Romero-Fresco, P., & Pöchhacker, F. (2017). Quality assessment in interlingual live subtitling: The NTR model. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *16*, 149–167.

Quality assessment in interlingual live subtitling: The NTR Model

Pablo Romero-Fresco

Universidade de Vigo, Spain
promero@uvigo.es
University of Roehampton, United Kingdom
P.Romero-Fresco@roehampton.ac.uk

Franz Pöchhacker

University of Vienna, Austria
Franz.Poechhacker@univie.ac.at

This article presents a model for assessing the quality of interlingual live subtitling, as carried out by way of respeaking and automatic speech recognition or other means. The first part reviews the requirements and formulas for quality assessment in intralingual live subtitling and presents the NER model, developed by the first author, which serves as a foundation for the new model. Drawing on insights from the literature on both prerecorded interlingual subtitling and simultaneous interpreting, the authors then introduce the NTR model, an NER-based formula for calculating the accuracy rate of (interlingual) live subtitles. The model features a set of categories for scoring the accuracy of content and form as well as a three-level grading scale for translation and recognition errors. The application of the NTR model is exemplified, highlighting the role of the accuracy rate as part of an overall qualitative assessment.

1. From intralingual to interlingual live subtitling

Intralingual live subtitling (or "real-time captioning", as it is known in the United States) is commonly regarded as one of the most challenging modalities within media accessibility – one in which errors and delays are almost inevitable. The International Telecommunication Union (ITU) (2005) defines it as "the real-time transcription of spoken words, sound effects, relevant musical cues, and other relevant audio information" (p. 5) to enable deaf or hard-of-hearing persons to follow a live audiovisual programme. Ever since they were introduced in the United States and Europe in the early 1980s, live subtitles have been produced through different methods: standard QWERTY keyboards, dual keyboard, Velotype and the two most common approaches, namely, stenography and respeaking (Lambourne, 2006). Stenography uses a system of machine shorthand in which letters or groups of letters phonetically represent syllables, words, phrases and punctuation marks. This is the preferred method to produce live subtitles in the United States and Canada, and it is also used in other countries such as Italy and Spain, especially for transcriptions and subtitles in courtrooms, classrooms, meetings and other settings. Respeaking (or "real-time voice writing", as it is also known in the United States) refers to the production of written text (such as subtitles) by means of speech recognition. This is currently the preferred method for live subtitling around the world, described more thoroughly as

> a technique in which a respeaker listens to the original sound of a (live) programme or event and respeaks it, including punctuation marks and some specific features for the deaf and hard-of-hearing audience, to a speech-recognition software, which turns the recognized utterances into subtitles displayed on the screen with the shortest possible delay. (Romero-Fresco, 2011, p. 1)

When respeaking was first introduced in Europe, in 2001, subtitling companies knew what they wanted (i.e., professionals who could produce intralingual live subtitles using speech recognition), but in many cases they did not know how to identify qualified practitioners, since there was no academic training or research in this area. As a result, respeaking practices differ greatly across countries (Romero-Fresco, 2011). In the United Kingdom, live subtitlers work individually and correct their own errors once they have been displayed on the screen (Marsh, 2006), whereas in France they work in teams of three or even four people who correct the errors before the subtitles are broadcast to the viewers (CSA, 2011). These interesting developments "on the ground" have not been matched by systematic research, which is still scarce. It accounts for only 4 per cent of all academic publications on media accessibility and 0.8 per cent of all publications about audiovisual translation in general (Romero-Fresco, forthcoming). However, since 2006, and especially during the past five years, research on live subtitling and respeaking has developed significantly and can now be said to have caught up with the industry. (Progress has also been made in training, albeit to a lesser extent.) The main areas of research interest are the analysis of the respeaking process (Eugeni, 2006), the training of respeakers (Arumí Ribas & Romero-Fresco, 2008; Remael & Van der Veer, 2006), the analysis of respoken subtitles (Eugeni, 2009), including quality assessment (Romero-Fresco, 2016), their reception by the users (Eugeni, 2008; Rajendran, Duchowski, Orero, Martínez, & Romero-Fresco, 2012) and, finally, the application of respeaking to contexts other than TV (Romero-Fresco, 2011).

As societies have become more and more linguistically diverse, there has also been growing demand for *inter*lingual live subtitling to make programmes accessible to speakers of other languages. This puts the industry back to where it was for intralingual live subtitling at the turn of the century. Subtitling companies and broadcasters, who are equipped to provide intralingual live subtitles, are now testing different methods to produce interlingual live subtitles: bilingual stenographers subtitling from English into Spanish in Spain, subtitlers respeaking TV programmes from Welsh to English in Cardiff and teams of three people (an interpreter, a subtitler and a respeaker) in Flanders, where an antenna delay is used to allow the team to edit the subtitles before they are broadcast for the viewers. However, this professional transition from intralingual to interlingual live subtitling is not being accompanied by academic research, which means that there is little knowledge about how interlingual live subtitlers can be trained and especially how the quality of their output can be assessed.

The aim of this article is to present a model that can be used to assess quality in interlingual live subtitling. The first two sections offer a review of quality assessment in intralingual live subtitling (with special emphasis on the NER model) as well as in (pre-recorded) interlingual subtitling and in (simultaneous) interpreting, all of which are thought to be relevant to the practice of interlingual live subtitling. The second part of the article is devoted to the presentation and discussion of the NTR model.

2. Quality assessment in intralingual live subtitling

2.1. Requirements and formulas

Quality assessment in intralingual live subtitling varies greatly across and even within countries. As explained in Romero-Fresco (2016), what may be expected from these models of assessment is that they meet at least some of the following requirements: (1) they are functional and easy to apply, (2) they take into account not only the linguistic accuracy of the subtitles but also the comparison to the original speech,[1] (3) they account for the possibility of reduced and yet accurate subtitles depending on the different national editing conventions, (4) they provide information about not only the accuracy of the subtitles but also other aspects of quality such as delay, position, speed, character identification, etc., (5) they account for the fact that not all errors have the same origin or impact on the viewers' comprehension and (6) they provide an assessment of quality as well as an overall idea of aspects to be improved, in other words, food for thought as far as training is concerned.

Most models of assessment for intralingual live subtitling are based on the basic principles of WER (word error rate), which have traditionally been applied to the analysis of accuracy in speech recognition (Dumouchel, Boulianne, & Brousseau, 2011). These are product-oriented, error-based models that identify three types of error: deletions (D), substitutions (S) and insertions (I) (see Figure 1).

$$\text{Accuracy rate} \frac{N - \text{Errors } (D + S + I)}{N} \times 100 = \%$$

Figure 1: Formula used by the US National Institute of Standards and Technology to calculate word accuracy

Although useful for assessing speech-recognition output, this type of model is less suited to intralingual live subtitling, since it does not meet the third requirement outlined above and therefore penalizes any discrepancy between the source and target texts, even if the meaning of the original audio is retained in the subtitles:

(1) (Audio) Well, you know, you have to try and put out a good performance, I mean, yeah, it's kind of a stepping stone, isn't it, really?
(Subtitles) You have to try to put out a good performance. It's a stepping stone.

