Edition’ or ‘Recognition’, classifies it as ‘serious’, ‘standard’ or ‘minor’, and assigns it a score of 1, 0.5 and 0.25 respectively. Words which have been omitted but have no impact on comprehension may be labelled ‘correct errors’.

- Serious errors are defined as those which change the meaning of the text (e.g., ‘alms’ instead of ‘arms’ or ‘15%’ instead of ‘50%’).

- Standard errors are those which result in the omission of information. They include recognition errors, where the speech recognition software has failed to recognise a word and produced an entirely different word or words (‘hell of even’ instead of ‘Halloween’).

- Minor errors, which may not be noticed by a viewer, include a lack of capital letters from proper names.

As well as producing a realistic accuracy measurement, the NER Model can provide useful feedback to respeakers and suggest ways for them to improve their accuracy. Romero-Fresco and

[23] An outline of the NER model can be found here: http://roehampton.openrepository.com/roehampton/bitstream/10142/141892/1/NER-English.pdf (All quotes in this section are from this document.)

Martinez give the example of a caption file where contractions have not been conveyed property (e.g. 'we're' has become 'were'). In this instance, the respeaker should be advised not to use contractions.

To make using the NER model easier, a software tool called NERstar has been developed by Swisse Text and d'Accessibilitat i Intel•ligència Ambiental de Catalunya (Centre for Research in Ambient Intelligence and Accessibility in Catalonia, CaiaC)[24]. This compares the transcription and captions, and highlights discrepancies between the two. The user manually enters the error type and seriousness rating for each discrepancy, and the software then automatically calculates the accuracy rate.

The NER model has been used in Spain, Italy, Switzerland and Germany, and is the model the UK regulator Ofcom has asked broadcasters to use to measure the quality of captioning on selected live programs as part of its two-year quality measurement exercise (see section 5.1.2).

The model has also been adopted by Australian access company Ai-Media, which provides captioning for subscription television and became the caption service provider for the Nine Network in January 2014.[25]

6.4 Evaluating the caption quality measurement models

To perform a truly rigorous and comprehensive analysis of the quality of a captioned program, an accurate, verbatim transcript of the program is required. This is expensive to produce (while creating a completely accurate transcript presents problems of its own). A comparison of the transcript and caption file must then be performed, which is also an expensive and labour-intensive process. NCAM's Caption Metrics project, which aims to create an automatic process which does not require an accurate, verbatim transcript, is therefore very appealing. The initial results of the model are promising, and if perfected, it could theoretically be incorporated into the caption production process, allowing quality reports to be automatically generated for a broadcaster's entire live captioned output.

Further research and testing of the model was put hold as NCAM awaited the announcement of new caption quality standards by the FCC (see section 5.3).[26]

The chief advantage of the NER Model is that it has been specifically designed to evaluate respeaking captioning, which is now the most common form of live captioning in Europe, and used extensively in Australia. This means that, as well as producing a meaningful accuracy percentage, it can also be used as a training tool for respeakers as it highlights areas where they can improve their performance.

The disadvantages of the NER Model are that it requires an accurate, verbatim transcript, and using the model is very labour intensive. People who have used the model have told Media Access Australia that evaluating the quality of a captioned program takes between 10 and 15 times the

[24] http://www.speedchill.com/nerstar/
[25] http://ai-media.tv/downloads/file/2013%2011%2019%20Ai-Media%20Leads%20with%20International%20Quality%20Standard%20for%20TV%20Captions.pdf
[26] Personal communication from Larry Goldberg, Founder and In-House Advisor at NCAM.

length of the program. Given the commercial constraints on caption producers, it is therefore unreasonable to expect that the model be used to evaluate large volumes of captioned output.

That does not mean that it cannot be a useful tool. Poor quality live captioning can be the result of problems associated with an individual program, or of systemic issues such as an inexperienced captioner or one ill-suited to the task, inefficient work practices, problems with software or equipment, or captioners not being given adequate preparation time. Using the NER model to evaluate a representative sample of a broadcaster's captions could identify these systemic issues and lead to improvements in the service.

7. Commentary on caption quality issues and their treatment in different standards

7.1 Word speed (presentation rate)

Setting any word speed or presentation rate for captions must take into account the fact that their users can be dividing into two groups: Deaf people (whose first language is often sign language and whose reading skills may be lower than the general population) and the hearing impaired (many of whom have lost hearing as they have aged).

