# An empirical analysis on the efficiency of five interlingual live subtitling workflows

Pablo Romero-Fresco – Luis Alonso-Bacigalupe

DOI: 10.18355/XL.2022.15.02.01

**Abstract**
Interlingual Live Subtitling (ILS) is an innovative translation and accessibility method where a written text in one language is produced live from an oral source in another language. ILS can be provided through different methods, some of which involve the participation of one or more humans, whereas others are fully automatic.
Speech-to-text interpreting (STTI) is a form of human-mediated ILS that is situated at the crossroads of audiovisual translation, media accessibility and simultaneous interpreting, as well as between human-mediated translation and automatic language processing systems. One of the most promising forms of STTI is interlingual respeaking. It builds upon intralingual respeaking (the most common form of speech-to-text captioning, which does not include language transfer) and involves the participation of a human interpreter plus speech recognition software.
Although interlingual respeaking is in great demand, there are other approaches to STTI -with different degrees of human intervention- which are currently being used by broadcasters and conference organizers. The purpose of this research is to test the efficiency of five of those methods, namely, (1) interlingual respeaking, (2) simultaneous interpreting plus intralingual respeaking, (3) simultaneous interpreting plus automatic speech recognition, (4) intralingual respeaking plus machine translation and (5) automatic speech recognition plus machine translation.
The results provide a useful insight into the current efficiency of five different ILS methods and strengthen the idea that efficiency is not restricted to accuracy, but includes factors such as delay and the type of resources (either human or machine) required. It is hoped that this research may help provide the industry with tools to make informed choices between different forms of ILS (at least for the language combination English-Spanish) while offering employment opportunities for simultaneous interpreters and respeakers in the digital era.
**Key words:** Interlingual live subtitling (ILS), speech to text interpreting (STTI), intra and interlingual respeaking, accessibility, automatic speech recognition (ASR), machine translation (MT)

## 1. Introduction

As we move into the second decade of the 21st century, the consolidation of globalization and the increasing diversity of contemporary societies have brought up new communicative needs across cultures and languages. The COVID-19 pandemic has only made these needs more pressing, drawing on the development of ever-advancing technological innovations to enable remote communication where on-site and face-to-face communication was not possible. As a discipline that tracks, analyzes and fosters cross-cultural communication, Translation and Interpreting is in a (perhaps inevitable) state of flux. Although prototypical conceptions of translation and interpreting are still upheld (for instance in most university degrees and standardization committees), the boundaries between the two areas are increasingly blurring. Interpreting is no longer only concerned with the spoken and signed modalities, as it now includes written target texts (Pöchhacker, 2019), which links it to audiovisual translation. Traditionally concerned with dubbing and subtitling for films, audiovisual translation now covers the provision of intra- and interlingual

translation for live programmes and live events for users who have no access to the original version (be it for linguistic, sensory-related or other reasons), thus overlapping with media accessibility.

This paper looks at Interlingual Live Subtitling (ILS) as a communication-enabling service in response to specific social requirements –the need to produce a written translation for a live event or programme conducted in a different language so as to provide access for people with and without hearing loss. It is thus a technology-enabled hybrid modality of translation (Davitti & Sandrelli, 2020) at the crossroads of interpreting, audiovisual translation and media accessibility (see Figure 1).