For Example 1, the model would yield an accuracy rate of 52 per cent. However, in most countries this example would be considered as 100 per cent correct, given that the words omitted do not have an impact on the viewers' comprehension (and allow the respeaker to catch their breath and keep up with the original speaker).

In order to solve this problem, the Centre de Recherche Informatique de Montréal (CRIM) adapted the above model by including a human operator who reviews the subtitle output and decides whether or not the deletions have caused loss of information (Dumouchel et al., 2011). This model meets requirements 1 to 3, but 4, 5 and 6 remain neglected.

In the United States in 2010, the Carl and Ruth Shapiro Family National Center for Accessible Media introduced the so-called "weighted word error rate" (WWER), a formula that classifies subtitle errors on the basis of their seriousness and automates the process so that no human intervention is needed to assess subtitle quality (Apone, Brooks, & O'Connell, 2010). The WWER formula is useful in that it accounts for the fact that not all subtitle errors have the same impact on the viewers. However, the lack of human intervention means that the omission of two words would automatically yield an error rate of only 0.05 per cent, even for a subtitle that introduces misleading information. Therefore, the omission of "the US and" in Example 2 below would result in an accuracy rate of 99.95 per cent, despite the fact that many deaf viewers with no access to the audio would not notice the error and would therefore take the information to be true.

(2) (Audio) There are a number of questions in the UK, but the big one is the US and whether it's about to slip into recession.
(Subtitles) There are a number of questions in the UK, but the big one is whether it's about to slip into recession.

2.2. The NER model

The NER model was first introduced in Romero-Fresco (2011) and developed further in Romero-Fresco and Martínez (2015) in order to meet the six requirements mentioned above. The model draws on the basic principles of WER calculations but highlights the need for human intervention in assessing subtitle quality. The NER model grades errors of edition (E) and recognition (R) according

to different degrees of severity while also accounting for editing that can be considered an appropriate, or "correct", choice.

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

CE (correct editions):
Assessment:

Figure 2: Formula used by the NER model to calculate accuracy

In this model, N is the number words in the subtitles. However, it is important to note that the NER model is meaning-focused and therefore the unit used for scoring errors is the idea unit, defined by Chafe (1985) as a "unit of intonational and semantic closure" (p. 106). Idea units can be identified because they are spoken with a single coherent intonation contour, preceded and followed by some kind of hesitation, made up of one verb phrase along with whatever noun, prepositional or adverbial phrase may be appropriate; they usually consist of seven words and take about two seconds to produce (Chafe, 1985, p. 106).

Edition errors (EEs) may often be the result of the subtitler's strategic decision-making, as in the case of an original with an excessively high speech rate, which may force the subtitler to omit information. Since deliberate editing is difficult to identify, such losses of idea units would be scored as errors. The same applies to the subtitler's decision to add idea units or paraphrase the original text in a way that introduces incorrect information, which may be due to miscomprehension of the original text. EEs are identified by comparing the subtitles against the transcript of the audio and may be classified as "minor", "standard" or "serious", scoring 0.25, 0.5 and 1, respectively.

REs are usually misrecognitions caused by mishearing and/or mispronunciation on the part of the respeaker, or by the specific technology used to produce the subtitles, be it speech recognition or stenography. These errors may be deletions, insertions or substitutions and are identified by comparing the respoken text against the transcript of the audio. They are also classified as minor, standard or serious, scoring 0.25, 0.5 and 1, respectively. Correct editions (CEs) are deviations from the original that do not involve a loss of information – as illustrated in Example 1. Given the difficulty involved in producing verbatim live subtitles (Romero-Fresco, 2009), omitting redundancies and hesitations is often considered good practice in Europe as long as the coherence and cohesion of the original discourse are maintained. CEs are not factored into the accuracy score but form an integral part of the overall assessment.

2.3. Error classification and grading

In the NER model, the classification of errors by degree of severity is based on the extent to which a lack of correspondence between the subtitles and the original audio affects viewers' access to the original meaning, analysed in terms of (independent and dependent) idea units. This ranges from largely inconsequential deviations (minor errors) to isolated information loss (standard errors) to the creation of an utterance with an entirely new meaning that could make sense in the new context (serious errors).

Serious errors (penalized as 1) introduce factual mistakes or misleading information in the respeaking or the recognition phase. An example of a serious RE is the substitution of 15% by 50% in the sentence "The government has reduced public spending by 15%". A serious EE was illustrated in Example 2 above. Standard errors (penalized as 0.5) do not create a new meaning but nevertheless result in the omission of an idea unit from the original text. Standard REs disrupt the coherent flow of the text. The resulting subtitles are evidently erroneous but make it difficult to figure out what was originally meant (e.g., when "Halloween" is recognized as "hell of even"). Finally, minor errors

(penalized as 0.25) allow viewers to follow the meaning or flow of the original text and sometimes even to reconstruct the original words. Minor REs may be caused by the presence or absence of capital letters, apostrophes, insertions of small words, etc. Minor EEs often involve the omission of a dependent idea unit that does not render the remaining unit meaningless or nonsensical.

The difference between standard EEs and minor EEs is based on the distinction between independent and dependent idea units. An independent idea unit, such as "The blaze started this morning at the front of the house", is the oral equivalent of a sentence. It makes sense as a complete, independent utterance and is usually composed of several dependent idea units, such as "this morning" and "at the front of the house". A dependent idea unit is often an adjunct (Eppler & Ozón, 2013), that is, an optional part of the sentence that is not required by the verb and which provides information about the "when", the "where", the "how", etc. of an independent idea unit. Standard EEs often entail the omission of an entire independent idea unit (which may not be noticed by the viewers) or of a dependent idea unit that renders the remaining unit meaningless or nonsensical (e.g., the omission of "last night" in "The rain started last night"), whereas the omission of "this morning" in "The blaze started this morning at the front of the house" would be classified as a minor EE.

2.4. Accuracy and assessment

In the NER model, subtitles are required to reach an accuracy rate of 98 per cent in order to be considered acceptable. Beyond the accuracy rate itself, however, the NER model includes an overall assessment that comprises (a) an analysis of the accuracy rate, including the number of CEs; (b) comments on different issues that are not normally included in the formula (such as the speed and delay of the subtitles, the overall flow of the subtitles on screen, speaker identification, the coherence between the original image/sound and the subtitles, whether or not too much time has been lost in the corrections, etc.); and (c) a final conclusion. Although the conclusion is usually closely linked to the accuracy rate obtained in the formula, in the case of a discrepancy (e.g., when the delay of highly accurate subtitles is considered unacceptable), it is the final conclusion and not the accuracy rate that represents the quality of a given set of subtitles as assessed using the NER model.

