Universal Access in the Information Society (2021) 20:741–751
https://doi.org/10.1007/s10209-020-00735-6

LONG PAPER

Negotiating quality assessment in media accessibility: the case of live subtitling

Pablo Romero-Fresco[1,2]

Published online: 17 July 2020

Abstract

Given the recent shift in scholarly attention from the quantity to the quality of media accessibility (MA), this paper aims to provide an account of how the assessment of quality in MA may be negotiated across researchers, regulators, companies and users. Taking as an example the use of the NER model to assess live subtitling quality, it focuses on what each party needs to do and compromise on, so that a consensus can be reached and quality in MA can be assessed. It covers firstly the role played by researchers, who are normally expected to develop rigorous models of assessment while, at the same time, ensuring that they are transferable to the industry. The focus is then placed on media regulators, which must often make a choice between adopting a soft or a hard approach. The former tackles quality only superficially and may have little effect on professional practice or on the viewers' experience. The latter requires investing time and resources to conduct research, but it also results in tangible improvements in quality. As for companies, they must be willing to accept external reviews and even to modify their training, but they can also benefit from the improved performance of their staff. Finally, in order to have a voice in this area, the users may be expected to engage with the method and the results of the quality assessment. This can enable them to have more informed and realistic demands that can lead to effective changes and, ultimately, to better access.

Keywords Media accessibility · Quality assessment · NER model · Live subtitling

1 Introduction

Over the past years, an increasing number of publications [14, 35, 46] on audiovisual translation (AVT) and especially media accessibility (MA) have been using as a starting point the fact that scholarly attention in these areas is shifting from quantity to quality. This has also been confirmed by other key actors in the field, such as access service providers, user associations and governmental regulators. The UMAQ project[1] (which aims to provide a unified theoretical framework for understanding quality in MA), its conference and this issue of UAIS offer further evidence of the key place that the issue of quality now occupies in this area. However, there is much less consensus regarding how quality can be approached, which in turn reflects the difficulty in agreeing on what quality really means:

> What is quality in translation? Quality is about as elusive an idea as 'happiness', or indeed, 'translation'. Quality means very many different things depending on your perspective. To those in translation management, the concept is often associated with processes, work flows and deadlines. To professionals, quality is often a balancing act between input and efficiency. To academics, it is often a question of equivalence and language use [44].

Despite the complexity involved in specifying what constitutes quality, "many people have to judge translation quality on a daily basis: revisers, editors, evaluators, teachers, not to mention the subtitlers themselves, and of course: the viewers" (ibid). In order to do this, quality assessment methods are often needed. This only compounds matters further, given the difficulty involved in finding models that can be accepted by all parties and that can obtain comparable results that have an impact on professional practice

✉ Pablo Romero-Fresco
p.romero-fresco@roehampton.ac.uk; promero@uvigo.es

[1] GALMA, Facultad de Filología y Traducción, Universidade de Vigo, Campus universitario Lagoas, Marcosende, 36200 Vigo, Pontevedra, Spain

[2] University of Roehampton, Roehampton, UK

[1] http://pagines.uab.cat/umaq/.

and ultimately on the experience of the users. First of all, researchers are not always willing to propose this type of assessment model, given that there is no certainty that it will produce reliable results in the medium or long term. Secondly, even if an assessment model is accepted by the research community, its impact on the industry is far from guaranteed. As will be explained in Sect. 4, regulators are sometimes more keen to encourage an increase in quality in general terms than to set up a monitoring system with a specific model. In turn, as shown in a report by the UK regulator Ofcom [37], pp. 9–11, companies often prefer to stick to their own in-house quality control methods than to be subject to an external assessment process that may be time-consuming and that can trigger comparisons with their competitors.

However, examples such as the use of the NER model [48] to assess live subtitling quality around the world show that finding consensus across different stakeholders regarding quality assessment in MA is by no means impossible. Subtitling is one of the modalities included within MA, along with others such as sign language interpreting or audio description for blind and partially sighted people. Unlike pre-recorded subtitling, live subtitling has to be produced in real time, which leads to delay and errors. Accuracy in live subtitling is thus an important factor and the NER model is one of the several methods to assess it. First introduced in 2015, the NER model is now being used by universities, broadcasters and companies in Europe, Australia, Africa and America [52]. It is also recommended by the International Communication Union, the United Nation's agency for information and communication technologies, as the preferred method to assess the quality of remote captioning [32].

Based on the experience with the NER model and on the research conducted so far regarding quality assessment in translation (Sect. 2), this article aims to provide an account of how the assessment of quality in MA may be negotiated across researchers (Sect. 3), regulators (Sect. 4), companies and users (Sect. 5). More specifically, it focuses on what each party needs to do and compromise on so that a consensus can be reached and quality in MA can be assessed.

2 Quality assessment in translation

Over the past few years, researchers have been calling for a wide and universalistic view of media accessibility that concerns not only persons with sensory disabilities but anyone who cannot, or cannot completely, access audiovisual content in its original form [28, 51]. This is in line with the recent EU Audiovisual Media Services Directive (2016), which is targeted at both persons with sensory impairments and older people, and with the latest international standard on subtitling, ISO/IEC DIS 20071-23 [57], which cites as its main target users not only persons with hearing loss but also persons with learning or cognitive difficulties, persons who cannot hear the audio content due to environmental conditions (noisy surroundings or circumstances where the sound is not available or not appropriate) and persons watching a movie in a non-native language. In other words, MA may be seen to encompass AVT. However, MA has traditionally been regarded as a sub-unit within AVT [42] VIII. In turn, AVT has only recently been seen to come of age and become independent from the parent discipline of translation studies [19], where translation quality assessment (TQA) has traditionally been a key issue of debate [31].