In the early days of captioning, word speed tended to be set lower to cater to the first audience. The first news program ever captioned was *The Captioned ABC Evening News*, a rebroadcast of the ABC's 6 p.m. evening news which went on air on PBS in the US on 3 December 1973.[27] A team of five captioners worked to create captions for the rebroadcast at 11 pm, writing the captions so that they were no higher than a sixth-grade reading level, with passive verbs eliminated, short words substituted for longer ones, and potentially confusing idioms replaced.

In 1999, Dr Carl Jensema and Dr Robb Birch published the results of a US government-funded study, *Caption Speed and Viewer Comprehension of Television Programs.*[28] During the study, 1,102 subjects were shown captioned television clips with speeds ranging from 80 wpm to 220 wpm. In their conclusion, the authors wrote:

> *The bottom line is that this study indicates that caption viewers are likely to be able to absorb facts and draw conclusions from captions that are presented as fast as 220 wpm for short periods of time. In general, this suggests that caption viewers are capable of keeping up with most verbatim captioning, since normal speech rates are unlikely to exceed 220 wpm.* (p. 30)

Since this study appeared, the trend in the US, Canada and Australia has been to make captions closer to verbatim. WGBH's *Captioning FAQ* states that 'much of the caption-viewing audience prefers to have a verbatim or near-verbatim rendering of the audio'. However, it also advises that

[27] http://www.vitac.com/news_blog/vitac-blog.asp?post=390
[28] http://www.dcmp.org/caai/nadh135.pdf

captioners should ‘strive for a reading speed that allows the viewer enough time to read the captions yet still keep an eye on the program video’.

Both the Canadian English-language and French-language standards state that ‘captions must be verbatim representations of the audio,’ and ‘speech must only be edited as a last resort, when technical limitations or time and space restrictions will not accommodate all of the spoken words at an appropriate presentation rate’.

European standards, on the other hand, tend to set maximum word speed rates. However, the French and Spanish standards do not specify whether their recommended maximums (which I have estimated equate to 180 wpm and 173 wpm respectively) apply to live programs as well as pre-recorded programs. Ofcom’s code is alone in making a distinction between the two (recommending a maximum of 180 wpm for pre-recorded programs and 200 wpm for live programs).

While verbatim captioning of live programs is often mentioned as an ideal (both in standards issued by regulators and by consumer advocates demanding full access), in a 2009 paper, ‘Respeaking the BBC News’[29], Carlo Eugeni finds that it is not possible for respeakers to produce completely verbatim captions. Eugeni looked at the strategies that respeakers were using to caption live programs by analysing the captions for 8 hours of BBC news broadcasts. He found that the average speech rate of the source text (i.e. the actual audio track of the programs) was 183 wpm. He found that the main strategy used by respeakers was to attempt to repeat dialogue (while also adding punctuation and colouring information).

> *This is possible when the speech rate is slow enough, where lexical density, turn taking and visual compensation are low enough, and where given information is sufficient to guarantee no other constraints are exercised on respeaking. However, not even ideal conditions enable respeakers to produce verbatim subtitles for a period longer than five macro-idea units. It seems that keeping the same pace as the original is a cognitively consuming and unsustainable activity.* (p. 43)

This suggests that there is an in-built maximum word speed for captions produced by respeakers (although the actual figure for any given program will vary according to its nature, the skills of the respeaker and other factors), It should be noted though that experienced stenocaptioners can produce captions much faster than those produced by respeaking. In the article ‘Confessions of a Television Captioner’[30], Jessica Bewsee writes that she was once stenocaptioning a sports program and realised that she was writing at a rate of 380 wpm. In practice, though, stenocaptioners are generally considered to be fully trained when they can achieve 96 per cent to 98 per cent accuracy at a speed of 180 to 200 wpm.

Eugeni also points out that, quite apart from word speed, there are advantages for comprehension in captions not being verbatim. One of these is that ‘it is very difficult, if not impossible, to follow subtitles if read as a completely oral text, from a cohesive reading and grammatical point of view’ (p. 65). In practice, as well as dropping unnecessary words, phrases and repetitions, most experienced respeakers learn to routinely ‘clean up’ the grammar of the people they are captioning, and this can only help caption users.

---

[29] Eugeni, Carlo, Respeaking the BBC News: A Strategic Analysis of Respeaking on the BBC’, *The Sign Language Translator and Interpreter* 3 (1), 2009, 29-68.
[30] http://www.hearinglossweb.com/Issues/Access/Captioning/Television/conf.htm

After all these factors are taken into consideration, we do not believe there is any point in including maximum word speeds for live captions in standards. There are too many variables which are out of a captioner's control, including the speed at which people in the program are speaking, while it is clear that there are inherent upper limits to the speed of captions that can be produced by respeakers.