The NER model is now being used by regulators, broadcasters, subtitling companies and universities in a number of European countries as well as in Australia, South Africa and the Americas. It is being used in research projects that have provided data about two key criteria to assess its robustness: inter-annotator agreement and the correlation between viewers' subjective rating of subtitling quality and the scores obtained using the model. In the United Kingdom, a two-year project set up by the government regulator Ofcom to assess the quality of live TV subtitles with the model (Ofcom, 2015) found discrepancies between the scores of different annotators of only 0.1 per cent. As for the correlation with subjective rating, recent studies in Poland (Szczygielska & Dutka, 2016) show that NER scores correspond well with viewers' subjective impressions of the quality of live subtitles.

While the NER model has therefore been successfully applied and established for intralingual live subtitling, the question is to what extent it can also be employed in assessing the quality of interlingual subtitles produced live by way of "respeaking" in another language or other methods. Whereas respeaking, as an intralingual operation, can be conceived of as a form of translation in the sense of Jakobson's (1959) intralingual variant, the task of interlingual live subtitling constitutes "translation proper" and can therefore be assumed to share common ground with other forms of (interlingual) translation. With regard to the process, "interlingual respeaking", if one could call it that, is really a form of simultaneous interpreting, while the product, in the intralingual as well as the interlingual variant, is a set of subtitles. With a view to adapting the NER model to the assessment of interlingual live subtitling, we therefore look to audiovisual translation and interpreting studies for relevant insights and models.

3. Quality assessment in subtitling and interpreting

3.1. Quality assessment in (pre-recorded) interlingual subtitling

Much of the work on quality in subtitling can be traced back to early studies on quality in translation, from Juliane House's 1981 *Model for Translation Quality Assessment* to discussions about the units of assessment (Van Leuven-Zwart, 1989) or notions of equivalence (Toury, 1995) to be used in assessing translation quality. As Pedersen points out, though, the notion of translation quality is as elusive as it is conditioned by perspective and point of view:

> 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. (Pedersen, 2017, p. 210)

Official standards regulating quality in translation, such as the European Standard for Translation Services EN 15038, focus mainly on the process and requirements of translation. This is essential for managing translation quality, but the aim of this article is to assess the quality of the translated product. Metrics-based models such as the LISA QA metric, applied to machine translation (MT), and the EU Multidimensional Quality Metrics are useful systems for analysing the translation product, but in their current form they cannot account for the specificities of subtitling.

In a recent survey on quality assurance and quality control in the subtitling industry, Robert and Remael (forthcoming) outline the key quality parameters in subtitling, based on the literature on quality parameters in translation revision and, especially, on the seminal Code of Good Subtitling Practice by Carroll and Ivarsson (1998). They identify four translation quality parameters and four technical parameters for subtitling quality (Robert & Remael, forthcoming):

Translation quality parameters:
- Content and transfer (including accuracy, completeness and logic);
- Grammar, spelling and punctuation;
- Readability (ease of comprehension and coherence between individual subtitles);
- Appropriateness (socio-cultural features of the audience).

Technical parameters:
- Style guide
- Speed
- Spotting
- Formatting.

Though mainly concerned with the different phases and notions of quality assurance as they are implemented in the subtitling industry, the study by Robert and Remael (forthcoming) offers interesting findings regarding, for example, differences in the importance that subtitlers and clients attach to the above parameters. Whereas subtitlers claim to focus on all of them, including content, grammar and readability, clients focus much more on the technical than on the linguistic parameters.

As for product-oriented models for quality assessment in subtitling, perhaps the most common are the in-house guidelines used by subtitling companies. They demonstrate that, although elusive and complex, quality assessment of the subtitled product is not only possible but is in fact carried out daily. Unfortunately, most of these models are for internal use only and have yet to be reviewed and tested in systematic empirical research. More accessible, by contrast, are models used for training in subtitling at university, which can be sets of criteria (rubrics) assigning different weights to individual parameters (see Table 1) or error-based grids that outline different types of error and indicate how they should be penalized (see Table 2).

Table 1: Subtitling evaluation grid designed by Kruger (2008) for North-West University (South Africa) and Macquarie University (Australia)

Parameter	Explanation	Weight
Translation/ editing	Level of equivalence between subtitle and dialogue	20
Division of subtitles	Line-to-line, subtitle-to-subtitle	10
Linguistic	Correct spelling, grammar, research	20
Punctuation	Accuracy, obstruction factor, including the use of dialogue dashes	10
Cueing	Minimum vs maximum length of one-line and two-line subtitles (time allowed to read the subtitle and to take in the image)	20
	Relation to visual/sound rhythm of film, respecting of boundaries (shot, scene, music)	20

Table 2: Linguistic section of the subtitling scoring system designed by José Luis Martí Ferriol and Irene de Higes at Jaume I University (Spain) in 2015

Type of error	Score
Not same sense	0.25
False sense	0.0
No sense	1.00
Contresens	1.00/1.50
Omission	0.50/1.00
Spelling	0.50/1.00
Grammar	0.50
Lexis	0.50
Dialect	0.25
Register	0.50
Pragmatics	0.25/0.50
Textual coherence	0.50
Semiotics	0.50
Style	0.10/0.25
Typography	0.50/1.00
Format	0.50/1.00
Duration (time)	0.50/1.00
Synchronization	0.50/1.00
Gap between subtitles	0.50/1.00
Reading speed	0.50/1.00
Timecodes	0.25

Another interesting proposal is the FAR model recently put forward by Pedersen (forthcoming), which is partially based on the NER model. Pedersen distinguishes between F errors (those related to functional equivalence or how the message/meaning is rendered in the subtitled translation), A errors (those related to acceptance or how the subtitles adhere to target norms) and R errors (readability issues). The FAR model is interesting in several respects: (1) it is viewer-centred, (2) it yields individual scores for different areas that can be used for training and feedback, and (3) it can be localized so that it is applied according to specific norms and guidelines.

The parameters and models used by Carroll and Ivarsson (1998), Robert and Remael (forthcoming), Pedersen (forthcoming) and both Macquarie University and Jaume I University provide useful insights for developing a model of assessment for interlingual live subtitles, not least regarding the distinction between content and form and between linguistic and technical issues, as well as error grading in the assessment of subtitling quality. However, these models have been designed for pre-recorded subtitling and therefore cover aspects that are not so relevant to interlingual live subtitling (such as spotting or segmentation). Indeed, since the process of respeaking is more similar to simultaneous interpreting than to subtitling, an attempt to develop a quality-assessment model for interlingual live subtitling suggests that insights gained from work in the field of interpreting should be drawn on.

3.2. Quality assessment in (simultaneous) interpreting

Attempts at assessing the quality of simultaneous interpreters' output date back to early empirical research by experimental psychologists (Barik, 1969, 1975; Gerver, 1969). Their assumption that "translation" should ideally result in complete correspondence between source-language (SL) and target-language (TL) units was akin to the notion of linguistic equivalence, which dominated the discourse on written translation at the time. The focus on source–target correspondence gave rise to assessment models that comprised various types of deviation, or departure, of the interpretation (target text) from the original (source text).