As explained by Doherty [22], TQA aims to "ensure a specified level of quality is reached, maintained, and delivered to the client, buyer, user, reader, etc., of translated texts". TQA is thus important for translator training, professional certification and for the analysis of the performance of translation technologies such as machine translation. In translation studies, TQA has been approached "from a theoretical and case study perspective" (ibid:132), dealing with issues such as source text versus target text focus or the purpose and function of translations, as well as with challenges such as subjectivity [10, 11, 33, 55], lack of systematic approaches [10, 30, 54] and inconsistency in terminology [12]. However, in Doherty's view, the main issue in TQA is "the lack of explicit operationalization of concepts" and the "non-adherence to established standards upheld in test theory, namely those related to validity, reliability, and the selection of evaluators" (2017, p. 132). Validity is generally regarded as the extent to which a model/assessment method measures what it is supposed to measure [53]. Internal validity concerns the aspects assessed as part of the test and external validity refers to whether or not the results can be extrapolated beyond the test [20]. Reliability refers to the degree of consistency of the test results across different evaluators [9, 15]. This leads to the notion of interrater agreement and to the importance of the selection and training of evaluators.

Other key concepts for the purpose of this article may be derived from the use of TQA in the industry and, more specifically, from the analysis of machine translation output, which often presents a choice between human and (semi-) automatic assessment. Human evaluation offers the benefit of rationale judgement but also the disadvantages derived from subjectivity [33] and time [22]. Automatic evaluation with metrics such as BLEU—BiLingual Evaluation Understudy [43], GTM—General Text Matcher [59] and TER—Translation Edit Rate [56] is less time-consuming but cannot make sophisticated judgements about idiomaticity, naturalness, etc. Drawing on a large-scale survey [21], Doherty explains that TQA in the industry tends to resort to a combination of customised and customer-focused human evaluation and semi-automated methods that deduct errors of different severity (i.e. minor, major, critical) in a selected sample and provide a result that may or may not achieve a previously set quality threshold and that is presented to the users in a simple form (i.e. traffic light or star system).

Although there are differences between TQA and quality assessment in MA, many of the aspects mentioned here (validity, reliability, focus on the customers/viewers, etc.) apply to the discussion presented in this article, as will be shown in the following sections.

3 The researchers' role in quality assessment

Drawing on the lessons learnt from TQA and on the experience obtained with the introduction and development of the NER model, this section includes some of the basic requirements that a model for the assessment of quality in MA may be expected to meet. From the point of view of the researcher, such model would need to be, at the very least, rigorous (research-informed, valid, reliable, user-focused) and transferable (straightforward, flexible and valid for training).

3.1 Rigour

If a model is *research-informed*, it may help to dispel the fears of subjectivity that are often attached to what are regarded as prescriptive models based on the individual experience of the researcher. In the case of the NER model, its formula (number of respoken words – edition errors – recognition errors/number of respoken words × 100) draws first of all on the basic principles of WER (word error rate) as applied by the US National Institute of Standards and Technology and on its adaptation by the Centre de Recherche Informatique de Montréal (CRIM) [50]. The classification of errors in terms of severity (minor, standard and serious) is based on the research project set up in 2010 by the Carl and Ruth Shapiro Family National Center for Accessible Media [8] and especially on the findings of the EU-funded DTV4ALL project [47]. The results of the questionnaires and focus groups conducted as part of the latter project yielded a distinction between three different types of errors as far as impact on viewers' comprehension is concerned: minor errors that do not impact on comprehension (which prompted respondents to note "we are deaf, not stupid"), standard errors that cause the viewers to lose information and often to be perplexed at the oddity of the subtitle output (which led some users to say that "we now can speak English and teletext") and finally serious errors that may not be noticed by the viewers but that create a different (and credible) meaning in the subtitles (considered by viewers as "lies").

In order to have *validity*, the model must measure the dimensions that determine quality in the specific MA modality at hand. In the case of the NER model, the dimensions covered are accuracy, speed and delay, as agreed following official consultations by governmental regulators in the UK and Australia with broadcasters, subtitling companies, researchers and user associations [40]. One of the key aspects of accuracy is editing or, more precisely, the extent to which the condensation that is often found in live subtitling causes or not loss of information. At the moment, this can only be done manually, hence the need to have a human evaluator using the NER model. This adds a degree of subjectivity that can be mitigated if the model is reliable.

The *reliability* of the model requires the calculation of the inter-rater (dis)agreement between evaluators, who, prior to this, must be selected and trained. For the official assessments of live subtitling quality in the UK and Canada, conducted in collaboration with the governmental regulators and the national broadcasters, bespoke sessions were organised to train professional subtitlers from the different companies involved to become NER-certified evaluators. For LiRICS, the first official certification of respeakers, the assessment is being carried out by a team of NER-certified external evaluators belonging to the research group GALMA (Galician Observatory for Media Accessibility) [49]. The average inter-rater disagreement found for the Ofcom project, that is, the average discrepancy amongst the 30 evaluators assessing a sample of 78,000 subtitles from 300 programmes over 2 years, was 0.09% [50]. This is the equivalent of 0.25 on a 0/10 scale and can thus be considered a low (and satisfactory) inter-rater disagreement.

Finally, a rigorous model for the assessment of quality in MA may be expected to be *user-focused*, in line with the second of the three shifts produced by the accessibility revolution according to Greco: the change from a maker-centred to a user-centred approach [28]. The exact role played by the users will be discussed more thoroughly in Sect. 5, but suffice it to argue that in the case of the NER model, as noted above, the different degrees of error severity (and thus the final score) are determined by the impact that an error may have on the users. In other words, unlike marking systems for translation or MA courses at university [13], the score does not take into account other factors such as, for instance, the difficulty involved in subtitling a specific programme. This is valued and noted in the assessment, but the accuracy rate produced by the NER model is based exclusively on the experience of the user. A second involvement of the users, which perhaps refers to whether a model is user-informed, rather than user-focused, lies in the extent to which the results produced by a model are in line with the users' view of quality. This has traditionally been an issue in the area of TQA [45]. As far as the NER model is concerned, recent research has found positive correlations between NER scores and the users' views of the quality of live subtitling output in countries such as Canada and Poland [1, 58].