While there is no consensus across the world about appropriate words speed rates for captioning on adult programs, there is more agreement about children's programs. A number of studies have been undertaken which demonstrate that Deaf and hearing impaired children benefit from lower caption word speeds[31]. For example, in their 1980 study *Captions and Reading Rates of Hearing-Impaired Students*[32], Edgar Schroyer and Jack Birch found that the average speed of normal extempore speech – and of speech in films and TV programs – is about 159 wpm. However, the mean wpm reading speed of the primary school students they studied was 125.7 wpm.

The DCMP's *Captioning Key*, the Canadian English-language standards and Deafness Forum's draft standards all indicate that a word speed of 120-130 wpm is appropriate for children's programs. While they do not specify a maximum word speed, Ofcom's code and the BBC guidelines also state that lower speeds are appropriate for children's programs.

7.2 Time lag in live captions (latency)

The delay between the speech (or other sounds) and the corresponding captions appearing on the screen in a program which has been captioned live is one of the most often voiced caption complaints made by Deaf and hearing impaired TV viewers. In a 2013 survey conducted by Action on Hearing Loss in the UK[33], it was considered to be the biggest problem by the greatest number of respondents (24 per cent), with the next most serious problem being inaccurate captions (22 per cent).

The time lag in live captions is determined by several factors, including the production systems in place, the reaction time of the stenocaptioners and respeakers, and the complexity of the dialogue. Ofcom's code sets a target of 3 seconds, which is often considered the shortest time lag that can be reasonably expected in most circumstances. This may change, however, when the results of Ofcom's two-year caption quality measurement exercise are in.

The Canadian English-language and French-language caption standards allow for a longer time lag (6 and 5 seconds respectively). The French charter calls for a lag of less than 10 seconds (which most caption users will find frustratingly long), reflecting a greater emphasis on ensuring that spelling and punctuation are correct.

It should be noted that there is ambiguity in all of these standards, as none of them make it clear whether the maximum time lag target applies to individual captions or the average over the whole program. Ofcom appears to be thinking about this issue, as in its measurement exercise, it has asked broadcasters to measure both the average latency and range of latencies.

---

[31] The DCNP has provided a useful summary of this research which is available at: http://www.captioningkey.com/captioning_presentation_research.pdf
[32] http://www.dcmp.org/caai/nadh131.pdf
[33] http://www.actiononhearingloss.org.uk/

The time lag in live captions is such a crucial factor for consumers that we feel that all caption quality standard should acknowledge this. We believe that 5 seconds is appropriate as a *target* which is achievable in most cases. It should be made clear that this target applies to the average time lag over the length of the program.

To overcome the problem of delay in live captions, it is often suggested that the broadcast of programs be delayed to allow time for the creation of the captions. This idea has generally been resisted by broadcasters, although it has been adopted by the Flemish public broadcaster VRT in Belgium, which delays its live programs by up to 10 minutes. And, as noted above, Ofcom will continue to discuss the idea with UK broadcasters.

7.3 Accuracy rates and measurement

Canada's CRTC is so far the only government regulator to have introduced a percentage accuracy rate (95 per cent in the English-language standard, 85 per cent in the French-language standard). The model it has chosen to determine the accuracy rate is the WER Model (see section 6.1) which basically counts every deviation in the caption file from the original audio without considering the seriousness of the errors. There are obvious problems with this approach.

The chief obstacle to including accuracy targets in caption standards has been finding a system which will take into account the relative seriousness of errors. The NER Model seems to be the most promising developed so far, especially as it has been geared toward evaluating respeaking captioning, now the most common form of live captioning in Europe. It therefore has potentially great benefits as a measure of accuracy for both regulators and caption suppliers, although the costs and labour time involved in it mean that it is not feasible to use it for more than a sampling of programs.

We therefore believe that Ofcom is right to have adopted the NER Model for its two-year live caption quality initiative to be completed by 2015. This will allow for thorough testing of the model on a variety of program genres, captioned by different access providers, and ultimately enable Ofcom to decide whether it would be practical and beneficial to introduce a caption accuracy metric into its code.