Gerver's (1969) assessment scheme, for instance, included omissions and substitutions as well as corrections, which could affect words, phrases or longer passages. Interestingly, Gerver's categories were inspired by the "error categories" employed by Carey (1968) in the assessment of monolingual shadowing (i.e., verbatim repetition of auditory input in the same language). Carey had posited four types of error: omissions, substitutions, additions and distortions of words. In adapting these categories to the interlingual task of interpreting, Gerver proposed modifications along two lines. First, he argued that substitutions or corrections in the output were not necessarily errors and therefore opted for the broader term "discontinuities" (Gerver, 1969/2002, p. 54), mainly to accommodate corrections under the same heading as errors. Secondly, though his assessment approach was to count the "number of words correctly translated", he saw the need to go beyond words as units of scoring and allow for paraphrasing. The reason he gives for this could be seen as encapsulating the age-old translation-theoretical dichotomy between literal and free, as Gerver (1969/2002) explains that "a word-for-word translation was not expected and, indeed, would not have been a good translation from the interpreter's point of view" (p. 56).

Gerver's early efforts point to two of the major challenges in the assessment of interpreters' performance quality. One is working with "paraphrase", which goes beyond the lexical level and is an essentially meaning-based category of analysis. Gerver himself encountered this problem in subsequent work and had this to say, parenthetically, about the scoring of a (monolingual) rewriting test, in which candidates for interpreter training being screened for aptitude had to rewrite sentences in two different ways: "In practice, the scoring of this test was found to be subjective and difficult to accomplish, since it involved equating of meaning at the sentence level" (Gerver, Longley, Long, & Lambert, 1989, p. 728)0 In other words, even intralingual paraphrasing proves difficult to score, as demonstrated also in Russo's (2014) work on simultaneous paraphrasing as an aptitude test.

The other major challenge, closely related to that of meaning-based scoring of verbal data, is goodness, or quality, "from the interpreters' point of view" – the subjective judgement of an interpreter's performance as shaped by what Chesterman (1993) referred to as "professional norms".

As highlighted by early controversies around Barik's (1969) work, interpreters have traditionally insisted that they render the speaker's intended message rather than (all) the words actually expressed. The use of appropriate strategies and techniques, such as compression, generalization or stalling for time, is considered an important part of interpreters' professional skills, many of which are needed to cope with the high processing load associated with the task (e.g., Gile, 2009). While these claims were articulated in the interpreting community as early as the 1950s and 1960s, empirical research into the nature of conference interpreters' professional norms regarding output quality started only in the mid-1980s. Many studies were conducted as surveys among interpreters, using the list of quality criteria proposed in the pioneering work of Bühler (1986).

What has emerged from this line of research is that consistency of meaning (or sense) between source and target texts (also referred to as fidelity, or accuracy and completeness) is the most highly valued criterion in assessing the quality of an interpretation, but that aspects of form and delivery, such as fluency and terminological and grammatical correctness, also play an important part. The idea that content- and form-related criteria contribute variably to overall quality judgements, depending on contextual conditions, is now widely accepted (e.g., Gile, 2009). Likewise, it has been stressed that interpreting should be seen – and evaluated – not only as a textual product but also as a professional service (see Pöchhacker, 2001). Where the focus is on the product, criteria such as the accuracy and completeness as well as the correctness of TL expression are central to assessment.

Quality of service, on the other hand, is evaluated more broadly, relying on such criteria as communicative effectiveness and user satisfaction. Between these two conceptual poles of product and service, the emphasis in interpretation quality assessment has tended to shift away from error analysis (see Falbo, 2015) towards a more functional approach to evaluation which takes account of situational requirements and constraints (e.g., Moser-Mercer, 1996). With reference to the three different conceptualizations of quality identified by Grbić (2008), this marks a shift away from seeing quality as perfection (i.e., a product without defects) to the view of quality as "fitness for purpose", as in the ISO definition of a product or service whose inherent characteristics need to fulfil specified requirements. For practical purposes, however (such as certification testing or summative assessment in interpreter education), these two different notions of quality can be reconciled by the appropriate specification of requirements. If, for instance, a testing authority or panel of examiners assumes that certain deviations from the original meaning make the interpreter's product less effective in meeting service users' needs, assessment with an eye to service quality may still (have to) involve error analysis, whether implemented in actual error counts or in ratings of fidelity which penalize errors of content. Such error-based assessment schemes often hark back to Barik's (1969, 1975) typology of omissions, additions and substitutions, the main categories of which are much less controversial than his misguided attempt to subcategorize error types according to different criteria – including the size of the source text unit affected, the probable reason for the deviation (e.g., miscomprehension, delay, "closure") and the degree of severity of the error (e.g., "mild", "substantial", "gross"). Overall, then, a concern with source–target correspondence, or accuracy, based on meaning units rather than words, remains an integral part of assessment in interpreting studies (see Liu, 2015), and can even be seen to form the core of quality-assessment models (Pöchhacker, 2001, p. 413), albeit within a complex understanding of quality as a multidimensional concept in relation to the interpreting product and service.

3.3. Shared ground and specific needs

The NER model, which serves as our point of departure for developing an assessment model for interlingual live subtitling, has altogether different roots compared to quality-assessment approaches in audiovisual translation (subtitling) and interpreting studies. Its origins in attempts to measure accuracy in automatic speech recognition (see section 2.1) imply a focus on individual lexical units (words) and on the relation between spoken input and written output in the same language. In translation (and interpreting) studies, in contrast, notions such as quality or accuracy are applied to linguistic (including semantic) relations between languages. Even so, a concern with words is

evident, and perhaps unavoidable, also in translation studies (and not least in early work on simultaneous interpreting by experimental psychologists), if only as indicators of conceptual meaning relations. Where the actual scoring of interlingual source–target relations is attempted, words have served as the basic elements of "meaning units", with all the theoretical and analytical complications this implies. The complex relations between words and meaning, between linguistic units and idea units, or conceptual information, also pose a challenge to the NER model, even though it deals only with intralingual correspondence. As soon as paraphrasing comes into play, so does meaning as the *tertium comparationis*, and the need for subjective judgement by the analyst. When applied to an interlingual task, the challenge is greatly exacerbated, and gaining some degree of control over subjectivity in scoring becomes a primary concern.

Drawing on insights from work in audiovisual translation and simultaneous interpreting, a viable solution for quality assessment would involve a set of criteria or parameters and/or a definition of observable phenomena. It would also involve a way of grading or weighting their relevance to the goal of the assessment task, which is to issue a judgement on performance quality – that is, the degree to which the product fulfils the requirements specified for the service. The NER model already combines the concern with accuracy, along several dimensions, with an overall, user-oriented view of performance quality. The new model needs to go beyond the notion of "editing" and stipulate criteria for assessing accuracy in the translational relationship between the original audio and the recognized text (subtitles). This involves both source–target correspondence in terms of content and quality of expression in the TL (i.e., form). For both of these main components of translation quality, a set of descriptive categories can be defined, but they must be kept sufficiently clear and concise to ensure ease and consistency of application. Content-related phenomena (corresponding to such quality criteria as sense consistency and completeness) can be accounted for in terms of omissions, additions and substitutions (Barik, 1975); form-related criteria might include correct terminology and grammar as well as appropriate style (Bühler, 1986). All these deviations from source–target correspondence or formal adequacy, which, in the tradition of the NER model, might be referred to as "errors", are to be scored for severity. A number of other aspects of quality were mentioned in the above review of the literature, but it goes without saying that both pre-recorded subtitles and simultaneous interpretations involve features that apply to interlingual live subtitling to a different extent, if at all. Intonation in interpreting is a case in point, whereas accent and diction do come into play in the recognition phase.

