SENDEBAR

Sendebar. Revista de Traducción e Interpretación
https://revistaseug.ugr.es/index.php/sendebar
Publisher: Universidad de Granada

Original Articles · Artículos originales

Joining Forces for Quality Assessment in Simultaneous Interpreting: the NTR Model

Aunando esfuerzos para la evaluación de la calidad en interpretación simultánea: el Modelo NTR

Luis Alonso-Bacigalupe 0000-0002-3020-4536

Universidade de Vigo

ABSTRACT

Quality in simultaneous interpretation (SI) has always been an elusive concept, and the literature has usually argued that it is not easy to establish a valid instrument for the assessment of any interpretation assignment. The provision of accessibility services for the deaf and hard-of-hearing in live audiovisual subtitling, however, has routinely demanded quality assessment procedures of both the intralingual and interlingual live subtitles shown on the screen. In speech-to-text interpreting (STTI), which includes both intra- and interlingual respeaking (i.e. live subtitling without and with translation respectively), the NTR Model was proposed for the assessment of interlingual respeaking (Romero-Fresco & Pöchhacker, 2017). The purpose of this contribution is (i) to argue why the NTR Model might be conceptually valid for the assessment of SI, (ii) to present the results of a small-scale analysis of an SI task evaluated with four instruments aimed at scrutinizing the benefits of this Model and its applicability to SI, and (iii) to advocate for the use of objective assessment systems for student and professional performance in SI.

Keywords: simultaneous interpretation (SI), quality assessment, intralingual and interlingual live subtitles, intralingual and interlingual respeaking, speech-to-text-interpreting (STTI), the NTR Model

RESUMEN

El concepto de calidad en interpretación simultánea (IS) ha sido siempre difuso y la literatura se ha dedicado a reflexionar sobre por qué no es posible establecer un instrumento válido para medir la calidad de cualquier encargo de interpretación. Por su parte, los servicios de accesibilidad para personas sordas en el campo del subtitulado audiovisual han demandado métodos de medición de la calidad de los subtítulos en directo que se proyectan en pantalla. En Interpretación de Voz a Texto, actividad que comprende tanto el rehablado intra- como el interlingüístico (es decir, subtitulado en directo sin y con traducción) se ha propuesto el Modelo NTR (Romero-Fresco y Pöchhacker, 2017) para la evaluación del rehablado interlingüístico. El objetivo de este artículo es: a) argumentar por qué el NTR podría ser conceptualmente válido para su aplicación a la IS; b) presentar los resultados de un pequeño análisis de una tarea de IS que fue evaluada con 4 instrumentos distintos, en un intento por desvelar los beneficios del modelo NTR y su aplicabilidad a la IS; y c) defender la necesidad del uso de métodos objetivos en la valoración del rendimiento de estudiantes y profesionales.

Palabras clave: interpretación simultánea (IS), evaluación de la calidad, subtitulación en directo intralingüística e interlingüística, rehablado intralingüístico e interlingüístico, interpretación de voz a texto, el Modelo NTR

Article info

Corresponding author:
Luis Alonso-Bacigalupe
lalonso@uvigo.es

Publication history:
Received: 15/01/2023
Reviewed: 26/02/2023
Accepted: 01/07/2023

Authorship Contribution Statement:
All authors have contributed to the manuscript equally.

Conflict of interest:
None.

Funding:
The author(s) received no financial support for the research, authorship, and/or publication of this article.

How to cite:
Alonso-Bacigalupe, L. (2023). Joining Forces for Quality Assessment in Simultaneous Interpreting: the NTR Model, *34,* 198-216.
https://doi.org/10.30827/sendebar.v34.26860

1. Introduction and background

The issue of how to define –let alone how to measure– quality in interpretation is complex. Interpreting is about immediate, mediated and situated communication between people or groups of people who do not understand each other, which means that apart from the linguistic, terminological, conceptual and technical issues and hurdles of all kinds which may emerge in the course of an international meeting, there are a number of other factors which may contribute to felicitous or infelicitous communication beyond mere linguistic transfer or the efficient conveyance of propositional information.

This is particularly relevant and pressing in the case of simultaneous interpreting, an activity that is strongly constrained by two crucial limitations: (1) the availability of processing capacity in the interpreters' minds, and (2) time, which means that resorting to strategies that may involve a certain loss of information or some kind of trade-off with other aspects of live communication may be seen as a legitimate operation, often applied in practice by professionals in real-life simultaneously interpreted assignments.

The purpose of this paper is to advocate for the application of objective assessment instruments for the evaluation of SI, and, more specifically, to argue why the instruments already developed for the evaluation of interlingual live subtitling may effectively be extrapolated to SI.

1.1. Quality issues in interpreting

According to Daniel Gile (1995, 2009), simultaneous interpreting (SI) is an activity strongly constrained by the availability of cognitive resources. Human capacity for information processing is not unlimited, and some of our mental operations consume a certain amount of energy, i.e. of cognitive resources. SI includes three basic non-automatic mental operations or efforts: a listening and comprehension effort, a memory effort and a production effort, all of them coordinated by the operations of a central control system or coordination effort. Whenever the processing capacity available is below the amount of effort that needs to be devoted to each one of the efforts, or to all the efforts at the same time, there will be flaws in the flow of information. Problem-triggers include segments with high information density –for example, those containing proper names, acronyms and complex figures–, a fast speech rate, a foreign accent or an unclear diction, among others. Therefore, interpreters are forced to provide immediate professional high-quality solutions to complex problems of all kinds on the spot, while the source text 'unfolds' (see Shlesinger 1995 below) relentlessly. Consequently, SI is (at least partly) approached as a problem-solving activity where it is legitimate to resort to radical crisis-management strategies (including, for example, omissions or generalizations) aimed at avoiding interruptions in the constant flow of information that is to be expected from professional interpreters, despite the fact that such immediate solutions might not be the best choices, which leads us to the issue of time.

Miriam Shlesinger (1995) argued that SI is strongly constrained by three basic limitations: time, the linearity of the ST (or the 'unfolding-text') and the amount of knowledge shared by delegates and interpreters. The interpreter has no control over the time aspect, since it is the speed of delivery of the original speaker which will determine the speed at which the interpreter will be required to process the text. On the other hand, the unfolding-text limitation involves

that in SI the interpreter will produce his/her interpreted version based on short processing units (often at the clause level), unaware, thus, at that point of what the full contents of the ST will be. Finally, the unshared knowledge limitation means that interpreters are confronted with the paradox of establishing communication between top-level experts in all types of specialist fields of knowledge with the handicap that the interpreter is most of the times a non-expert (sometimes the only non-expert in the room) on the topic in hand.