7.4 Captioning of late-delivery programs

One of the frustrations for caption consumers in recent years has been an increase in the live captioning of pre-recorded programs. Broadcasters have argued that this is inevitable given tighter production schedules which lead to programs being delivered late, the 'fast-tracking' of programs produced overseas so that they are screened here much earlier than previously, and an increase in overall captioning levels putting a strain on resources and meaning there is less opportunity to pre-prepare captions. During meetings held by the ACMA as it put together its caption guidelines, some consumer representatives argued that a cut-off time between a program's completion and broadcast should be nominated, below which it is acceptable to caption it live.

We believe the danger in setting a cut-off time would make it more likely for any program delivered in less time than this to be live captioned. It appears that this has happened in Canada (see section 5.2.2).

7.5 Hybrid captioning

Media Access Australia has for a long time argued that the most effective way to achieve an overall increase in caption quality is to keep live scrolling captions to an absolute minimum.
In the context of news captioning, this means using the hybrid method of pre-preparing captions from scripts and other materials as far as possible, then sending them as block captions. We note though that even if the pre-prepared captions are sent as scrolling captions, there are potential benefits in reduced time lag and improved accuracy.
The hybrid method is currently used in the UK, Australia and France, and we would strongly encourage its use whenever it is technically possible to do so.

7.6 Converting live captions to block captions

There has been some research done on how scrolling captions affect viewers. A study of eye movements by Martínez and Linder[34] found that viewers spend 88 per cent of their time reading scrolling captions, compared to 67 per cent of their time watching block captions. Another study by Rajendran, Duchowski, Orero, Martínez and Romero-Fresco concluded that 'according to eye movement metrics, text chunking by phrase or by sentence reduces the amount of time spent on subtitles, and presents the text in a way that is more easily processed.'[35]

Software now exists which can convert live captions created by respeakers or stenocaptioners to block captions rather than having them transmitted as scrolling captions. The drawback of this is that it increases the time lag between the captions and the audio. Consumers dislike the time lag in live captions, but they also dislike scrolling captions and the extra time that needs to be spent reading them. We therefore believe it would be worthwhile for caption suppliers to explore this technique and invite feedback from consumers and regulators.

8. General conclusions

Captioning has existed as a commercial service for many decades and the maturing of the industry has seen a gradual shift away from justifying the need or scope of captioning, to more practical matters covering issues like quality. Captioning did not appear in the same way, at the same levels of coverage, spontaneously across the world. It grew rapidly in large English-speaking markets, such as the UK and US and then more recently in other languages. The process of developing and formalising captioning standards has followed this uneven development of captioning.

As captioning spread across the world there has been a corresponding gradual increase in communication, sharing of information between advocates for captioning, regulators, suppliers and their clients and practical application of standards to real-world situations.

---

[34] Martínez, J., & Linder, G. (2010, October 6-8). 'The Reception of a New Display Mode in Live Subtitling'. In *Proceedings of the Eighth Languages & The Media Conference* (pp. 35–37).
[35] Rajendran, D.J., Duchowski, A.T., Orero, P, Martínez, J and Romero-Fresco, P, 'Effects of Text Chunking on Subtitling: A Quantitative and Qualitative Examination'. (Page 11). This can be downloaded from http://andrewd.ces.clemson.edu/research/vislab/docs/perspectives.pdf

There has always been a tension between formality (wanting to provide specific metrics/descriptions of what should be presented) and practical reality (what is delivered by suppliers due to broadcast environments, human capability and viewer feedback). This is not unique to captioning quality regulation but it has been refined with more practical input and acceptance of compromises by viewers.

The predominant area of caption quality issues remains programs that are captioned live. This is due to the inherent nature of live programming which amplifies any shortcomings in a caption production process, particularly the very limited ability to correct errors as they occur.

The issues that always impact on quality are: training, experience of the captioner, interaction with the broadcaster, resources devoted to the task (including being able to cope with peaks of demand), and quality control of output.

A significant evolution is that regulators have become more practically focused over time and less fixated on legalistic measurements. This also reflects better communication between regulators and the broadcasters that they regulate, and a better understanding of how content is produced and delivered across a range of platforms.

There is a trend to more universality in captioning standards reflecting the better communication of standards and the internationalisation of suppliers and their broadcast clients (with a couple of notable exceptions, such as the French colouring standards). Encouraging this process helps to better align the objectives of cost-effective delivery of program content and consistently delivery of high quality access.