The aim of remaining as close as possible to both the basic architecture (see section 2.2) and the scoring conventions of the NER model is not only a pragmatic choice but one that corresponds to specific industry requirements. Whereas the model may well be applicable to interlingual speech-to-text interpreting more generally, the initial target group are users in the live subtitling industry, where the NER model is quite well established. This industry also specifies a high accuracy rate as a requirement for subtitles of acceptable quality (see section 2.4), so that a similar metric or formula needs to be used.

4. The NTR Model

In the light of the above considerations, the NTR model draws heavily on the NER model. It uses a similar formula and the same error grading, but the E for EEs is exchanged for T, which accounts for errors of translation:

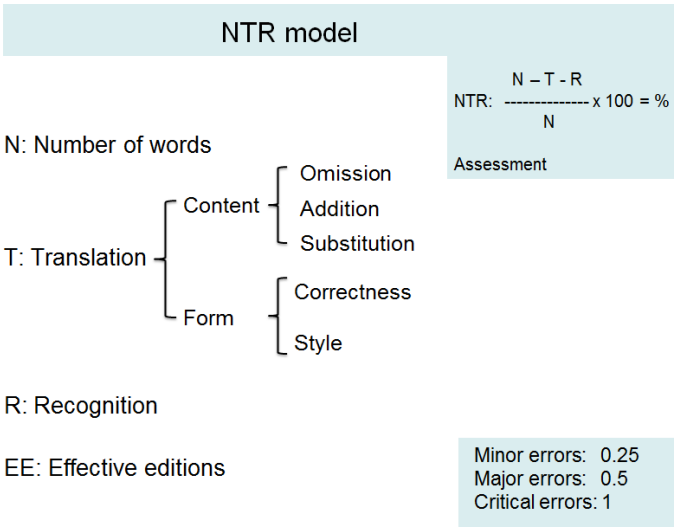

Figure 3: The NTR model

As in the NER model, N is the number words in the subtitles. The main distinction is now made between T and R, which provides information about whether the issues found in the subtitled text are related to the subtitler's ability to render the content of the source text in an adequate TL form (T) or to their interaction with the software (R). In respeaking-based live subtitling, a T error lies in what the "respeaker" has said, whereas an R error is not in what has been said but rather in how it has been (mis)recognized by the speech-recognition software. T errors are subdivided into content and form: content errors can be omissions, additions or substitutions (typically mistranslations); form errors can affect the correctness of the subtitles (grammar, terminology) or their style (appropriateness, naturalness, register). Both translation errors (TE) and REs are classified by their degree of severity, using the three-level grading system and scoring rules of the NER model. However, two of the three grades have been relabelled, in line with the terminology of the LISA QA metric: "standard errors" in the NER model are now referred to as "major errors", capitalising on the antonymic relation to "minor errors"; and "serious" errors are now labelled as "critical", which offers better contrast with "major" than "serious". The category for deviations from the source text that do not involve a loss of information or that even enhance the communicative effectiveness of the subtitles is now called "effective editions" (EE). Such instances of presumably strategic solutions on the part of the subtitler are not factored in to the formula for the calculation of the accuracy rate, but form part of the overall assessment.

As in the NER model, 98 per cent is set as the minimum accuracy rate for interlingual live subtitles to be acceptable with the NTR model. However, since such high percentages are more common in intralingual than in interlingual live subtitling, the new model recalculates the accuracy rate to a more standard 10-point scale (Table 3).

As with the NER model, assessment based on the NTR model comprises several different aspects, namely: (a) an analysis of the accuracy rate; (b) comments on issues not included in the formula, such as effective editions, the speed, delay and overall flow of the subtitles, speaker identification, the coherence between the original image/sound and the subtitles, whether too much time has been lost in the corrections, etc., and (c) a final conclusion. In the case of a discrepancy between the accuracy rate and the conclusions from the broader assessment (for example, when the delay of highly accurate subtitles is considered unacceptable), it is the final conclusion and not the accuracy rate that represents the quality of a given set of subtitles as assessed with the NTR model.

Table 3: Recalculation of accuracy rate to a 10-point scale

Accuracy (%)	10-point scale
< 96	0/10
96.4	1/10
96.8	2/10
97.2	3/10
97.6	4/10
98.0	5/10
98.4	6/10
98.8	7/10
99.2	8/10
99.6	9/10
100	10/10

5. Discussion

Beyond the rationale provided in section 3.3 for the adoption of the NTR model as an elaboration of the NER model and for the proposals for quality assessment in (audiovisual) translation and, in particular, interpreting studies, to be incorporated, there are three issues that merit some further discussion: (1) the move from intralingual edition to (interlingual) translation, (2) the error-grading scheme, and (3) the issue of subjectivity.

5.1. From E (edition) to T (translation)

The NTR model replaces the category of edition (E) with that of translation (T). This is done to acknowledge that the move from intra- to interlingual processing results in a significantly different cognitive task and should not be viewed as a process that is essentially similar to the added complication of language switching. (The latter view would be reminiscent of early psychological studies equating the process of simultaneous interpreting to that of monolingual shadowing.) In a previous attempt to adapt the NER model to interlingual live subtitles, Soria (2016) proposed NERT, therefore making a distinction between E (editing errors) and T (translation errors). It is difficult to see, however, how (interlingual) editing can be distinguished from changes in meaning or content. For instance, compression-related "editing" is likely to result in some loss of meaning and therefore will affect the translated content. Moreover, the notion of editing carries the implication of something that is done deliberately – as in the case of the "effective editions" accounted for in the NTR model.

Since "interlingual respeaking" can be assumed to involve a distinctly higher cognitive load, time pressure is likely to demand recourse to strategic compression (Kalina, 2015), but it would not be evident whether something was left out deliberately or inadvertently as a result of processing overload (Gile, 2009). The NTR model largely avoids such analytical uncertainty (E vs T), whereas the distinction between T and R serves to fulfil the requirement (number 6 in section 2.1) of providing an overall idea of aspects to be improved, which can suggest two different types of remedial action, depending on whether the subtitler has experienced difficulty with translating or with being recognized.