Shlesinger (1999: 66) claimed that many of the rules of thumb of SI are based on a number of specific unwritten norms. For instance, those which contend that brevity and, above all, the use of condensation strategies is preferred to excessive verbatim translation, where minor omissions or even the loss of unnecessary utterances that may have a negative effect on the clarity of the information conveyed are sanctioned or 'condoned' by the audience (Shlesinger, 1999: 68), or where examples can be simply transformed into the main message intended by the speaker: "it is legitimate for an interpreter to telescope examples, to convey the order of magnitude of a series of numbers rather than the numbers themselves" (Padilla and Martin, 1992: 200). Shlesinger (1999: 70) concluded the following: "Judging by the responses of the AIIC User Expectations Survey (Moser, 1995), it appears that users do want to receive the full substance of the speaker's text, but do not want to be bombarded with words".

Therefore, a successful interpretation performance does not rest exclusively on the ability of interpreters to convey all the detailed information. Rather, there are other factors beyond propositional information that are equally relevant for felicitous communication during a simultaneously interpreted event, thus, truly meeting users' expectations and needs.

Looking into the specific realm of media interpreting, Kurz and Pöchhacker (1995) analysed user preferences among TV people compared with the preferences of users of conference interpretation services. They concluded that the former considered that full information transfer may not be as important as other aspects such as fluency of delivery, native accent or a pleasant voice. This idea was shared by other researchers such as Russo (1995: 343), for whom:

> The TV viewers' and radio listeners' expectations are so high that the interpreter ought to become a performer rather than just a linguistic/cultural mediator. Paramount importance is attached to factors such as: voice quality, a cohesive and coherent language and a lively and self-confident performance, often to the detriment, if necessary, of the fidelity or completeness of the original message.

In the same vein, Mizuno (1997: 192) proposed a brief list of the main functions ("interpreters are required to") of TV interpreters and did not include any reference to completeness of information, while issues revolving around quality of the output took a leading position in his catalogue of priorities.

This was also underlined in Collados' 1998 research, dealing with quality in international conferences. She put to the test the relevance of a factor different from propositional meaning in an attempt to identify aspects to which conference delegates assign particular importance. By applying a mixed approach where users' expectations were checked against users' judgements of specific interpretation tasks, she introduced the Granada paradigm in interpretating studies (Pöchhacker, 2013: 39), and found that marked intonation was considered an important component of quality.

Finally, Kahane (2000) wrote about that elusive concept of quality in interpreting, referring once more to Moser-Mercer's 1995 study, and acknowledged no further progress in attempts to measure quality in SI:

> we live in an era obsessed with quality control. With that in mind, in 1995 AIIC launched a study on the quality expectations of interpretation users (1) The amazing thing is that there is no such consensus. Granted, users and interpreters agree on certain quality criteria, but significant differences remain as to nuances, and especially as to the very essence of the elusive concept of quality; quality for whom, assessed in what manner?

However, no matter how much complex this discussion may be, audiences do know when they are exposed to a high level of quality in simultaneously interpreted events. Also, interpreters are fully aware of the features of successful communication in SI-mediated contexts. Interpreting scholars, on their side, are not only aware of this, but rather they need to be able to define and describe quality and to produce an assessment model to measure quality for the benefit of their students, not just because the quality of their work will be assessed every day in the SI lab, but also because they need to know how their performance will be evaluated and judged by clients during professional practice.

1.2. Defining and Measuring Quality

The first attempts to define quality came with the initial stages of the profession. For example, Hebert (1952) and Seleskovitch (1978) already made lists of the features or requirements of prospective successful interpreters. These were not a catalogue of potential interpreter errors, but pointed to the main problems that may be encountered by professionals in the processing of texts in real-life situations; for example, the need to have 'nerves of steel', which refers to the high levels of stress that interpreters are subject to and to the requirement of showing maximum confidence and full control of the situation, without for example mumbling or hesitating. There have been references about the trade-off between accuracy and style (Herbert, 1952: 34), or potential differences in quality indicators depending on the kind of meeting at hand (Herbert, 1952: 82).

The earliest attempt to measure quality objectively was probably that of Barik (1971, 1973). He analysed the quality of interpretation in terms of three types of potential errors: omissions, additions and substitutions, therefore, disregarding aspects such as the quality of the verbal output (diction, hesitations) or the quality of the language used.

Bühler (1986) studied user expectations through a survey to professional interpreters: 16 different parameters were evaluated, including native accent, pleasant voice, fluency of delivery, logical cohesion of utterances, sense consistency with the original message, completeness of the interpretation, correct grammatical usage, use of correct terminology, use of appropriate style, thorough preparation of conference documents, endurance, poise, pleasant appearance, reliability, ability to work in a team and positive feedback from delegates.

Kopczynski (1994) approached interpretation from a pragmatic perspective. Bearing in mind that interpretation is situated communication, he contended that quality cannot be seen in a vacuum, as an absolute value, but rather as one that is relative and strongly constrained by contextual aspects: "context 'complicates' the problems of quality in that it introduces situational variables that might call for different priorities in different situations of translation"

(Kopczynski, 1994: 190). Nonetheless, despite different perceptions of quality, Kopczynski identified a number of general trends which seemed to indicate that the focus of both audience and speaker is information transfer, with terminological accuracy taking the second position, and fluency and style coming next in their list of preferences.

Shlesinger (1997: 128) proposed a three-tier method to analyse quality, which included: (i) intertextually – i.e. consistency between source text and target text, (ii) intratextually –as a product on its own right based on its acoustical, linguistic and logical features; and (iii) instrumentally– as a consumer service based on the target text's usefulness and comprehensibility.

Daniel Gile (1995, 2009) also looked at quality and concluded that the very concept of quality may have different readings depending on the role of the agents involved (be it speakers, audiences or interpreters themselves) and returning to the idea of "quality for whom and under which circumstances". However, he understands that there is an unwritten agreement on the basic indicators of quality, which, in his view, are: clarity of the ideas expressed, linguistic acceptability, terminological accuracy, fidelity and professional behaviour.

Martin and Abril (2002) formulated an assessment model with established percentages of the values allotted to the three main parameters to be assessed –contents (60%), language (20%) and presentation (20%)–, but did not provide a quantitative model with a grading system for the different types of errors.

Viezzi (2003: 151-154) approached interpretation as a transcultural interlinguistic service activity and proposed four quality goals: equivalence, accuracy, appropriateness and usability.

Alonso-Bacigalupe (2013) proposed three basic indicators of quality, very much in tune with those of Martin and Abril: content, expression and presentation. Content includes the right, accurate and complete transfer of the information contained in the ST (including the author's intention) plus all the details of the original text (thoroughness). Content is, thus, about fidelity or, to use the functionalist terminology, intertextual consistency. Expression analyses the quality of the language used, including grammar, register, acceptability, collocations and terminology. Finally, presentation describes the external components of the TT –except linguistic expression– i.e., all the features which make the interpreters' output comprehensible and fluent, credible and, above all, convincing and cogent, such as clarity of articulation or diction, appropriate speech rate and intonation, absence of disturbing noises, mumbling and hesitations, etc.