9. MAA recommendations

It is clear from the research that a simplified set of recommendations saying that presentation speeds should be set at particular levels or specific colours should be used for sound effects are not very helpful in delivering more consistent, high quality captions. Instead the evidence suggests that effort is better focused on communication, sharing of information and grounding standards discussion in practical reality. Finally, the ultimate arbiter of quality, the viewer, should be a key party to those discussions and deliberations.

More specifically the following recommendations are made:

1. Greater discussion across national boundaries and information sharing between regulators about their approaches will help standardise high quality.
2. That the move to towards practical experimentation (such as Ofcom's approaches) and field-testing of different approaches and evaluation models be encouraged.
3. That consumers are involved in decision-making about what works for them (i.e. the trade-off between time lag and block captions for live programs).
4. That the use of the hybrid method of captioning be encouraged for news and current affairs programs.

Acknowledgements

I would like to thank Larry Goldberg, Tia Muller, David Padmore, Chris Howe, Adrian Chope, Pablo Romero-Fresco, Michelle Kwan, Nicole Gilvear, Margaret Lazenby and Laura Greeves for help in compiling this paper.

Appendix: Metrics in international caption standards

Table 3 Metrics in international caption standards

Country	Document	Accuracy rate	Time lag in live captions (latency)	Word speed (presentation)
UK	Ofcom code	Not specified.	No more than 3 seconds	Pre-recorded programs: 160 – 180 wpm. Live programs: no more than 200 wpm. Slower speeds appropriate for young children.
UK	BBC guidelines	Not specified.	Not specified	Adult programs: 140 wpm (but can go up to 180 wpm). Lower speeds for children's programs.
Canada	CRTC English-language standards	For pre-recorded programs, the target is 100%. For live programs: at least 95%.	Must not exceed 6 seconds	Adult programs: Verbatim – speech must only be edited as a "last resort". Children's programs: 120 – 130 wpm.
Canada	CRTC French-language standards	For pre-recorded programs, the target is 100%. For live programs: at least 85%.	Must not exceed 5 seconds	Adult programs: Verbatim – speech must only be edited as a "last resort".
Canada	CAB standards	For pre-recorded programs, the target is 100%. For live programs: at least 95%.	Must not exceed 6 seconds	Adult programs: Verbatim – speech must only be edited as a "last resort". Children's programs: 120 – 130 wpm.
France	CSA charter	Not specified.	Less than 10 seconds	144 – 180 wpm.

Country	Document	Accuracy rate	Time lag in live captions (latency)	Word speed (presentation)
Spain	AENOR standards	Not specified.	Not specified	15 characters per second (approx. 173 wpm).
Germany	ZDF standards	Not specified	Not specified	Adult programs: 147 wpm Children's programs: 115 wpm
Germany	ARD standards	Not specified	Not specified	Adult programs: 115 wpm Children's programs: 77 wpm
US	DCMP Captioning Key	Goal of 100%.	Not specified	Adult programs: 150 – 160 wpm. Lower to middle-level educational media: 120 – 130 wpm. Upper-level educational media slightly higher than 130 wpm.
US	WGBH Captioning FAQ	Not specified	Not specified	Not specified, although verbatim where possible.
Australia	Deafness Forum draft standard	Not specified	Not specified	180 wpm for adult programs. 120 – 60 wpm for children's programs.

Glossary

ASR transcript	A transcript produced using automatic speech recognition software.
Block captions	Captions which appear in-screen 1 – 3 lines at a time.
Closed captions	Captions that can be switched on and off by the viewer (as opposed to open captions which are always on-screen).
Hybrid captioning	Method of news and current affairs captioned where all captions are as far as possible pre-prepared before broadcast, with stenocaptioners or respeakers captioning the remainder.
Latency	A term used to denote the delay between the audio and the corresponding caption appearing on-screen.
Offline captioning	The captioning of pre-recorded programs.
Online captioning	The captioning of live programs, where captions are created in whole or part as the program is broadcast.
Pop-on captions	US term for block captions.
Presentation rate	A term often used to denote the speed of captions (usually expressed as words or characters per minute).
Respeaking	A method of live captioning where captioners repeat the dialogue into a microphone as the program is broadcast, and speech recognition software converts this to captions.
Roll-up captions	Captions which roll up from the bottom of the screen and are usually visible three lines at a time. Often seen on US live programs. Also called scroll-up captions.
Scrolling captions	Live captions which appear on the screen one word a time.
Stenocaptioner	A captioner who creates captions for live programs using a stenographic keyboard.
Subtitles	The term used for captions in Europe (sometimes expanded to ‘subtitles for the deaf and hard of hearing’).