3.2 Transferability

The above requirements can contribute to guarantee that a model is solid and as objective as possible, and that its results are replicable. However, they do not suffice to guarantee that it will have an impact on society, instead of simply sitting in an academic publication. For a model to be taken up by the industry, it needs to be *transferable*, which is regarded here as being straightforward, flexible and valid for training. This involves difficult decisions for the researcher, who may need to simplify elements of the model to make it more accessible for the evaluators without necessarily compromising its rigour.

The need for a model to be informed by previous research and to have both validity and reliability may lead to complex methods that can be perfectly valid for scientific purposes and can produce transferable results, but that can also prove too complicated or time-consuming for the industry to apply it regularly. It thus becomes necessary to ensure that the model is, to a certain extent, *straightforward* to understand and apply. The NER model, for example, was designed to be as easy to use as possible, with only two types of errors (edition and recognition) and three levels of severity (minor, standard, serious). The downside is that the information it provides is not as thorough as, for example, the model devised by Eugeni [24], which can yield a much more detailed analysis of the causes and types of errors in live subtitling. The advantage of the NER model is that it is relatively easy to understand. This is essential considering that in large-scale projects, researchers are unlikely to be able to train all evaluators, which means that some of them will be receiving the model second-hand. Despite the simplicity of the NER model, the fact that it is based on the comparison between the original audio and the subtitles and thus needs a transcript of the programme to be analysed makes it appropriate for the analysis of random samples but too time-consuming for daily use at a company.

Another key aspect of the simplicity of a model lies in the results it produces, which should be measurable and recognisable. For the sake of argument, these results can be based on metrics or not based on metrics and, in the case of the former, these metrics may be regarded as cold or warm. Not using metrics allows for more nuanced assessments and may be particularly suited to certain arts-related disciplines, but it can also lead to subjective, inconsistent and non-transferable results. In contrast, metrics have the potential to produce results that are easy to follow as long as the final scores are presented in an accessible manner. There are, however, some attached risks, as was revealed by the experience with the NER model. Firstly, this model has a quality threshold of 98%, which means that, what for the average user may be a very high score (i.e. 96%), for the reviewers equates to virtually unreadable subtitles. This made it necessary to adapt the scores to a 1/10 scale, where the different values were classified in four groups: poor, good, very good and excellent subtitles [50]. Secondly, the focus on the numbers (and especially on the final accuracy rate) by the media that reported on the results of the Ofcom project meant that the assessment that accompanied every final score went unnoticed. The model was then seen to provide cold metrics and to reduce the complex issue of quality to a single number. In reality, though, in the NER model it is the assessment that matters, as it contextualises the score and provides a thorough analysis of its impact and causes.

The transferability of a model may also be helped by its *flexibility*. When first used in Canada, where live subtitles are mainly produced by stenotyping, as opposed to respeaking, the NER model was modified [18]. From a terminological point of view, a proposal was made to change the labels used for the different types of errors to terms that were deemed more accessible to the users. More substantially, two of the error types were slightly modified to account for the errors that normally occur with steno-made subtitles as opposed to respoken ones. Here, the researcher may need to weigh up the benefits of having a single model that can produce comparable results across countries against the need to localise it so that it can account for the different national practices.

Finally, another important element for a model to be transferable is the extent to which it is valid for training. Even though the need for rigour may make it impossible for a model to be used on a daily basis in the industry, if it is valid for training, it will have an impact not only on how media accessibility is assessed but also on how it is practiced in the first place. The NER model is not used on a daily basis, but the fact that many of the main live subtitling companies in the world are using it for training (for example, the distinction between minor, standard and serious errors) creates a degree of consistency (i.e. respeakers are assessed on the model they have been trained with) and may contribute to the comparability of the results.

4 Regulators

Media regulation can be described as the process whereby governments and other political and administrative authorities control or guide, through established rules and procedures, different kinds of media activities such as press, radio, television, film, recorded music and the Internet [34]. In the case of broadcasting, regulation takes many forms, from laws and clauses in national constitutions to administrative procedures and technical specifications. It often involves intervention in broadcasters' activities with the following aims: ensuring universal accessibility to broadcast services, equitable broadcast concessions, diversity in social, political, cultural and local/regional terms, high quality of content with particular reference to information, education, advertising, culture, taste and decency and adherence to the basic interests of the state in matters of security and good order [27, 29]. Accessibility is, thus, at the very core of media regulation, which means that the provision of MA services is within the remit of media regulators.

A quick look at the role played by media regulators in different countries shows at least two different approaches to the assessment of quality in MA, especially regarding live subtitling: the soft, complaint-driven approach adopted by the Federal Communications Commission (FCC) in the US and the Australian Communications and Media Authority (ACMA) in Australia or the hard, audit-based approach chosen by the Office of Communications (Ofcom) in the UK and the Canadian Radio-television and Telecommunications (CRTC) in Canada.

4.1 Soft approaches: the US and Australia

As part of the Telecommunications Act, the FCC adopted in 1996 the first set of rules regulating the provision of closed captioning on television, the Closed Captioning of Video Programming on Television [25], to ensure that all Americans can have access to video services and programmes. The initial focus of these closed captioning obligations was quantity rather than quality. The FCC admits that they expected video providers to introduce captioning quality controls through their agreements with captioning companies, which would then result in high-quality captions. However, as shown by hundreds of complaints received by the FCC every year, "the marketplace alone has not provided effective incentives for all providers to maintain good quality captioning" [25], p. 16). The widespread frustration with the poor quality and inconsistency of captioning quality led the FCC to publish in 2014 an Order dealing specifically with closed captioning quality as a means to ensure that broadcast material is "fully accessible through the provision of closed captions" [26], p. 18), in accordance with the mandates of Section 713 of the Communications Act. The Order identified four parameters to assess closed captioning quality [26], p. 20): accuracy (when captions reflect the content of the programme and identify speakers), synchrony (when captions match the dialogue and sounds and are displayed at a readable speed), completeness (when captions are provided for the entire programme) and adequate placement (when captions do not obscure or are not obscured by other on-screen information).