San-Bin Lee (2015) wrote an interesting paper on consecutive interpreting assessment and admitted that "there has been little work on how an instrument for IPA [interpreter performance assessment] can actually be developed in practical terms" (2015: 226) as a justification for his stated purpose of "developing an analytic scale for assessing undergraduate students' consecutive interpreting performances" (2015: 226-227). Such model was later on developed (Lee, 2019), and contained -once again- the same catalogue of quality indicators traditionally applied in practice by interpreter trainers (content, delivery and form), pointing, again, to the validity of such catalogue, but also leaving certain room for more general aspects, such as personal bias and the overall impression that the speech may potentially cause in he audience.

Also, Chao Han (2022) offered a thorough review on interpreting testing and assessment (ITA) and described in detail some of the different scoring methods that can be applied to it. However, in terms of assessment criteria he resorted, once more, to Buhler's 1986 classifica-

tion (see above) and acknowledged the following: "Over the years, researchers seem to agree on three major quality criteria: content, delivery, and language quality" (Han 2022: 38). He added that the application of such criteria may be subject to ample variations and that "information completeness" or "equivalence" is a crucial component of the criterion of content. However, neither does he offer any further indication regarding the application of the criteria proposed in practice, nor any scoring scale which may help in the assessment of interpreters' performance, or some kind of practical evidence of the applicability of such methods.

In summary, despite the mountains of literature on quality and assessment methods –as shown above– and the urgent need for interpreting instructors and students to have an assessment tool which might help them determine quality in interpreting exercises in the lab and during professional assignments, there seems to be no standard model for the evaluation of quality in SI, at least not one that has been widely embraced by the interpreter community. Pöchhacker (1994: 242) proposed a dynamic vision of the concept of quality and called for "quality under the circumstances", pointing to the idea that quality must be analysed in context, bearing in mind all the variables involved. However, I couldn't agree more when he added that "we must develop a concept of quality that can apply to all the different actors involved in the interpreting process, answering the question, what quality, for whom?" (Pöchhacker, 1994: 242).

1.3. New trends in Translation and Interpreting (TI): hybridization and computer-assisted TI

One of the most interesting developments in TI studies over the last few years comes with the introduction of information and communication technologies (ICTs) for TI. Remote interpreting (Mouzorakis, 2008), for example, is helping break the basic rule of physical co-presence of interpreters and delegates on the spot, whereas automatic speech recognition (ASR) and machine translation (MT) systems (Pöchhacker, 2019: 54-56) are helping –and sometimes even replacing– translators and interpreters in professional assignments.

But such transformations also stem from changes in the field, with the emergence of mixed audiences that may have different –and perhaps hybrid– needs, as it may be the case of the elderly and hard-of-hearing, who conform a considerable share of the population and are a crucial part of TV audiences. They often demand live subtitles –be it intralingual, i.e. with no code-switching activity involved, or interlingual, where the subtitles are the translated version of the original– for TV broadcasts, not because they do not speak the language of the country, but because of their hearing losses and limitations.

The field of audiovisual translation may be a case in point, with public and private broadcasters in Europe, the US and Australia actively working to respond to such demands, whereas international institutions are already testing systems for the provision of interlingual live subtitling. And both private and public stakeholders are demanding, in turn, objective quality assessment instruments to monitor the quality of their subtitles. As a matter of fact, the industry has already been testing different workflows and making decisions on their methods of choice, whilst researchers (Eugeni, 2020; Dawson & Romero-Fresco, 2021; Pagano, 2022; Romero-Fresco & Alonso-Bacigalupe, 2022) lead efforts to produce solid evidence of the ef-

ficiency of the systems tested, be it in terms of accuracy, time, and, of course, from a financial standpoint.

The purpose of this contribution is not, however, to delve into live subtitling research, but rather to justify why it may be useful to resort to the quality assessment methods employed in intralingual and interlingual live subtitling and apply them to SI. More than 70 year after the introduction of professional SI we continue to struggle with the evaluation of SI performance in the absence of a valid and objective tool, adding uncertainty to the already evanescent and uncertain realm of interpretation.

1.4. Accessibility

Accessibility in language services can be defined as the provision of both written titles and oral discourse for people with some type of physical or cognitive limitation. Accessibility services in the area of human communication –and more specifically in audiovisual translation– encompass a broad variety of areas –which may or may not require translation–, among which is live subtitling for the deaf and hard-of-hearing. The term coined for the provision of live subtitling services on the screen from an oral source where there is no translation activity is intralingual respeaking, whereas the term interlingual respeaking is reserved to instances where translation from one language into another language is required as well.

However, accessibility is not limited to just this. Accessibility in film-making, for example, encompasses audio description (or the oral narration of what is shown on the screen for the blind), or the production of 'special' accessible titles (surtitles, subtitles, intertitles) on the screen for the deaf, in such a manner that the titles are more readable and do not result in the loss of most of the visual information for users of the service, as it is normally the case (Romero-Fresco, 2019).

Nonetheless, for the purpose of this paper we will limit our scope to the provision of live subtitles on the screen for the deaf and hard-of-hearing, as well as for people we do not speak or understand the language of the source text, and will not dwell into those services that are not live.

1.4.1. Intra- and Interlingual Live Subtitling and Intra- and Interlingual Respeaking

Intra- and Interlingual Live Subtitling is a service through which live subtitling of a live event, normally broadcast on TV, is provided for those with special needs. It can be put in practice by applying different methods, some of which involve the use of technology (Automatic Speech Recognition and Machine Translation software principally), whereas others are exclusively human. Intra- and Interlingual Respeaking are just two of those exclusively human live subtitling methods.

Intralingual respeaking can, thus, be defined as the provision of accessible communication services for the hard-of-hearing and the deaf, but also for migrants who may not master their foreign language(s) in the oral mode but could probably make do with a written version of it. The target text (in the form of written subtitles on the screen) is in the same language as the original text, therefore, there is no translation. The respeaker receives an oral input through his/her headset and then simultaneously respeaks (or reformulates) the target text through the microphone into a speech recognition software, which automatically produces the subtitles on

the screen. The respeaker's text must be a condensed version of the ST (otherwise the audience will not be able to keep pace with the subtitles), devoid of any unnecessary elements that may be an obstacle for effective communication. The respeaker has to add punctuation marks to the TT as well (which are just pauses in oral discourse and cannot be added automatically by the software) and colour tags to identify who says what whenever more than one person are talking and visible on the screen.

Although some of those features may appear to be minor issues for the non-expert, intralingual respeaking is hard work in terms of use of cognitive resources: it requires maximum concentration, perfect articulation skills, confidence and assertiveness, extensive preparation of materials (including names and acronyms that need to be fed into the dictionaries of the ASR software), and the mental flexibility and 'the knack' (to borrow Seleskovitch's 1978 terminology) that is to be expected from conference interpreters. Another major requirement of respeaking is the need to avoid marked intonation, as it makes the TT unclear to the ASR software, and to produce a target text that sounds as neutral and 'robotic' as possible, contrary to one of the main precepts and tenets of SI. As can be observed, intralingual respeaking is an activity similar to SI, being the crucial difference between the two that in the former there is no translation, and that certain diction and pronunciation features need to be adapted to the specificities and capabilities of the software.