5.2. Error grading

Despite the use of new labels, the error-grading system remains the same as in the NER model. Minor translation errors cause a certain loss of meaning, but some of the content is still there. Major translation errors deprive the viewer of a correct understanding of an idea unit, the source-text content being lost. Critical errors change the meaning of the source text and introduce factual mistakes or misleading information that could make sense in the new context. This approach to error grading effectively combines two perspectives on the analysis, that is, the intertextual one, which focuses on deviations in meaning, and the reception-oriented one, which deals with what users (can) do with the (erroneous) meaning.

While some assessment models for subtitling and interpreting assign different weights to content and form, typically more to the former than the latter, the three-level grading scale of the NTR model is applied regardless of whether there is an issue of translation (content or form) or recognition. Differences in the relative prevalence of errors of a certain category or grade can nevertheless be expected. Therefore, research on the NER model (Ofcom, 2015) indicates that the percentage of "minor errors" is higher for R than for E, and the analyses carried out to date with the NTR model show that TEs of form (whether related to correctness or style) tend to be classified as minor, which suggests that the viewers find them fairly easy to handle. Potentially, TEs of form can also be major if they cause the viewers to lose the content of the source text unit completely. What would not really be conceivable, though, are TEs of form classified as critical. By definition, critical errors change the meaning of the source text, introducing a new meaning that could make sense in the new context. They would therefore be regarded as errors of content rather than of form. By way of example, the rendering of "Tercera Guerra Mundial" (Third World War) as "third world war" in a context dealing with developing countries represents a serious (content) error that is caused by an error of form (i.e., the use of lower case instead of capitals). Unlike TEs of form, which are therefore either minor or major, TEs of content normally require consideration across the three grades of severity.

A relevant issue that arose with the NER model and which also applies to the NTR model concerns the scoring of errors affecting several dependent idea units within an independent unit, as in the case of the omission of ingredients in a recipe. In this case, the omission of three out of, say, nine ingredients will be penalized as 0.75, corresponding to three minor omissions (as instances of TEs of content). If the subtitler had omitted the whole sentence (i.e., an independent unit), however, this would be marked only as a single major TE of content (omission; 0.5). Regarding the score for this element, a subtitler would therefore be better off omitting the sentence altogether than conveying it with three of the ingredients missing. The rationale behind this is that if the entire sentence is omitted, the viewers cannot get the wrong recipe. If they do get the sentence containing the three errors, every error has an impact on the message they are receiving, and the model strictly penalizes the individual omissions. On the other hand, when subtitlers are seen to resort to omissions of full independent idea units frequently, this is both reflected in the accuracy score and considered in the overall assessment. In cases where the subtitler is forced to omit parts of a sentence, it may be a good idea to use the symbol "(…)", which is currently used by some broadcasters in the United Kingdom to let viewers know that some information is missing. After all, one of the most important factors regarding viewers' comprehension is whether or not they are aware that an error has occurred, and the use of parentheses makes it clear that it is not the speaker but the subtitler who has left the sentence unfinished.

5.3. Subjectivity

As is evident from the literature on quality assessment, subjectivity in scoring is one of the key challenges to any assessment model. In the NER model, subjectivity was tackled by means of evaluator training as well as by analysing inter-annotator agreement and the correlation between NER scores and users' subjective ratings. As explained above (section 2.4), inter-annotator agreement as well as correlations between user ratings and NER assessments have been found to be

high. Initial testing of the NTR model with 10 evaluators has shown promising results, with an average discrepancy of 0.3 on a scale from 1 to 10. Most of the discrepancies concerned errors of form, and in particular style. Since most of these errors are minor (penalized as 0.25), they have little impact on the overall accuracy rate. In a ten-minute programme with an average of 1500 words, for instance, a discrepancy between evaluators regarding one minor error would cause a variation of only 0.025 in the score across the 1–10 scale. In other words, for the score to change from, say, 6/10 to 7/10, the evaluators would need to disagree in 40 instances of minor errors.

6. Application of the model

This section offers an example of how the NTR model can be applied to assessing interlingual live subtitling. The material used is made up of an excerpt from a narrative stretch of discourse in the film *Shawshank Redemption* (1995) respoken live from English into Spanish by a master's student in Audiovisual Translation at the University of Roehampton. The exercise was part of the Respeaking module, the last unit of which was devoted to interlingual respeaking. The students had the opportunity to preview the excerpt before undertaking their respeaking task. The coding used in the analysis is shown in Table 4.

Table 4: Error coding scheme

Translation errors				
of Content:	– omission	T (Cont-omiss)	Minor (MinT)	(0.25)
	– addition	T (Cont-add)	Major (MajT)	(0.5)
	– substitution	T (Cont-subs)	Critical (CritT)	(1)
of Form:	– correctness	T (Form-corr)	Minor (MinT)	(0.25)
	– style	T (Form-style)	Major (MajT)	(0.5)
Recognition errors				
			Minor (MinR)	(0.25)
			Major (MajR)	(0.5)
			Critical (CritR)	(1)
Effective editions		EE		

Table 5 : Example

Note: Effective editions are underlined and errors appear in italics, followed by a parenthesis with the correct form.

Original text (transcribed audio)	Respeaking-based subtitles	Errors
Not long after the warden deprived us of his company, I got a postcard in the mail. It was blank. But the postmark said, “Fort Hancock, Texas”. Fort Hancock, right on the border. That's where Andy crossed. When I picture him in his own car with the top down, it makes me laugh all over again ... Andy Dufresne, who crawled through a river of shit and came out clean on the other side. Andy Dufresne, headed for the Pacific. Those of us who knew him best talk about him often. I swear, the stuff he pulled ... Sometimes it makes me sad, though, Andy being gone. I have to remind myself that some birds aren't meant to be caged. Their feathers are just too bright ... and when they fly away, the part of you that knows it was a sin to lock them up does rejoice. But still, the place you live is that much more drab and empty that they're gone. I guess I just miss my friend.	Poco después de que te fueras (el alguacil nos privase de su compañía)[1], recibí una postal (-en el correo)[2]. Estaba en blanco, pero en el matasellos ponía: Fort Hancock, Texas. Por otra (Fort)[3] Hancock, justo en la frontera. Por ahí cruzó Andy. Cuando le imagino dirigiéndose al sur en su descapotable, no puedo evitar reírme. Andy Dufresne, que atravesó un río de mierda y salió completamente limpio por la otra orilla. Andy Dufresne, rumbo al Pacífico. Quienes mejor lo conocíamos hablamos de él a menudo. (-I swear, the stuff he pulled)[4]. Sin embargo, a veces me entristece que se haya ido y me recuerdo a mí mismo que hay pájaros que no están hechos para ser enjaulados. Sus plumas son demasiado brillantes. Y cuando salen volando, la parte de ti que sabía que era un pecado tenerlos encerrados se alegra, pero aun así el lugar en el que vives está mucho más triste y vacío porque se han ido. Supongo que sólo[5] echo de menos a mi amigo.	1. CritT (cont-subs) (1): Contresense: “you left” instead of “the warden left” or “the warden deprived us of his company”. The target text introduces a new, mis-leading meaning that makes sense in the new context. 2. EE: No relevant information is lost here by omitting “en el correo” (“in the mail”). 3. MinR (0.25): Despite the error, the viewer may be able to recognize the source text unit given that it was mentioned in the previous sentence. 4. MajT (cont-omiss) (0.5): Omission of a full independent unit/sentence. 5. MinT (form-style) (0.25): This is a calque from English, since the word “sólo” would not normally be used here. It is not incorrect but rather unnatural/unidiomatic and it does not prevent the viewers from understanding the source text.