With regard to live captioning, the Order takes into consideration "the challenges associated with accurately captioning such programming" (2014, p. 30). It also expects content providers to adhere to a set of best practices, such as using metrics to assess accuracy, synchronicity, completeness and placement of live captions, establishing minimum acceptable thresholds and performing regular sample audits. The Order even goes as far as to explain how accuracy and latency should be measured, proposing a formula that is very similar to that of the NER model. However, as noted by consumer groups, the fact that a particular process is followed in the production of a programme "cannot cure the program's inaccessibility if the process ultimately results in poor-quality captions" [26], p. 37. For these groups, metrics are necessary to assess "the ultimate quality of captions delivered to consumers" [26], p. 37. However, the FCC did not share this view [26], p. 44:

> this order rejects the need for the Commission to "identify clear metrics for determining the completeness, accuracy, readability, and synchronicity of programming." Rather, by (1) defining standards that comprise good quality captions, and taking into consideration the extent to which compliance with these standards can be achieved for various types of programming, and (2) requiring video programming distributors (VPDs) to obtain certification from programmers that they are either complying with the quality standards or with Best Practices, we focus more on the end result – *i.e.* the provision of captions that effectively convey video programming content for people who are deaf and hard of hearing to the same extent that the audio track conveys this content to people who do not rely on captions – than a strict numeric accuracy standard that may be more burdensome, yet less effective in ensuring that viewers can fully understand the captions.

In practical terms, this means that there is no real assessment of (live) captioning quality. Firstly, because this soft approach is based on reaction to user complaints(2014, p. 43):

> We will rely on consumers to bring any potential noncompliance with our captioning quality standards to our attention. We believe that a consumer-complaint-driven procedure, rather than an audit-driven one, is the most practical means to monitor industry compliance with our rules.

Secondly, and most importantly, the Order does not include information about how the above-mentioned parameters (which in fact refer to pre-recorded rather than live captions) can be applied to the assessment of quality. When does an omission in the captions qualify as an error? Are all errors equally serious? What kind of latency is considered acceptable or inacceptable? Even if the Order is to be applied to complaints, how can these parameters allow the comparison between two sets of captions from two different complaints? In short, the Order may be regarded as a timely and useful document to highlight the importance of captioning quality and to identify the aspects that determine this quality, but it is not designed to enable quality assessment. The risk here is that just as captioning quality decreased due to the lack of effective incentives following the 1996 FCC guidelines, this 2014 Order is unlikely to persuade content providers to tackle the complex issue of live captioning quality by themselves, which can potentially undermine the impact of the Order and the reason why it was drafted in the first place.

In Australia, following consultations with the television industry, television captioning service providers and community representative groups, ACMA published in 2013 its Television Captioning Standard, which aimed to ensure that "captioning services provided for television programs are meaningful to deaf and hearing-impaired viewers" [4], p. 2. ACMA ruled out the use of metrics and decided that "a holistic approach would be taken when assessing the quality of a captioning service provided for a program" [3]. The Standard identified three parameters to assess captioning quality [4], pp. 3–5: readability (legible colour and font, line break and segmentation, punctuation, position and number of lines), accuracy (transfer of content, tone of voice, sound effects and music and extent to which the captions are verbatim or, if not, reflect the actual meaning of the spoken content) and comprehensibility (speaker identification, reading time, synchrony with the audio, spelling, shot changes). It is a similar approach to that of the FCC in the US, and one that does not apply to live captioning, where accuracy and delay/synchrony cannot be assessed in the same way as it is done for pre-recorded captioning.

In view of this issue, the ACMA undertook in 2016 a review of its TV Captioning Quality Standard to consider how the differences between live and pre-recorded programs may affect the quality of television captions. For this review, ACMA put in place a consultation process with captioning providers, broadcasters, user associations and caption users, organised several events with key stakeholders, issued a consultation paper for public comment and studied different international approaches to the assessment of live captioning quality, especially in the US, the UK and Canada. After this extensive consultation process, ACMA decided to maintain the Standard's approach to assessing captioning quality. The only minor modification to the Standard is this note under paragraph 6 [5], p. 3:

> While noting that it is not authorised to determine that a lower quality of captioning service is acceptable for a kind of program or program material (see subsection 130ZZA(2B) of the Act), in determining this Standard, the ACMA has considered the differences (including time constraints for live content) between providing captioning services for live and pre-recorded television programs; and wholly live or wholly pre-recorded television programs and television programs that include both live and pre-recorded program material (see subsection 130ZZA(2A) of the Act).

ACMA considered the use of metrics and more specifically the NER model to assess captioning quality, but decided against it for being too onerous on live captioning providers, too lenient on pre-recorded captioning providers, too prescriptive and complex, and finally for not accounting for captions that are edited but do not necessarily lose information and for comparing audio and captions instead of focusing on viewer comprehension (2016, pp. 8–9). Some of these arguments may be regarded as questionable. Firstly, the NER model is not so much lenient but simply not suited to the analysis of pre-recorded captions. It was designed exclusively for live captions. Secondly, the possibility of edited captions is factored in the analysis, whether this edition causes the loss of information (editing error) or not (correct edition). Finally, the model does compare audio with captions but the classification of errors, the score and the final assessment are all based on how the quality of the captions impacts on the viewers' experience.

In any case, ACMA opted for a non-metric approach as the most appropriate way to ensure that the quality of a captioning service is meaningful to deaf and hearing-impaired viewers. This is the so-called 'meaningful access' test (2016, p. 20), which, according to ACMA, is best achieved by addressing the non-metric factors outlined in the Television Captioning Standard (accuracy, readability and comprehensibility). However, as in the US, since neither the standard nor its review include any information about how these non-metric factors can be applied to an assessment of quality and since the ACMA's modus operandi is to investigate only captioning complaints, there is no real assessment of (live) captioning quality. This is evidenced by the fact that the full captioning results published by ACMA over the past years [6, 7] include only information about captioning quotas (quantity), but not about quality.