Intralingual respeaking has become common practice for broadcasters in Central Europe, Australia, the US, Canada and the UK. However, in certain parts of Europe the practice of intralingual respeaking is almost non-existent, and probably unknown to the general public, as it has not yet been generally embraced by private or public broadcasters as a service routinely provided to viewers.

Interlingual respeaking, in turn, involves a further (and crucial) development from intralingual respeaking, since in the interlingual mode the respeaker translates simultaneously the ST in one language into a TT in a different language, while all other features of the task remain unchanged. Interlingual respeaking is, therefore, a form of simultaneous interpretation, save for the fact that the TT voiced through the mike by the interlingual respeaker will not be available to the audience through their headsets, as in SI, but rather, the TT is directly dictated to the ASR software, which will then produce the written subtitles on the screen automatically. The other main defining features of respeaking that need to be maintained include: maximum condensation of utterances (to avoid cognitive overload of the audience), thorough preparation of materials, the use of marked intonation is strongly discouraged, and punctuation marks and colour tags need to be added to the target text as well. Therefore, since additional tasks and skills are required for the job, interlingual respeaking may be seen as more demanding than traditional SI, or, to put it differently, it may be contended that the skills required for successful interlingual respeaking build on SI and may be a step beyond those required for successful SI.

1.4.2. Speech-to-Text Interpreting

Inter- and Intralingual Live Subtitling is not limited, however, to intra- or interlingual respeaking (where only humans do the job), as it can be provided through a wide variety of methods (or workflows), some of which also involve the participation of humans (doing SI, for example), whereas others involve the participation of machines exclusively. Among the latter

is the combination of Automatic Speech Recognition (ASR) and Machine Translation (MT). The former, however, include different potential systems and combinations, among others: (i) a single human agent, or (ii) the combination of two (or more) human agents, or else (iii) a human (or more) and a machine working in concert (Romero-Fresco & Alonso-Bacigalupe, 2022). Speech-to-Text Interpreting (Stinson, 2015) is, thus, an umbrella term proposed to name all the methods and systems that can be used for the production of live subtitles on the screen. Interlingual respeaking is just one form of speech-to-text-interpreting (STTI).

1.5. Measuring quality in intralingual and interlingual live subtitling

As said above, more than 70 years after the earliest efforts to define and measure quality in interpreting there seems to be no standard system, instrument, tool or model widely embraced by the profession to measure quality in SI. However, the landscape is different for accessibility services, perhaps because users of the service conform a massive audience (counted by the millions), perhaps because public broadcasters can be held accountable for the quality of the service they provide, or perhaps because private broadcasters struggle to meet the expectations of their fee-paying customers. Be it as it may, broadcasters in the USA, Canada, the US and Australia, Belgium and other places are demanding protocols for the evaluation of the quality of their subtitles (including live subtitles). And due to that demand, systems have been developed for the objective analysis of the quality of intra- and interlingual live subtitling.

Also, most recently the European Parliament has been testing the quality of their live subtitles in an initiative to provide this service for the deaf and hard-of-hearing in parliamentary meetings through a system that will run in parallel with the Parliament's translation and interpretation services.

1.5.1. Assessment models for intra- and interlingual live subtitling: WER and NER

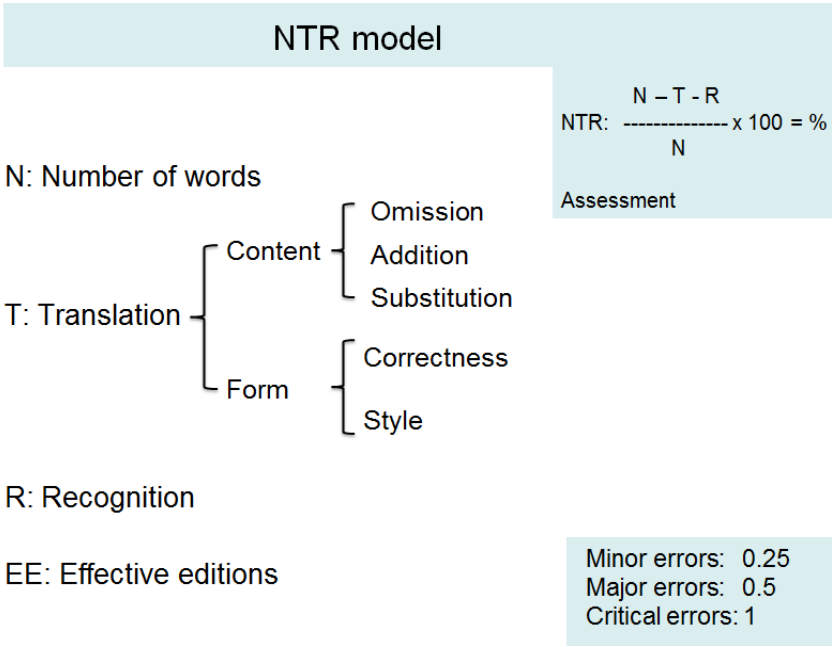
Most models of assessment for intralingual live subtitling are based on the principles of the WER (word error rate) methodology, which have been applied to the analysis of accuracy in speech recognition (Dumouchel, Boulianne, & Brousseau, 2011). This error-based model was developed to identify three types of information errors: deletions (D), substitutions (S), and insertions (I). The model, however, focuses exclusively on information transfer and is therefore not sensitive to language or presentation issues. Also, this model focuses on words, and, thus, penalizes differences between source and target text, regardless of whether or not the spirit of the message has been conveyed.

The NER model was introduced in Romero-Fresco (2011) and developed further in Romero-Fresco and Martínez (2015) and is based on the principles of the WER model. This model "grades errors of edition (E) and recognition (R) to different degrees of severity while also accounting for editing that can be considered an appropriate, or 'correct', choice" (Romero-Fresco & Pöchhacker, 2017: 151-152). This model is sensitive to the number (N) of words in the subtitles and is meaning-focused, and therefore the unit used for scoring errors is the idea unit, understood as a "unit of intonational and semantic closure" (Chafe 1985: 106).

1.5.2. The NTR Model

The NTR model (Romero-Fresco & Pöchhacker, 2017) is a further development from the NER model and is used for the assessment of interlingual respeaking and interlingual speech-to-text-interpreting. The NTR (Number of words, Translation, Recognition) Model uses a formula and an error-grading system that resembles those of previous models. However, in NTR the E for the errors of edition is replaced by a T, which accounts for translation errors. Therefore, a difference is established between instances where the speech-to-text interpreter provides a wrong translation (T), and those instances where the software cannot recognize what s/he says (R). Also, T errors are subdivided into content and form. The former includes omissions, additions and substitutions, and the latter grammar, terminology, register and appropriateness. All errors are classified by their degree of severity. Also, in this model N is the number of words in the subtitles, which means that it is sensitive to the length of the ST.

Figure 1. The NTR model (Romero-Fresco & Pöchhacker, 2017: 159)



1.6. In search of common ground: the extrapolation of assessment models

Taking into account (i) the previous review of the many points in common between SI and interlingual STTI, to such an extent that it may well be claimed that interlingual STTI is just one development from SI, (ii) that STTI already enjoys the benefits of a standard model for the evaluation of subtitles on the screen, and (iii) that no such instrument has yet been developed for SI, one might well ask oneself about the potential benefits of the extrapolation of the NTR to SI.