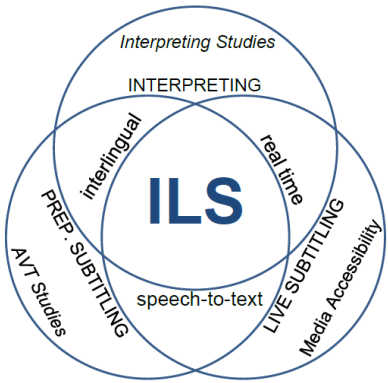


**Figure 1: The geographical range of ILS within translation studies**

ILS is firstly an example of diamesic translation, which consists of "a change in language mode; i.e. from speech to writing or vice versa" (Gottlieb, 2017:51). Drawing on Gottlieb, Eugeni (2020: 29) defines the diamesic translation involved in ILS as "any process, or product hereof, in which a combination of spoken and non-verbal signs, carrying communicative intention, is replaced by a written combination reflecting, or inspired by, the original entity". Eugeni (ibid.) adds that this type of diamesic translation may be concerned with reporting every sound of the original (phonetic transcription, automatic transcription, phone tapping), every word (court reporting, captioning for TV commercials) or its meaning (live subtitling, parliamentary reporting, film subtitles for the deaf and hard of hearing).

Although the notion of diamesic translation goes some way towards describing ILS, it does not account for its interlingual nature. This explains why ILS is being referred to as a form of "real-time interlingual speech-to-text transfer" (Davitti & Sandrelli, 2020: 104) or "real-time speech-to-text translation", defined by Pöchhacker and Remael (2019: 131) as "the real-time rendering of a spoken source-language utterance into a written target-language text". As it corresponds to a service that sits at the crossroads between interpreting, audiovisual translation and media accessibility, ILS targets viewers who do not have sufficient knowledge of the source language, viewers with hearing loss and with specific communicative needs, and those who, for contextual reasons, may not have access to the original audio. ILS is currently being used in a wide range of settings (e.g. TV, conferences, workplace, political, educational,…), event types (e.g. breaking news, business meetings, parliamentary debates, classroom interaction, museum tours,…) and formats (e.g. one speaker, multi-party interaction, etc.) (Davitti & Sandrelli, 2020). Depending on the setting, the output may consist of subtitles on a screen (TV programmes) or running text displayed on separate screens, mobile devices or even glasses in a live event, which explains why authors such as Pöchhacker and Remael (2019: 133) refer to live titling (not necessarily displayed under images) rather than live subtitling.

ILS can be delivered on-site or online and, crucially for the purpose of this paper, it can be produced with a range of methods or workflows, all of which involve different degrees of human-computer interaction (Hewett et al., 1992). The most commonly used form of ILS is speech-to-text interpreting (STTI). Initially conceived as an intralingual communication service for deaf and hard-of-hearing people (Stinson, 2015), STTI is now also used interlingually. In interlingual STTI, an interpreter listens to the original soundtrack of a programme or event in the source language and then produces a written translation using a keyboard, a stenotyping machine or speech recognition software. The latter option requires interlingual respeaking or transpeaking (Pöchhacker and Remael, 2019), that is, an interpreter (interlingual respeaker) who translates what is being said, adding punctuation marks, to the speech recognition software, which displays the output as text on screen. Interlingual STTI can also be done by combining different professionals: an interpreter who translates the source language audio into target language audio and a second professional who types or respeaks the target language audio into target language written text. And then, more generally, ILS can also be produced with methods that do not involve (human) STTI, for instance by combining automatic speech recognition (ASR) to turn source language audio into source language text with machine translation (MT) to turn source language text into target language text.

The aim of this paper is to compare the efficiency (in terms of accuracy, delay and human/machine resources required) of five ILS methods involving different degrees of human-computer interaction: from several forms of STTI to a completely automated workflow. Before describing the experiment and discussing its main results, the next section provides an account of previous studies that have tested some of these methods and that can be seen as precedents to the pioneering analysis presented here.