In other words, soft approaches to quality assessment in MA such as those adopted by the US and Australia in the area of live captioning have value because they raise awareness, trigger discussions amongst key stakeholders and often result in the inclusion of content on quality in official captioning standards. However, this content is often vague and not suitable for the assessment of actual captions, which means that these soft approaches may end up having very little or no impact on the quality of the captions and the experience of the viewers.

4.2 Hard approaches: the UK and Canada

In the UK, Section 3 of the Communications Act (2003) establishes that Ofcom should have regard to the needs of persons with disabilities and requires the regulator to prepare and periodically review a code guiding broadcasters on how they can "promote the understanding and enjoyment of their services by people with sensory impairments, including people with hearing impairments" [38]. Until 2013, Ofcom had published several updates of its guidance [41]. They included some aspects related to quality, but the focus was mainly on quantity, that is, setting increasingly high subtitling quotas for broadcasters to meet. Once these quotas were met and approached 100%, Ofcom turned its attention to quality and particularly to live subtitling, the most challenging form of subtitling and the source of most complaints. In May 2013, Ofcom published a consultation document in order to gather views from broadcasters, researchers, subtitle providers and user associations as to how to improve the quality of live subtitling on UK TV [37]. Although this is a very similar process to the ones started in the US and Australia by the FCC and the ACMA, Ofcom chose to adopt a more research-based approach, taking time to learn about the latest findings in live subtitling in an attempt to establish "the facts behind the understandably non-scientific observations of subtitle users" [36], p. 6. This enabled Ofcom to identify three key dimensions that are essential in, and specific to, live subtitling (speed, latency and accuracy), in contrast to the more general captioning dimensions established in the US (accuracy, synchrony, completeness and placement) and Australia (readability, accuracy and comprehensibility). Acknowledging the complexity involved in live subtitling, Ofcom noted that there was scope for "small but significant improvements", which, taken together, could result in "an appreciable difference over time to the quality of the viewing experience for those relying upon subtitles to understand and enjoy television" (2013a, p. 6).

Following its consultation, Ofcom decided to undertake a two-year assessment of the quality of live subtitling on UK TV. From March 2013 until March 2015, broadcasters were asked to measure and report on the quality (accuracy, latency and speed) of their live subtitles, based on four samples of news programmes, entertainment and chat shows. Broadcasters were asked to apply the NER model for this analysis, given its use by researchers and companies in other countries, and a team of experts at the University of Roehampton was asked to validate the measurements provided by the broadcasters from a third-party standpoint. The results of this project show a steady increase in the accuracy of live subtitles, both per genre and per broadcaster, and a significant decrease in latency [50]. As noted by Ofcom [39], other noticeable improvements relate to an increased awareness amongst programme producers of the need to collaborate with (and deliver content earlier to) subtitling companies, which resulted in a significant increase in subtitling quality. Finally, Ofcom highlights that the research project has provided a great deal of knowledge about the trade-offs involved in live subtitling, which means that broadcasters, companies and users can now take more informed decisions regarding the improvement of live subtitling quality.

In Canada, the Broadcasting Regulatory Policy 2012-362 drafted by the CRTC in 2012 approved the compulsory quality standards for English-language closed captioning. The standards included information on how to monitor accuracy rate and a requirement for broadcasters to provide the Commission, every 2 years, with evidence of efforts made to improve the accuracy of captioning. In their first biennial report to the Commission [23], the broadcasters complained that the required accuracy rate of 95% for live captions was not achievable under the assessment model chosen by the CRTC. This model calculated accuracy exclusively by comparing the on-screen captions with a verbatim transcription of program audio. Every discrepancy was thus penalised as an error, thus ruling out the possibility of correct editions in captions. This model may be regarded as measuring reduction rate, rather than accuracy rate. As a result, the broadcasters explained that, as per the findings of this assessment, all broadcasters were in noncompliance with their conditions of licence relating to the accuracy rate for the closed captioning of live English-language programming.

In view of this, the CRTC started a consultation process in 2015 calling for comments on alternative methods to assess quality in live captioning [16]. This prompted the creation of the so-called 2016 Working Group, made up of broadcasters, captioning providers and user associations, which proposed a trial to adapt the NER model to suit Canadian circumstances and to test its validity to measure the accuracy of live captioning on Canadian TV [17]. From April 2017 to June 2018, a series of live-captioned programmes were assessed, using the Canadian version of the NER model, by 11 evaluators trained by the author of this article. Each program was evaluated by two separate evaluators. The objective of the trial was twofold: to ascertain if the NER model can produce consistent results (different evaluators using the system should come to the same conclusions) and to find out if these results are in line with subjective impressions of caption quality. As for the first objective, the study concludes that, provided that evaluators are trained in the use of the model, its results are "reliable and useful, i.e. NER results are "repeatable" within reasonable ranges" [1]. An additional consideration here is that the results show no evidence of "home-team" bias: the broadcaster's affiliated evaluators did not favour their own shows when their assessment was compared to that of non-affiliated evaluators. Consistency and lack of bias were the attributes required by the Commission to demonstrate that a measurement system is "objective".

As far as the second objective is concerned, i.e. the extent to which the NER model meets the needs of the users, the trial showed that, in general, the NER scores correspond to user satisfaction, which does not apply to the old verbatim scores [1]. During the course of this trial, the CRTC asked the Captioning Consumers Advocacy Alliance to conduct consumer research into the subjective impressions of quality amongst captioning users. The results show that NER results and consumer perceptions align well. In general, the 2016 Working Group concluded that "NER, as adjusted with the Canadian Guidelines, is an effective tool for measuring caption accuracy, and can be used with a good level of confidence" [1], p. 14. It also added, very much in line the research project conducted by Ofcom in the UK, that more research is needed to understand the trade-offs involved in live captioning and the impact they may have on live captioning quality. The Group considers that the findings obtained during this trial will be of use to researchers "for years to come" [2], p. 9.