In theory, there is no apparent reason why it could not be used for it, since what is subject to evaluation in the NTR is exactly the same catalogue of quality indicators used for the evaluation of SI: (1) the amount and accuracy of the information conveyed, (2) the quality of the language used, and (3) the comprehensibility of the message, which draws heavily on the clarity (or the absence of it) of the formulation of the target text. But it is precisely in this final aspect where major differences emerge.

In the case of SI, a recurrent quality indicator that has been proposed by scholars and instructors (beyond content and language) is presentation, which includes good diction (or clarity of articulation) and appropriate intonation, as well as fluency, a pleasant voice, absence of noises, hesitations or false starts and other elements that may make the TT unclear, confusing or hard to understand, and that are different from language as such.

With the NTR model what is first analysed is translation errors, which are then subdivided into content errors, and form errors. And the third indicator is recognition errors. The difference, therefore, is to be found exclusively in that third indicator at stake, and has to do with the fact that the speech-to-text interpreter does not speak directly to an audience –as in SI– but to an speech recognition software, in such a way that the final product in SI is an oral text, whereas it is a written text in STTI. However, if we look carefully into the issue, it may be fair to contend that instances identified as recognition errors with the NTR in STTI contexts can be rightly compared to the presentation errors identified in SI, since in both cases they deal with obstacles in communication stemming not from wrong translations or language problems, but from the lack of clarity of the message reaching the TT audience. Therefore, whereas in SI the message may be difficult to understand when, for example, the interpreter's diction is unclear or her/his speech rate is excessively fast, or when her/his message is full of hesitations and false starts, in STTI those instances of lack of clarity occur when the subtitles contain errors stemming from the limited recognition capacity of the ASR engines to understand messages well when they have not been clearly dictated to the machine.

2. The research

Once it has been conceptually established why it may be feasible in theory to use the NTR for the evaluation of simultaneous interpretation tasks, the next step would be to design and implement an empirical analysis which may help demonstrate its usability also in practice. This analysis includes two parts: one pilot test and one case study.

2.1. The pilot test

2.1.1. Materials and methods

Alba Fernández (2022) took one original real-life source text in English (17 min.), delivered on 28 October 2021 by Vivian Schiller, former president and CEO of National Public Radio (US) and used it as the source speech for her test. Also, since the speech had been simultaneously interpreted by a professional conference interpreter into Spanish, this rendition was used as the target text of this pilot.

Fernández downloaded ST and TT and then graded the quality of the TT using three different evaluation instruments (2 developed for SI and the NTR for STTI) in search of correlations of the scores after application of the three instruments to the same TT. What follows is a description of the three assessment methods used in Fernández's comparative analysis.

The method of Alonso-Bacigalupe (LAB, 2013) is a detailed and complete attempt to assess SI objectively, and includes a classification of errors –graded in terms of severity– and the negative points allotted to each type of error that can be applied to each one of the dimensions analysed, (1) information, (2) language and (3) presentation, weighed 33% each for the

final score. Errors range from critical errors (for example, in the case of information, those which result in serious deterioration of the information conveyed, including misleading or false information, 10 points), to major errors (those with a strong impact, including important deviations in information or major losses of information, 5 points) and minor errors such as small omissions of details (3 points). In terms of language, incomprehensible or nonsensical utterances receive the maximum penalization of 10 points to be taken from the initial credit of 50 points in each of the dimensions assessed, while smaller language problems, such as wrong collocations and wrong literal translations, are penalized with 5 points, and problems of register and terminology with 3 points. In the dimension of presentation maximum penalties are applied wherever the flow of information is interrupted abruptly leaving unended ideas or sentences, or wherever diction is so unclear as to make the TT incomprehensible (minus 10 points), whereas most other problems, such as false starts, mumbling or problems of concordances are penalized with three points. In-between the two there is a five-point penalty that is applied when the speech rate is too fast or too slow, or when hesitations and mumbling are repetitive and particularly disturbing. This model has proved to be useful in providing an objective assessment of students' performance in the SI class (Alonso-Bacigalupe 2013). However, the model is limited in that it is not sensitive to the number of words in the source text, which means that adjustments need to be made whenever a ST longer than 800 to 1,000 words (length for which the model was originally developed) has to be evaluated. On the other hand, this model had been applied only intuitively and its usefulness has never been confirmed with an empirical analysis.

Martin and Abril's model (M&A, 2002) also analyses the same parameters, but they are weighed differently, and priority is given to information (60%), whereas language is 20% and presentation the remaining 20%. This model, however, is limited as well in that it does not grade errors according to their severity. It is, therefore, subject to a high level of subjectivity and variability, and the final results depend largely on individual preferences and/or bias of the evaluator.

The NTR, as described above, could serve the purpose of closing the gap in this area, as it is sensitive to the number of words of the ST and includes a typology and a taxonomy of the severity of errors comparable to other methods in terms of detail and depth, therefore providing a thorough analysis of successful (or unsuccessful) performance.

As one of those methods (M&A) did not include numerical scores, those of LAB were applied to the three dimensions in their model. The difference in the analysis carried out with these two instruments therefore lays in the fact that the scores achieved for each dimension would be weighed 33% each for the final score in LAB, whereas they were 60% information, 20% language and 20% presentation in M&A. The third instrument would be the NTR, taking into consideration that the recognition errors would be understood and considered equivalent to the presentation errors of the other two models, as justified above.

2.1.2. Results and discussion

Fernández found that all three instruments rendered similar results in the analysis of the quality of this SI task: the three provided a grade between 7 and 8 points in a 10-point scale. Particularly remarkable was the level of consistency found between M&A and NTR, as they

provided exactly the same score: 8 points in the 10-point scale. As shown in the table below, completeness of information (CONT), language quality (EXP), and quality of delivery (PRES) were the three quality indicators at stake.

Table 1. Results of the pilot test

Model	CONT	EXP	PRES	POINTS	SCORE
M&A (+LAB)	76.2 (60%)	22 (20%)	22.8 (20%)	121/150	8
LAB	127 (33%)	110 (33%)	114 (33%)	117/150	7
NTR	9.25 (T = CONT+EXP)		4	99.5	8

Although Fernández's research was just a pilot test with a small sample of just one ST interpreted by a professional interpreter in a real-life scenario, results were all the most interesting and encouraging, as it turned out that only minor differences were found in the final scores after application of these three different evaluation methods.

This could probably be seen as preliminary evidence of consistency, and, thus, of a potential correlation in terms of efficiency and relevance of all three instruments, and, above all, of the applicability of the NTR to SI as well.

2.2. The case study

In an attempt to confirm the validity of the NTR in its application to SI a further step was then taken. In Fernández's pilot test above the performance of a professional interpreter had been analysed using the above-mentioned three models in search of a correlation of results. This time the idea was to try to confirm if the scores obtained by a small group of students were consistent or inconsistent after application of the three models, thus validating –or else refuting– the correlation and, therefore, the usability of the NTR for SI.