Accuracy rate	
MinT: (1 × 0.25 = 0.25) (form-style) MajT: (1 × 0.5 = 0.5) (cont-omiss) CritT: (1 × 1 = 1) (cont-subs) --- Total: 1.75	MinR: (1 × 0.25 = 0.25) MajR: 0 CritR: 0 --- Total: 0.25
$\text{NTR accuracy rate} = \frac{162 - 1.75 - 0.25}{162} \times 100 = 98.76\% \ (7/10)$	

Assessment
The quality of the subtitles is good. In terms of accuracy, these would have been very good subtitles if it were not for the occurrence of a critical translation error that introduces misleading meaning in the first paragraph. Still, the rest of the translation conveys the meaning and largely the style of the source text. The average delay of the subtitles (5 s) is acceptable and there are no corrections and no issues regarding speaker identification.

7. Conclusion

Quality in translation and interpreting is a complex construct that has triggered many scholarly discussions over the past decades. Models of quality assessment developed in academia are sometimes criticized for being overly complicated and difficult to operationalize. The NTR model presented in this article, which represents an adaptation of the widely established NER model developed by the first author, is designed to assess the quality of interlingual live subtitling, a new translational modality that is currently in high demand and greatly in need of methods for training and quality assessment.

Rather than errors of edition, the NTR model focuses on translation and considers both informational deviations from source-text content (in terms of omissions, additions and substitutions) and errors of form (correctness, style), as gleaned from a review of quality-assessment models in subtitling and interpreting. While the distinction of various error types and categories does not have a differential impact on the final score, it helps to identify problem areas requiring specific remedial actions. These could range from improving comprehension of the source text and the way its content has been dealt with in translation (translation errors of content) to fine-tuning correctness and stylistic appropriateness in the TL (TEs of form) and to engaging in further training with the speech-recognition software to improve the subtitler's voice profile (REs). The nature of the new interlingual live subtitling modality may sometimes make it difficult to establish whether an error was made by the subtitler (T) or by their interaction with the software (R). In those cases, it will be necessary to listen to the recording of the subtitler's spoken output in order to identify the error correctly.

The error typology and grading system in the NTR model is kept relatively simple and concise. Adding further subcategories or finer grading scales would make the model's application even more susceptible to subjective scoring decisions. This subjectivity, which is also present in the NER model but is greatly exacerbated by the interlingual nature of the new modality, must be acknowledged as an intrinsic challenge in any assessment of translation performance, whether it is expressed as a numerical score (like the accuracy score in the NTR model) or a more holistic appreciation of performance features based on relevant professional expertise. The latter is an integral part of quality assessment with the NTR model, and permits the analyst to consider a range of issues that are not formalized in numerical terms. The NTR model can, of course, be developed to incorporate some of these issues as part of the accuracy rate formula. This was done with the NER model by regulators and subtitling companies in the United Kingdom and Australia, which factored subtitling speed and instances of subtitles blocking important visual elements on the screen into the score. At any rate, and even though the experience with the NER model to date has shown that the focus is usually placed on the accuracy rate, it is important to emphasize that in both the NER and the NTR model it is the overall assessment and not the accuracy rate that represents the quality of the subtitles under scrutiny.

As with most quality-assessment schemes, the task ahead is to enhance the reliability of the model's application by improving training in how to use the model and by investigating both inter-annotator agreement and the correlation between NTR scores and users' subjective ratings. Initial analyses are showing promising results, not least because most of the discrepancies in scoring concern minor errors, which have a limited impact on the final score. Clearly, though, much work

remains to be done to test and fine-tune the model for application to different language pairs. Meanhile, it should be borne in mind that the model is not an end in itself but a tool designed to help us assess the quality of a new form of translation that can support accessibility for different user groups and cultural communities.

References

Apone, T., Brooks, M., & O'Connell, T. (2010). Subtitle Accuracy Metrics Project. Subtitle viewer survey: Error ranking of real-time subtitles in live television news programmes. WGBH National Center for Accessible Media, Boston. http://ncam.wgbh.org/invent_build/analog/subtitle-accuracy-metrics (last accessed 1 July 2015).

Arumí Ribas, M., & P. Romero-Fresco (2008). A practical proposal for the training of respeakers. *Journal of Specialised Translation, 10*, 106–127. Available online: http://www.jostrans.org/issue10/art_arumi.php (last accessed 19 November 2016).

Barik, H. C. (1969). *A study of simultaneous interpretation*. Doctoral dissertation, University of North Carolina, Chapel Hill.

Barik, H. C. (1975/2002). Simultaneous interpretation: Qualitative and linguistic data. In F. Pöchhacker & M. Shlesinger (Eds.), *The interpreter studies reader* (pp. 79–91). London: Routledge.

Bühler, H. (1986). Linguistic (semantic) and extra-linguistic (pragmatic) criteria for the evaluation of conference interpretation and interpreters. *Multilingua, 5*(4), 231–235.

Carey, P. W. (1968). *Delayed auditory feedback and the shadowing response*. Doctoral dissertation, Harvard University.

Carroll, M., & Ivarsson, J. (1998). *Code of good subtitling practice*. Berlin: European Association for Studies in Screen Translation.

Chafe, W. (1985). Linguistic differences produced by differences between speaking and writing. In D. Olson, N. Torrance, & A. Hildyard (Eds.), *Literacy, language, and learning: The nature and consequences of reading and writing* (pp. 105–122). Cambridge: Cambridge University Press.

Chesterman, A. (1993). From 'is' to 'ought': Laws, norms and strategies in translation studies. *Target, 5*(1), 1–20.

CSA. (2011). Charte relative à la qualité du sous-titrage à destination des personnes sourdes ou malentendantes, Paris: Conseil Supérieur de l'Audiovisuel. Available online: http://www.csa.fr/Espace-juridique/Chartes/Charte-relative-a-la-qualite-du-sous-titrage-a-destination-des-personnes-sourdes-ou-malentendantes-Decembre-2011 (last accessed 19 November 2016).

Dumouchel, P., Boulianne, G., & Brousseau, J. (2011). Measures for quality of closed captioning. In A. Şerban, A. Matamala, & J.-M. Lavaur (Eds.), *Audiovisual translation in close-up: Practical and theoretical approaches* (pp. 161–172). Bern: Peter Lang.

Eppler, E. D., & Ozón, G. (2013). *English words and sentences: An introduction*. Cambridge: Cambridge University Press.