Following this report, the CRTC has published a final call for comments [18] proposing an amendment of the English-language closed captioning mandatory quality standards for live programming so that (a) the (Canadian) NER model, with a threshold of 98%, replaces the verbatim test as the official method for live captioning quality assessment and (b) that two programmes, of which one must be news, are assessed every month, with results published every year. Broadcasters and captioning providers are expected to discuss any monitored program scoring below the accuracy threshold and to provide a report describing the remedial action to be undertaken.

To conclude, this section has shown two different approaches to quality assessment in MA by media regulators. The soft approach adopted in the US and Australia may lead to an inclusion of content on quality in official captioning standards, but it is based on user complaints and, as such, does not entail the assessment of live subtitling quality. In contrast, the examples of the UK and Canada show that, if the regulator is willing to invest time and resources to conduct research, audit-driven approaches (whether involving reporting, as in the UK, or regulating, as in Canada) can lead to tangible improvements in the quality of live captions and to a more thorough understanding by all stakeholders of what this complex issue involves.

5 Companies and end users

As well as researchers and regulators, companies and users (or user associations) also play an important role in quality assessment.

As shown in the case of the Ofcom project, it is not uncommon to find certain resistance amongst companies to the idea of an external quality review. During the consultation period, some content providers advised Ofcom to adopt a complaint-based system, arguing that the assessment proposed was too time-consuming, of limited public value and redundant given the daily use of existent in-house quality checks [36], pp. 9–11. In its reply, Ofcom explained that, although interesting and useful, a complaint-based approach is no substitute for a thorough quality assessment system, and encouraged companies to pursue both [36], pp. 13–14. Ofcom added that in-house quality assessment tools do not suffice, as they are not available to other broadcasters or viewers and their results are not comparable between broadcasters. Many of these internal tools are based mostly on the analysis of the subtitles, rather than on the comparison between the audio and the subtitles, as the latter requires spending a great deal of time transcribing the audio. It thus makes sense to use these methods for the daily or at least regular assessment of quality, saving the NER model for the analysis of specific samples. An alternative for this is the so-called NER at-a-glance, designed by the company AiMedia, which consists of applying the model live by comparing the subtitles with the audio and assigning errors in real time, with no need for transcription.

Another interesting contribution that companies can make to quality assessment in MA is to integrate the assessment model used in the assessment as part of their in-house training. This can allow the staff to internalise it and to be familiar with it as they gain professional experience. Following from the Ofcom project, several companies in the UK incorporated the NER model or some of its features to their in-house respeaking training programmes. Finally, companies can be even more involved in quality assessment in MA if, as is the case with AiMedia, they require their professionals to be certified as respeakers. AiMedia has done this through LiRICS (Live Respeaking International Certification Standard), a system that assesses the performance of professional respeakers with the NER model [49]. The data obtained so far from the candidates' tests bear a striking resemblance to the results of the Ofcom study, which shows that these respeakers are performing at a professional standard.

In any case, regardless of the degree of involvement in quality assessment, any effort in this regard can help a company offer a better service for the users and can allow the possibility of using quality as a distinctive selling point.

Finally, an analysis of quality assessment in MA cannot be completed without exploring the role played by the users. This is in line with the shift from a maker-centred to a user-centred approach in MA and with the "increasing attention towards users as bearers of valuable knowledge for the investigation of accessibility processes and phenomena" [28]. Models of assessment may in this regard be user-focused, user-informed, user-led and/or user-friendly. As explained in Sect. 3, the NER model is user-focused in that the final score is based on the potential impact that the different types of errors analysed may have on the viewers' comprehension. The model is also user-informed, as the classification of the different degrees of error severity (minor, standard and serious) is based on reception studies with viewers from several countries. The fact that, as found out by the projects carried out in Canada and Poland [1, 58], the model's scores are in line with viewers' subjective opinion of live subtitling quality, also shows that it is user-informed. The model is not, however, user-led, as the viewers have so far not been involved in its design or implementation. An exception here may be found in Canada, where the 2016 Working Group leading the live captioning project has decided to train users as NER evaluators. Time will tell whether this development, which places the users at the forefront of quality assessment, proves successful.

While the user-related characteristics of the model mentioned so far (and especially user-focused and user-informed) relate to its rigour, the last one (user-friendly) relates to its transferability. If a model is user-friendly, its methodology and especially its scores are accessible for the users, which means that they can learn from it and have a more informed view of quality. In the case of live subtitling in the UK, before the Ofcom project, the viewers' lack of knowledge about the complexity involved in the production of live subtitles often led them to make unrealistic demands ("subtitles should be 100% accurate, with no delay"), which were quickly dismissed by the industry [47]. An assessment method that is straightforward and whose results are easy to understand enables the viewers to get to grips with the intricacies of live subtitling and to make more informed demands that are more likely to have an effect on companies and regulators, which may be seen as one of the key roles played by the users.

6 Conclusions

Despite the widespread consensus in academia and in the industry about the need to shift our attention from the quantity to the quality of MA, any attempt to propose an actual method to assess this quality is bound to face considerable resistance, often expressed in arguments such as "it is too costly or time-consuming to be practical" or "quality cannot be defined, since there are as many views on quality as there are users".

This paper has attempted to show that quality assessment in MA can be possible if the different stakeholders involved agree that the inevitable compromises that they have to make are outweighed by the benefits they can obtain regarding quality. Researchers are normally expected to develop rigorous models of assessment that are research-informed, valid, reliable and user-focused, but in order to ensure that they are taken up by the industry, these models must also be transferable (straightforward, flexible and valid for training). Regulators must often make a choice between adopting a soft or a hard approach. The former tackles quality only superficially and has proved to have little or no effect on professional practice or on the viewers' experience. The latter requires investing time and resources to learn about existing studies and to conduct research, but it also results in tangible improvements in quality. Companies must be willing to accept external reviews and even to modify their training processes, but they can benefit from the improved performance of their staff and from the possibility of using quality as a selling point. Finally, in order to have a voice in this area, the users may be expected to make themselves available for reception research (so that assessment models are user-informed) and to engage with the method and the results of the quality assessment. This has proved successful in those countries in which it has happened and it has enabled users to have more informed and realistic demands that can lead to effective changes and, ultimately, to better access.