2.2.1. Materials and subjects

The materials used for this analysis included:

- One original real-life source text in English downloaded from the internet: Prime Minister Rishi Sunak's statement at the COP27 2022 summit in Egypt, delivered on 7 November 2022. The number of words was 682 and the duration of the video recording 6:15 min. Therefore, the speech rate was around 115 words per minute, which is deemed as a comfortable speech rate by interpreters.
- Four target texts in Spanish, which were the recordings of the simultaneous interpretations of the speech described above that the students had done a few weeks earlier for their final test in SI.

The subjects selected for this analysis were the four fourth-year students of the degree in Translation and Interpreting at the Universidade de Vigo (Spain) with the best grades (graded with M&A) in their final tests at the end of their training programme in advanced SI in the language pair English-Spanish. The rationale for choosing this small sample from a much larger potential sample of 12 to 14 relevant students in the class was that it is relatively easy to intuitively differentiate between good and poor performance in SI. What is not so easy is to effectively identify minor differences in performance that may help us grade students and

professionals accordingly, which is precisely the purpose of this research: the application of an assessment model that can actually distinguish and spot such small differences, therefore discriminating subtle differences of quality in SI.

2.2.2. Methodology

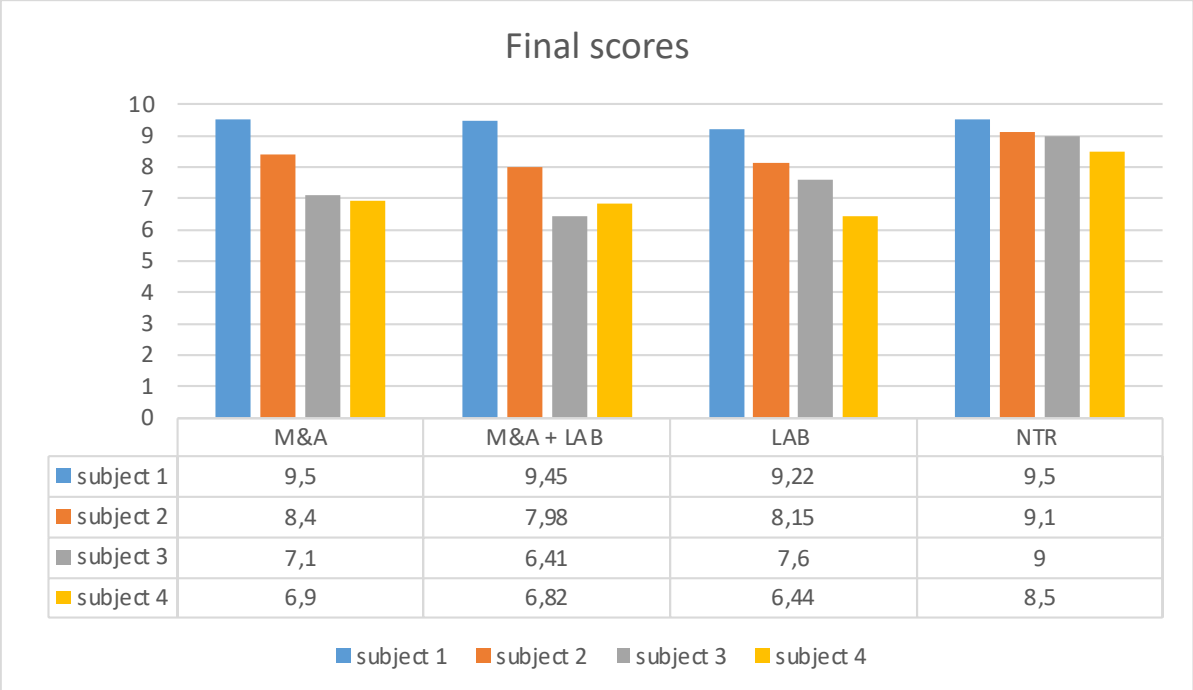
In this case study, however, there was a slight variation with respect to Fernández's research in the number and choice of the methods (or models) of analysis, since the subjects' performance had already been assessed with M&A for their final exam. Therefore:

- The reference model, or Model 0, was M&A. This was the model that had been used to grade these students for their final exam, as applied according to Martin and Abril (2002), i.e. without any kind of adjustment or modification in the form of numerical points or scales.
- Model 1 (M&A + LAB) involves the application of M&A's weighing factor (60% CONT, 20% EXP, 20% PRES) combined with LAB's scores, as in Fernández's.
- Model 2 is LAB's, as explained above (section 2.1.1).
- Model 3 is the NTR (section 1.5.2).

2.2.3. Results

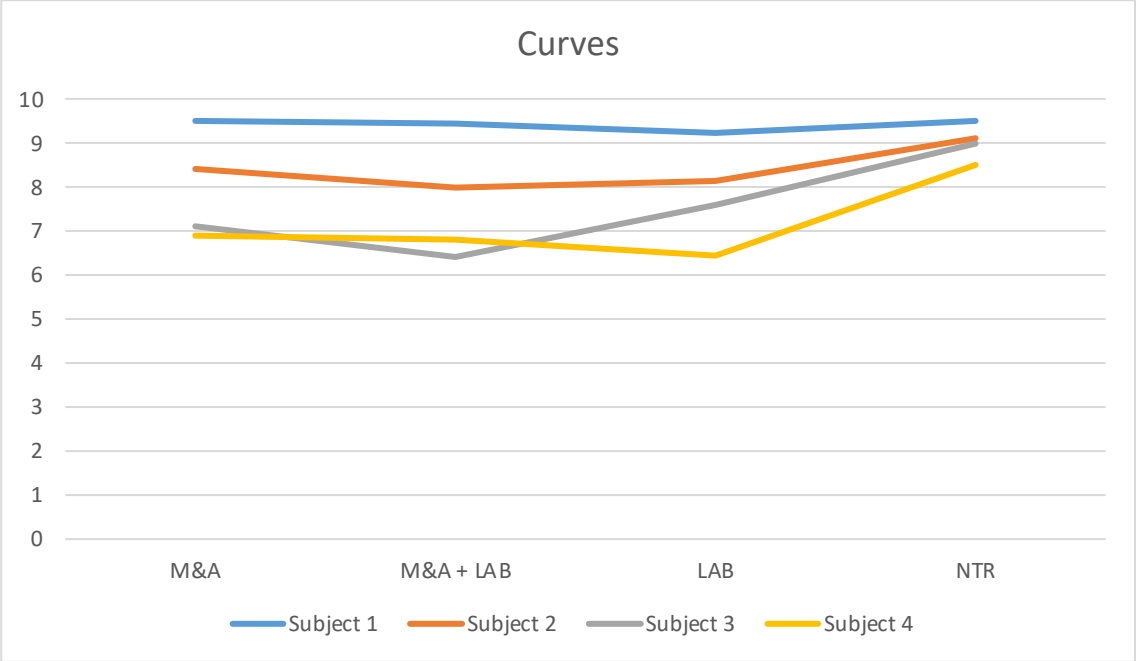
The table and chart below show the scores of the four subjects after application of all four methods of analysis (Figure 2). The chart in Figure 3 is the curve stemming from the graphical representation of the scores.