## 2. Prior research

Research comparing the efficiency of different ILS workflows is virtually non-existent. Some authors have, however, provided descriptions and accounts of ground-breaking ILS experiences in their countries. This is the case of Kurz and Katschinka (1988), who report on an experiment in ILS on Austrian television in the late 1980s for an arts programme featuring two English-speaking participants. Their utterances were subtitled live into German by a team made up of simultaneous interpreters (who translated the English audio into German audio) and a subtitler (who turned the German audio into written German subtitles). The experience involved a significant lag or delay in the subtitles, which put off the broadcaster from delivering this service regularly. Den Boer (2001) and de Korte (2006) describe a similar experience in the Netherlands, where the Dutch public broadcaster Nederlandse Omroep Stichting first trialled ILS in the late 1990s. Despite an unsuccessful first experience producing live Dutch subtitles for a live interview featuring president Clinton in 1999, the broadcaster persevered and managed to have a broadcast delay of 20 to 30 seconds in subsequent programmes (ibid). This allowed a much-needed breather for the two subtitling teams involved in the broadcast, which typically included a subtitler interpreting the English audio into Dutch audio and another subtitler typing the Dutch audio on a Velotype (special type of keyboard) to produce the subtitles as accurately and fast as possible.

One of the few existing analyses of the quality of different live subtitling methods was conducted by Daniela Eichmeyer in 2021. Eichmeyer compared the quality (in terms of accuracy and user preference) of the intralingual (German-German) subtitles provided by respeakers and typists for three different lectures held at the Klinikum Großhadern in Munich. Although there is no interlingual component here, it may be considered as a relevant example of analysis of intralingual STTI for the purpose of

this paper, especially as one of the methods analyzed is respeaking. An audience made up of 46 participants received on their iPads the subtitles produced by four teams of subtitlers: two of them used a conventional keyboard and the two other used respeaking. All subtitlers had an average professional experience of 2-5 years. The results show that the subtitles produced by respeaking were 12% more accurate (as per the WIRA model used by Eichmeyer) than those produced by a conventional keyboard. As for user preference, the results show a preference for respeaking and for the target text to be displayed as subtitles under the image of the speaker rather than as running text on a separate screen.

The most relevant research study carried out so far on the quality of different ILS methods is the one conducted by Carlo Eugeni (2020) at the 2019 Intersteno conference. Eugeni analyzed the speed (measured in words per minute), accuracy (assessed with the IRA method) and delay of the following five ILS methods:

- ILS1: the Italian-English translation of the opening ceremony by a simultaneous interpreter (Italian audio to English audio) and a stenotypist (English audio into English subtitles);
- ILS2: the German-English translation of the council meeting by an interlingual velotypist (German audio into English subtitles);
- ILS3: the English-French translation of the general conference by an intralingual velotypist (English audio into word-for-word English subtitles) and MT (English subtitles into French subtitles);
- ILS 4: the English-French translation of the council meeting by an intralingual respeaker (English audio into Plain English subtitles) and MT (Plain English subtitles into French subtitles);
- ILS 5: the English-French translation of the assembly by ASR (English audio into English subtitles), MT (English subtitles into French subtitles) and a live editor who corrected mistakes in the French subtitles.

The results (see Table 1) show that the use of MT increases the amount of words included in the subtitles (speed measured in words per minute) and decreases their accuracy. Eugeni focuses on the stark contrast in the results of ILS3 and ILS4, which involved a similar set-up. Although ILS4 causes more delay than ILS3, it is significantly more accurate, because MT performs better with texts that have been intralinguistically manipulated (as in ILS4, where English audio was turned into Plain English subtitles) than with texts that include a verbatim rendition of the original audio (as in ILS3).

| | ILS 1 | ILS 2 | ILS 3 | ILS 4 | ILS 5 |
|---|---|---|---|---|---|
| WPM | 110 (20.4%) | 114 (17.8%) | 141 (3.3%) | 118 (14%) | 132 (4.5%) |
| IRA | 97.3% | 95.8% | 71.2% | 92.1% | 86.9% |
| DELAY | 7.3″ | 4.8″ | 4.3″ | 6.8″ | 8.1 |

**Table 1: Results of Eugeni´s (2020) research on ILS: speed, accuracy and delay**

Finally, in a recent study, Fantinuoli and Prandi (2021) looked at the quality of what they describe as automatic speech translation (the same as Eugeni's ILS5 above but without a live editor) by comparing it to the manual transcription of the input produced by simultaneous interpreters. The results show a better performance by the human interpreters in terms of intelligibility (understood here as the clarity, comprehensibility, linguistic acceptability and stylistic correctness of the target text) and a better performance by the machine in terms of informativeness (semantic information when compared to the source text). The authors highlight some limitations, though, not least the fact that the analysis involves the comparison of written output (automatic speech translation) and transcribed oral output (human interpreting).