The journey to effective quality assessment in MA is long and busy, but some of the examples included here show that it can also be successful. There should be no excuse not to try it, especially because the other option is not to do anything, which, at this stage, should not be an option at all.

Funding This research has been conducted within the frameworks and with the support of the EU-funded projects ILSA: Interlingual Live Subtitling for Access (2017-1-ES01-KA203-037948) and EASIT: Easy Access for Social Inclusion Training (2018-1-ES01-KA203-050275) and the Spanish-government funded projects Inclusión Social, Traducción Audiovisual y Comunicación Audiovisual (FFI2016-76054-P) and EU-VOS Intangible Cultural Heritage. For a European Programme of Subtitling in Non-hegemonic Languages (Agencia Estatal de Investigación, ref. CSO2016-76014-R) and the Galician government -funded project Proxecto de Excelencia 2017 Observatorio Galego de Accesibilidade aos Medios.

References

1. 2016 Working Group: EBG NER trial: final report. https://crtc.gc.ca/eng/archive/2019/2019-9.htm (2018a). Accessed 25 June 2020
2. 2016 Working Group: Findings and proposal of the 2016 Working Group. https://crtc.gc.ca/eng/archive/2019/2019-9.htm (2018b). Accessed 25 June 2020
3. ACMA: Broadcasting services (television captioning) standard: explanatory statement (2013). Australian Communications and Media Authority, Canberra (2013). Accessed 25 June 2020
4. ACMA. Broadcasting Services (Television Captioning) Standard 2013. Australian Communications and Media Authority, Canberra. https://www.acma.gov.au/~/media/BroadcastingInvestigations/Issue for comment/pdf/Broadcasting Services Television Captioning Standard 2013.pdf (2013). Accessed 25 June 2020
5. ACMA. Review of the Television Captioning Standard Final report: Canberra: Australian Communications and Media Authority. https://www.acma.gov.au/theACMA/review-of-tv-captioning-standard (2016). Accessed 25 June 2020
6. ACMA. Captioning results for 2016–17: Canberra. https://www.acma.gov.au/Industry/Broadcast/Television/TV-content-regulation/captioning-results-2016-17?utm_medium=email&utm_campaign=Release of captioning compliance results&utm_content=Release of captioning compliance results+CID_837ec8c13ac33c1853044ae2 (2017). Accessed 25 June 2020
7. ACMA. TV captioning results in 2017–18: Canberra: Australian Communications and Media Authority. https://www.acma.gov.au/theACMA/tv-captioning-results-in-2017-18?utm_medium=email&utm_campaign=Annual captioning results for 201718&utm_content=Annual captioning results for 201718+CID_fd25d62e82f210c125180d927166adb0&utm_source=SendEmailCampaigns&utm_ter (2018). Accessed 25 June 2020
8. Apone, T., Brooks, M., O'Connell, T.: Caption Accuracy Metrics Project. Caption Viewer Survey: Error Ranking of Real-time Captions in Live Television News Programs. Boston (2010)
9. Bachman, L., Palmer, A.: Language Testing in Practice'. Oxford University Press, Oxford (1996)
10. Bassnett-McGuire, S.: Translation Studies. Routledge, London (1991)
11. Bowker, L.: A corpus-based approach to evaluating student translations. The Translator **6**(2), 183–209 (2000)
12. Brunette, L.: Towards a terminology for translation quality assessment: a comparison of TQA practices. The Translator **6**(2), 169–182 (2000)
13. Cerezo Merchán, B., de Higes Andino, I.: Using evaluation criteria and rubrics as learning tools in subtitling for the D/deaf and the hard of hearing. Interpret. Transl. Train. **1**, 68–88 (2018)
14. Chmiel, A., Vercauteren, G., Mazur, I.: Media accessibility training - Call for Papers. Linguistica Antverp., **18** (2018)
15. Clifford, A.: Discourse theory and performance-based assessment: two tools for professional interpreting. Meta J. Des. Traducteurs **46**(2), 365–378 (2001)
16. CRTC: Broadcasting notice of consultation CRTC 2015-325. https://crtc.gc.ca/eng/archive/2015/2015-325.htm (2015). Accessed 25 June 2020
17. CRTC: Broadcasting regulatory policy CRTC 2016-435. https://crtc.gc.ca/eng/archive/2016/2016-435.htm (2016). Accessed 25 June 2020
18. CRTC: Broadcasting notice of consultation CRTC 2019-9. Ottawa. https://crtc.gc.ca/eng/archive/2019/2019-9.htm (2019). Accessed 25 June 2020
19. Díaz Cintas, J., Neves, J.: Taking stock of audiovisual translation. Audiovisual Translation. Taking Stock, pp. 1–8. John Benjamins, Amsterdam and Philadelphia. https://doi.org/10.1002/9781405198431.wbeal0061 (2015)
20. Doherty, S.: The impact of translation technologies on the process and product of translation. Int. J. Commun. **10**, 1–23 (2016)
21. Doherty, S., Gaspari, F., Groves, D., van Genabith, J.: Mapping the industry I: Findings on translation technologies and quality assessment. Technical Report, GALA (2013)
22. Doherty, Steven: Issues in human and automatic translation quality assessment. In: Kenny, Dorothy (ed.) Human Issues in Translation Technology, pp. 131–148. Routledge, London (2017)
23. EBG: Report on efforts to improve the quality of closed captioning. https://crtc.gc.ca/eng/BCASTING/ann_rep/bmt_cbc_rm_sm.pdf (2014). Accessed 25 June 2020
24. Eugeni, C.: La sottotitolazione linguistica automatica: Valutare la qualità con IRA. CoMe. Studi Di Comunicazione e Medializione Linguistice e Culturale **II**(1), 102–113 (2017)
25. FCC: Closed captioning of video programming on television. Federal Communications Commission, Washington D.C. https://www.fcc.gov/general/closed-captioning-video-programming-television (1996). Accessed 25 June 2020
26. FCC: The 2014 closed captioning declaratory ruling, order, and notice of proposed rulemaking. Federal Communications Commission, Washington D.C. https://www.govinfo.gov/content/pkg/FR-2014-03-31/html/2014-06754.htm (2014). Accessed 25 June 2020
27. Feintuck, M.: Media Regulation, Public Interest and the Law. University of Edinburgh Press, Edinburgh (1999)
28. Greco, G.M.: The case for accessibility studies. J. Audiovisual Transl. **1**(1), 204–232 (2018)
29. Hoffman-Riem, W.: Regulating the Media. Guildford Press, New York (1996)
30. Hönig, H.: Positions, power and practice: functionalist approaches and translation quality assessment. In: Schaeffer, C. (ed.) Translation and Quality, pp. 6–34. Multilingual Matters, Clevedon (1998)
31. House, J.: Quality. In: Routledge Encyclopedia of Translation Studies, pp. 222–225. Routledge, London & New York (2009)
32. ITU: Overview of remote captioning services. Technical Paper. ITU (2018)
33. Koponen, M.: Comparing human perceptions of post-editing effort with post-editing operations. In: Proceedings of the 7th Workshop on Statistical Machine Translation, pp. 181–190. Montreal, Canada (2012)
34. Lunt, P., Livingstone, S.: Media Regulation. Governance and the Interests of Citizens and Consumers. SAGE, London (2012)
35. Neves, J.: Subtitling for deaf and hard-of-hearing audiences: moving forward. In: Luis Pérez-González (Ed.) The Routledge Handbook of Audiovisual Translation. Routledge, London (2018). Accessed 25 June 2020
36. Ofcom: Measuring the quality of live subtitling. London. https://www.ofcom.org.uk/__data/assets/pdf_file/0017/51731/qos-statement.pdf (2013a). Accessed 25 June 2020
37. Ofcom: The quality of live subtitling: a consultation. London. https://www.ofcom.org.uk/consultations-and-statements/category-1/subtitling (2013b). Accessed 25 June 2020
38. Ofcom: Ofcom publishes first results on quality of TV subtitles. https://www.ofcom.org.uk/about-ofcom/latest/media/media-releases/2014/ofcom-publishes-first-results-on-quality-of-tv-subtitles (2014). Accessed 25 June 2020
39. Ofcom: Measuring live subtitling quality: results from the fourth sampling exercise. London. https://www.ofcom.org.uk/__data/assets/pdf_file/0011/41114/qos_4th_report.pdf (2015a). Accessed 25 June 2020