Figure 2. Final scores of all four subjects after evaluation with all four methods



	M&A	M&A + LAB	LAB	NTR
subject 1	9,5	9,45	9,22	9,5
subject 2	8,4	7,98	8,15	9,1
subject 3	7,1	6,41	7,6	9
subject 4	6,9	6,82	6,44	8,5

Figure 3. Curve consistency of grades



As can be observed in the charts above, the curve-shapes stemming from the analysis of the scores of these 4 students after application of all four instruments to measure their performance are consistent all through the analysis regardless of the assessment model used. The student who obtained the best marks in the final exam (subject 1) would also get the highest scores with the other three evaluation tools tested here, whereas the subject with the lowest grades would rank last using all four models. And this is consistent for all the other positions in this ranking: although there are minor differences in the final scores, the fact is that no alterations or discrepancies have been found in terms of the positions attained. This appears to suggest that all four models subject to testing in this research may be valid and might be applicable for the evaluation of an SI task, including the proposal of an adjusted version of NTR, originally developed for the assessment of Speech-to-Text Interpreting or interlingual respeaking.

The only minor exception in terms of curve-shape is found in method M&A + LAB which was just a potential ad-hoc scenario, tentatively proposed exclusively for the purpose of this research, but never applied in practice in the SI class. However, it is worth noting that in Fernandez's pilot test there was a high level of consistency between NTR and this method as well.

Also, it is most salient that the three assessment models that are being used in practice (M&A and LAB for the evaluation of SI in the class, and the NTR applied to STTI in professional life) provide exactly the same curves, which can be understood as a high level of consistency, thus suggesting the validity of all three instruments.

3. Discussion

The main objective of this research –to test to what extent the NTR model might be applied to SI tasks without causing major disturbances or discrepancies– seems to be confirmed, as proved by the fact that in the case study the NTR curve is consistent with the curves stemming after application of the methods proposed and applied in practice by SI instructors (M&A and

LAB), which points to the feasibility of applying the NTR to SI as well. But there are other interesting effects derived from the use of this model. For example, the NTR can be considered to be quite generous and fair with students with lower levels of performance, whilst not punishing those with the best performance. In fact, these four subjects would have been ranked in exactly the same positions had their exam been graded with NTR in the first place, but their performance would have been deemed slightly better for the students with the lowest scores, while remaining nearly unaltered for those with the highest final scores.

Also, the NTR shows a high level of sensitivity to minor errors and is able to distinguish very small differences in performance, which supports its usability and efficiency for a fine-tuned assessment of interpreter performance. However, this high level of sensitivity, derived from both careful attention to detail and source-text segmentation into short units of meaning –which is one of the main strengths of this assessment model–, quickly turns into its main weakness, as applying the NTR is time-consuming and requires quite a lot of concentration and effort. Nonetheless, that weakness is equally shared by all the other instruments, a hurdle which can only be overcome if the systems can be made automatic in the future.

4. Conclusions

Measuring quality is always a complicated thing to do. But it is not impossible. Quality is dealt with as a priority these days and is being measured in all walks of life. Therefore, I can see no reason why it is not possible to measure quality in SI.

It is of course true that quality-measurement mechanisms are not always 100% accurate or efficient, and that flaws may appear in such evaluations. However, this is not a sufficiently powerful reason to contend that quality cannot be subject to objective analysis in SI: if the product of an Automatic Speech Recognition or Machine Translation system in interlingual live subtitling, or that stemming from human practice in interlingual respeaking or speech-to-text-interpreting (STTI) can be measured, and for whatever realm, subject, or topic, it can reasonably be contended that this is equally feasible for SI, since, as argued above, STTI builds on the principles and tenets, skills and limitations, and constraints and possibilities of SI.

The interpreter profession already has a long tradition and, thus, its own mechanisms to facilitate or deny access to the market, that is to say, to select whether or not an individual interpreter is up to the level expected from other professional interpreters and, above all, from an international audience. But that does not mean that professional interpreter performance cannot be (or should not be) objectively measured, particularly in today's world, where measuring the quality of any service is standard practice for businesses. But beyond this, we need objective methods for the assessment of students as well, not only because they need a mark to grade their performance for their academic record, but also because they need to have a list of the parameters and indicators of high and low performance, including the levels of severity of the errors and the grading system used to penalize such occurrences. Giving imprecise indications and subjective evaluations in the SI class does not probably help them as much as providing them with a detailed catalogue of errors, a taxonomy of the severity of such errors –including considerations on the communicative impact they might have– and the corresponding penalizations applied. This would inform them where they need to improve their performance, also providing them with threshold levels of acceptability that can help them set their

objectives and priorities, emphasizing the need of mastering the skills they do not master yet. And, in this regard, the NTR Model might perhaps have a contribution to make: as tentatively demonstrated in this research, it may be reasonably applied for the objective assessment of SI.

Moreover, in a world in a constant process of change, ICTs have come to transform traditional translation and interpretation practice. Interpreting is not any more just about situated communication, as interpreters, audience and clients may now be separated by long distances. And it is still interpreting. ICTs provide a wealth of new opportunities which should not be neglected or discarded just because they may shake some of the foundations of professional practice. In this paper STTI and interlingual respeaking are seen as a form of interpretation in its own right, as they maintain the main features of traditional interpretation practice, while revisiting the very way the service is provided to the audience with the support of ICTs, including the use of automatic processing systems, such as MT and ASR. Some interpreters are already enjoying the benefits of these breakthroughs, whereas others may be reluctant to follow suit.

Be it as it may, STTI is approached here as a thriving area for future growth and as a potential new source of employment opportunities for those interpreters who are prepared to take up the challenge of fine-tuning their already extensive professional skills. By embracing these emerging opportunities not only will interpreters benefit from the many advantages of developments in ICTs, but also they will help provide a new kind of linguistic services to a growing number of vulnerable users that are increasingly demanding them around the world.

References

Alonso-Bacigalupe, L. (2013). Interpretation quality: from cognitive constraints to market limitations. In R. Barranco-Droege, E.M. Pradas & O. García (Eds.), *Quality in Interpreting: Widening the Scope. Vol II* (pp. 9-33). Comares.

Barik, H. C. (1971). A description of various types of omissions, additions and errors of translation encountered in simultaneous interpretation. *Meta, 16*(4), 199-210.

Barik, H. C. (1972). Interpreters talk a lot, among other things. *Babel, 18*(1), 3-10. https://doi.org/10.1075/babel.18.1.01bar

Barik, H. C. (1973). Simultaneous interpretation: temporal and quantitative data. *Language and Speech,* 16, 237-270.

Bühler, H. (1986). Linguistic (semantic) and extra-linguistic (pragmatic) criteria for the evaluation of conference interpretation and interpreters. *Multilingua*, *5*(4), 231-235. https://doi.org/10.1515/mult.1986.5.4.231

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 University Press.