Eugeni, C. (2006). Introduzione al rispeakeraggio televisivo. In C. Eugeni & G. Mack (Eds.), *Intralinea*, Special Issue on Respeaking. Available online: http://www.intralinea.org/specials/article/Introduzione_al_rispeakeraggio_televisivo (last accessed 19 November 2016).

Eugeni, C. (2008). Respeaking the TV for the Deaf: For a real special needs-oriented subtitling. *Studies in English Language and Literature, 21*, 37–47.

Eugeni, C. (2009). Respeaking the BBC News: A strategic analysis of respeaking on the BBC. *The Sign Language Translator and Interpreter, 3*(1), 29–68.

European Committee For Standardization (2006). *European Standard EN 15038. Translation services – Service requirements*. Brussels: European Committee for Standardization.

Falbo, C. (2015). Error analysis. In F. Pöchhacker (Ed.), *Routledge encyclopedia of interpreting studies* (pp. 143–144). London: Routledge.

Gerver, D. (1969/2002). The effects of source language presentation rate on the performance of simultaneous conference interpreters. In F. Pöchhacker & M. Shlesinger (Eds), *The Interpreter Studies reader* (pp. 53–66). London: Routledge.

Gerver, D., Longley, P. E., Long, J., & Lambert, S. (1989). Selection tests for trainee conference interpreters. *Meta, 34*(4), 724–735.

Gile, D. (2009). *Basic concepts and models for interpreter and translator training* (Rev. ed.). Amsterdam: John Benjamins**.**

Grbić, N. (2008). Constructing interpreting quality. *Interpreting, 10*(2), 232–257.

House, J. (1981). *A model for translation quality assessment.* Tübingen: Gunter Narr.

International Telecommunication Union (ITU). (2015). *Accessibility terms and definitions: Series F: Non-telephone telecommunication services. Audiovisual services*. Geneva: ITU.

Jakobson, R. (1959/2000). On linguistic aspects of translation. In L. Venuti (Ed.), *The translation studies reader* (pp. 113–118). London: Routledge.

Kalina, S. (2015). Compression. In F. Pöchhacker (Ed.), *Routledge encyclopedia of interpreting studies* (pp. 73–75). London: Routledge.

Kruger, J.-L. (2008). Subtitler training as part of a general training programme in the language professions. In J. Díaz Cintas (Ed.), *The didactics of audiovisual translation* (pp. 71–88). Amsterdam: John Benjamins.

Lambourne, A. (2006). Subtitle respeaking. In C. Eugeni & G. Mack (Eds.), *Intralinea*, Special Issue on Respeaking. Available online: http://www.intralinea.org/specials/article/Subtitle_respeaking (last accessed 19 November 2016).

Leuven-Zwart, K., van (1989). Translation and original: Similarities and dissimilarities. *Target, 1*(2), 151–181.

Liu, M. (2015). Assessment. In F. Pöchhacker (Ed.), *Routledge encyclopedia of interpreting studies* (pp. 20–22). London: Routledge.

Marsh, A. (2006). Respeaking for the BBC. In C. Eugeni & G. Mack (Eds), *Intralinea*, Special Issue on Respeaking. Available online: http://www.intralinea.org/specials/article/Respeaking_for_the_BBC (last accessed 19 November 2016).

Martí Ferriol, J. L. & de Higes Andino, I. (2015). *Guía de corrección para subtitulación*. Unpublished class material. Universitat Jaume I, Castelló, Spain.

Moser-Mercer, B. (1996). Quality in interpreting: Some methodological issues. *The Interpreters' Newsletter, 7*, 43–55.

Ofcom (2015). Measuring live subtitling quality: Results from the fourth sampling exercise. London: Office of Communications. Available online: https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/tv-research/live-subtitling/sampling_results_4 (last accessed 19 November 2016).

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

Pöchhacker, F. (2001). Quality assessment in conference and community interpreting. *Meta, 46*(2), 410–425.

Rajendran, D. J., Duchowski, A. T., Orero, P., Martínez, J., & Romero-Fresco, P. (2012). Effects of text chunking on subtitling: A quantitative and qualitative examination. *Perspectives: Studies in Translatology, 21*(1), 5–21.

Remael, A., & van der Veer, B. (2006). Real-time subtitling in Flanders: Needs and teaching. In C. Eugeni & G. Mack (Eds.), *Intralinea*, Special Issue on Respeaking. Available online: http://www.intralinea.org/specials/article/Real-Time_Subtitling_in_Flanders_Needs_and_Teaching (last accessed 19 November 2016).

Robert, I. S., & Remael, A. (forthcoming). Quality control in the subtitling industry: An exploratory survey study. *Meta, 61*(3).

Romero-Fresco, P. (2009). More haste less speed: Edited vs. verbatim respeaking. *Vigo International Journal of Applied Linguistics, VI*, 109–133.

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

Romero-Fresco, P. (2016). Accessing communication: The quality of live subtitles in the UK. *Language & Communication, 49*, 56–69.

Romero-Fresco, P. (forthcoming). Respeaking: Subtitling through speech recognition. In L. Pérez-González (Ed.), *The Routledge handbook of audiovisual translation studies*. London: Routledge.

Romero-Fresco, P., & Martínez, J. (2015). Accuracy rate in live subtitling: The NER model. In J. Díaz Cintas & R. Baños (Eds.), *Audiovisual translation in a global context: Mapping an ever-changing landscape* (pp. 28–50). Basingstoke: Palgrave Macmillan.

Russo, M. (2014). Testing aptitude for interpreting: The predictive value of oral paraphrasing, with synonyms and coherence as assessment parameters. In F. Pöchhacker & M. Liu (Eds), *Aptitude for interpreting* (pp. 129–145). Amsterdam: John Benjamins.

Soria, E. (2016). *A proposal to assess accuracy for interlingual respeaking (EN>ES). Adapting the NER Model*. MA dissertation, University of Roehampton.

Szczygielska, M., & Dutka, Ł. (2016). Live subtitling through automatic speech recognition vs. respeaking: Between technical possibilities and users' satisfaction. Presentation at *Language and the Media*, 3 November 2016, Berlin.

Toury, G. (1995). *Descriptive translation studies – and beyond*. Amsterdam: John Benjamins.

Vinay, J.-P., & Darbelnet, J. (1958/2000). A methodology for translation. (Trans. J. C. Sager & M.-J. Hamel). In L. Venuti (Ed.), *The translation studies reader* (pp. 84–93). London/New York: Routledge.

Acknowledgement

This research has been conducted within the framework and with the support of the Spanish government-funded projects 'Inclusión Social, Traducción Audiovisual y Comunicación Audiovisual' (FFI2016-76054-P), '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 EU-funded projects 'MAP: Media Accessibility Platform for the European Digital Single Market' (COMM/MAD/2016/04) and 'Interlingual Live Subtitling for Access' (2017-1-ES01-KA203-037948).

---

1 Because of time constraints, the everyday internal quality assessment of intralingual live subtitles in many companies is based on the analysis of the subtitles only (Romero-Fresco, 2011). Comparison with the original speech is limited to spot checks when there is time to transcribe the programme.