40. Ofcom: Measuring the quality of live subtitling. London. https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/tv-research/live-subtitling (2015b). Accessed 25 June 2020
41. Ofcom: Ofcom's code on television access services. London. (2015c)
42. Orero, P. (ed.): Topics in Audiovisual Translation. John Benjamins, Amsterdam (2004)
43. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, pp. 6–12 (2002)
44. Pedersen, J.: The FAR model: assessing quality in interlingual subtitling. J. Spec. Transl. **28**, 210–229 (2017)
45. Ray, R., DePalma, D., Pielmeier, H.: The Price-Quality Link. Common Sense Advisory Ltd, Cambridge (2013)
46. Remael, A., Orero, P., Carroll, M.: Audiovisual Translation and Media Accessibility at the Crossroads. Rodopi, Amsterdam (2012)
47. Romero-Fresco, P.: The Reception of Subtitles for the Deaf and Hard of Hearing in Europe. Peter Lang, Bern (2015)
48. Romero-Fresco, P., Martínez, J.: Accuracy rate in live subtitling: the NER model. In: Díaz-Cintas, J., Baños, R. (eds.) Audiovisual Translation in a Global Context: Mapping an Ever-changing Landscape, pp. 28–50. Palgrave MacMillan, London (2015)
49. Romero-Fresco, P., Melchor-Couto, S., Dawson, H., Moores, Z., Pedregosa, I.: Respeaking certification: Bringing together training, research and practice. Linguistica Antverp. **18**, 216–236 (2019)
50. Romero-Fresco, P.: Accessing communication: the quality of live subtitles in the UK. Lang. Commun. (2016). https://doi.org/10.1016/j.langcom.2016.06.001
51. Romero-Fresco, P.: In support of a wide notion of media accessibility: access to content and access to creation. J. Audiov. Transl. **1**(1), 187–204 (2018)
52. Romero-Fresco, P.: Respeaking subtitling through speech recognition. In: Pérez-González, L. (ed.) The Routledge Handbook of Audiovisual Translation, pp. 96–113. Routledge, London (2018)
53. Rothe Neves, R.: Translation Quality Assessment for Research Purposes: An Empirical Approach. Universidade Federal de Minas Gerais, Brasil (2008)
54. Sager, J. (1989). Quality and standards—the evaluation of translations. In: Picken (Ed.) The Translator's Handbook, pp. 91–102. ASLIB, London
55. Snell-Hornby, M.: The professional translator of tomorrow: language specialist of all-round expert? Teaching Translation and Interpreting: Training, Talent and Experience, pp. 9–22. John Benjamins, Amsterdam (1992)
56. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of 7th Conference of the Association for Machine Translation in the Americas, pp. 223–231 (2006)
57. Standardization, I. O. for: Information technology – user interface component accessibility – Part 23: Guidance on the visual presentation of audio information (including captions and subtitles) (ISO/IEC DIS 20071-23: 2018). https://www.iso.org/standard/70722.html (2018). Accessed 25 June 2020
58. Szarkowska, A., Krejtz, K., Dutka, Ł., Pilipczuk, O.: Are interpreters better respeakers? Interpret. Transl. Train. **12**(2), 207–226 (2018). https://doi.org/10.1080/1750399X.2018.1465679
59. Turian, J., Shen, L., Melamed, I. (2003). Evaluation of machine translation and its evaluation. In: Proceedings of MT Summit IX, pp. 386–393