Collados, A. (1997). *La entonación monótona como parámetro de calidad en interpretación simultánea: la evaluación de los receptores.* Doctoral Dissertation. Universidad de Granada.

Collados, A. & Sabio, J.A. (Eds.) (2003). *Avances en la investigación sobre interpretación.* Comares.

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). Peter Lang.

Eugeni, C. (2020). Interaction in Diamesic Translation. Multilingual Live Subtitling. In D. Dejica, C. Eugeni & A. Dejica-Carţiş (Eds.), *Translation Studies and Information Technology - New Pathways for Researchers, Teachers and Professionals* (pp. 19-31). Editura Politehnica.

Fernández, A. (2022). *La evaluación de calidad en interpretación simultánea: análisis comparativo de la calidad del discurso a través de tres modelos de evaluación*. Trabajo de Fin de Grado. Facultade de Filoloxía e Tradución. Universidade de Vigo.

García, O., Pradas, E.M. & Barranco-Drodge, R. (Eds.) (2013). *Quality in Interpreting: Widening the Scope. Vol 1*. Comares.

Gile, D. (1995, 2009). *Basic Concepts and Models for Interpreter and Translator Training.* John Benjamins.

Han, C. (2022). Interpreting testing and assessment: A state-of-the-art-review. *Language Testing, 39*(1), 30-55. https://doi.org/10.1177/02655322211036100

Herbert, J. (1952). *The Interpreter's Handbook: How to Become a Conference Interpreter*. Librairie de L'Université Georg.

Dawson, H. & Romero-Fresco, P. (2021). Towards research-informed training in interlingual respeaking: an empirical approach. *The Interpreter and Translator Trainer, 15*(1), 66-84. https://doi.org/10.1080/1750399X.2021.1880261

Kahane, E. (2022). Thoughts on the quality of interpretation. AIIC.net. May 13, 2000. Accessed October, 27, 2022.

Kopczynski, A. (1994). Quality in Conference Interpreting: some pragmatic problems. In M. Snell-Hornby, F. Pöchhacker & K. Kaindl (Eds.), *Translation Studies --An Interdiscipline* (pp. 189-198). John Benjamins.

Kurz, I. & Pöchhacker, F. (1995). Quality in TV interpreting. In Y. Gambier (Ed.), *Audiovisual Communication and Language Transfer. Proceedings of the International Forum Strasbourg. Translation. FIT Newsletter* série XIV/3-4, pp. 350-358.

Lee, S. (2015). Developing an analytic scale for assessing undergraduate students' consecutive interpreting performances. *Interpreting, 17*(2), 226-254.

Lee, S. (2019). Holistic assessment of consecutive interpretation. How interpreter trainers rate student performances. *Interpreting, 21*(2), 245-269. https://doi.org/10.1075/intp.00029.lee

Martin, A. & Abril, M.I. (2002). Didáctica de la interpretación: algunas consideraciones sobre la evaluación. *Puentes, 1*, 81-94.

Mizuno, A. (1997). Broadcast interpreting in Japan. Some theoretical and practical aspects. In Snelling (roundtable discussion). In Y. Gambier, D. Gile & C. Taylor (Eds.), *Conference Interpreting: Current Trends in Research* (pp. 192-194). John Benjamins.

Moser, P. (1995). *Survey on Expectations of Users of Conference Interpretation: Final Report, January 1995.* International Association of Conference Interpreters (AIIC).

Mouzourakis, P. (2008). Remote Interpreter Training – Training for Remote Interpreting? ISCAP, Porto. https://multimedialinguas.wordpress.com/edicoes/ano-i-2010/0001-janeiro/panayotis-mouzourakis-%C2%ABremote-interpreter-training-training-for-remote-interpreting%C2%BB

Padilla, P. & Martin, A. (1992). Similarities and differences between interpreting and translation: implications for teaching. In C. Dollerup & A. Loddegaard (Eds.), *Teaching Translation and Interpreting: Training, Talent and Experience* (pp. 195-203). John Benjamins.

Pagano, A. (2022). *Testing Quality in Interlingual Respeaking and other Methods of Interlingual Live Subtitling*. Doctoral Thesis. Ph.D. in Digital Humanities Languages, Cultures and Digital Technologies. Department of Modern Languages and Cultures. Universitá Di Genova.

Pöchhacker, F. (1994). Quality assurance in simultaneous interpreting. In C. Dollerup & A. Lindegaard (Eds.), *Teaching Translation and Interpreting 2.* John Benjamins.

Pöchhacker, F. (2013). Researching quality: A two-pronged approach. In García, O., E. M. Pradas & R. Barranco-Drodge (Eds.), *Quality in Interpreting: Widening the Scope. Vol 1* (pp. 33-56). Comares.

Pöchhacker, F. (2019). Moving boundaries in interpreting. In H. Van Dam, M. N. Brøgger & K. Korning Zethsen (Eds.), *Moving Boundaries in Translation Studies* (pp. 45-63). Routledge.

Romero-Fresco, P. (2011). *Subtitling through Speech Recognition: Respeaking*. Routledge.

Romero-Fresco, P. (2019). *Accessible Filmmaking*. Routledge.

Romero-Fresco, P. & Martínez, J. (2015). Accuracy rate in live subtitling: The NER model. In Díaz Cintas, J. & R. Baños (Eds.) *Audiovisual Translation in a Global Context: Mapping an Ever-changing Landscape* (pp. 28-50). Palgrave Macmillan.

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.

Romero-Fresco, P. & Alonso-Bacigalupe, L. (2022). An empirical analysis on the efficiency of five interlingual live subtitling workflows. *XLinguae,* Volume *15*(2) April 2022, 3-16.

Russo, M. (1995). Media interpreting: variables and strategies. *Translatio. Nouvelles de la FIT- FIT Newsletter,* XIV (3/4), 343-349.

Seleskovitch, D. (1978). *Interpreting for International Conferences: Problems of language and communication*. Pen and Booth.

Shlesinger. M. et al. (1997). Quality in simultaneous interpreting. In Y. Gambier, D. Gile, D. & C. Taylor (Eds.), *Conference Interpreting: Current Trends in Research* (pp. 123-131). John Benjamins.

Shlesinger, M. (1995). Shifts in cohesion in simultaneous interpreting. *The Translator, 1*(2), 193-214. https://doi.org/10.1080/13556509.1995.10798957

Shlesinger, M. (1999). Norms, strategies and constraints. How do we tell them apart. In A. Álvarez & A. Fernández (Eds.). *Anovar/anosar: Estudos de Tradución e Interpretación, 1*, 65-77.

Stinson, M. S. (2015). Speech-to-text interpreting. In F. Pöchhacker (Ed.), *Routledge Encyclopedia of Interpreting Studies* (pp. 399-40). Routledge.

Viezzi, M. (2003). Interpretation quality: A model. In A. Collados & J.A. Sabio (Eds.), *Avances en la investigación sobre interpretación* (pp. 147-157). Comares.