Given the growing demand for ILS, the industry is currently testing different methods of ILS with more or less human intervention to ascertain what works better in terms of accuracy, delay and resources required. The studies mentioned here, and especially the one conducted by Eugeni (2020), pave the way for an empirical study that can compare the most commonly used methods in an empirical and controlled manner. This requires the use of the same source text and the same language combination for all methods and the selection of methods that use speech recognition, which, along with MT, is the most popular type of technology when it comes to ILS.
The next sections of this paper introduce the first study to compare the efficiency of five different speech-recognition-based ILS methods.

## 3. The study

### 3.1. Research question and methods

The research question around which the experiment was built is "How do different ILS workflows compare in terms of efficiency (understood here as a combination of accuracy, delay and resources required)?" Accuracy and delay have always been regarded as key factors in live subtitling, not least because of the impossibility (at least until now) of producing perfectly accurate and synchronous live subtitles and because they are closely connected (Romero-Fresco, 2011). In this trade-off between accuracy and delay, companies that prioritize accuracy often choose workflows involving a subtitler and an editor or corrector, or even several correctors, as in France (Romero-Fresco and Eugeni, 2020). This increases both the accuracy and the delay of the subtitles. In contrast, those companies that prioritize the reduction of delay tend to use single subtitlers/respeakers, who are instructed to display their subtitles on screen as they are being produced. This means that any errors and inaccuracies will only be corrected once the subtitles have been shown to all viewers. The third variable tested here, resources, is not normally included in academic studies of live subtitling quality, but since it is directly related to cost, it is an important factor for companies to decide for or against a particular ILS method.
The methods chosen for this experiment were selected as the most relevant and promising ILS workflows following consultation with the researchers involved in ILSA (Interlingual Live Subtitling for Access), the only EU-funded project on the subject, with professionals who were trained at the STTI course run by Universidade de Vigo and with leading international accessibility companies such as AiMedia and Red Bee Media. They range from different forms of STTI combining humans and machines (modes 1, 2, 3 and 4) to a fully automatic form of ILS not involving human STTI (Mode 5):

- Mode 1. Interlingual respeaking: a single STT interpreter who listens to the source text in English and dictates it into Spanish (English audio into Spanish subtitles).

- Mode 2. Simultaneous interpreting + intralingual respeaking: a simultaneous interpreter who translates the source text into Spanish (English audio into Spanish audio) and an intralingual respeaker (Spanish audio into Spanish subtitles).

- Mode 3. Simultaneous interpreting + ASR: a simultaneous interpreter who translates the source text into Spanish (English audio into Spanish audio) and an ASR engine (Spanish audio into Spanish subtitles).

- Mode 4. Intralingual respeaking + MT: an intralingual respeaker (English audio into English subtitles) and a MT engine (English subtitles into Spanish subtitles).

- Mode 5. ASR + MT: an ASR engine (English audio into English subtitles) and a MT engine (English subtitles into Spanish subtitles).

3.2. Materials

As a first study comparing the five above-mentioned ILS workflows, the two source texts selected were simple in terms of the parameters identified by Davitti and Sandrelli (2020) for ILS material: content (familiarity of subject matter and level of technicality), language (lexical density and syntactic complexity), delivery (speech rate and accent), context (single-speaker monologue vs multi-party interaction, use of visual aids, etc.) and sound quality. Both were freely available TED Talks produced by young female speakers in a real-life environment.

The first one was a speech by the Swedish activist Greta Thunberg entitled "School strike for climate – save the world by changing the rules" delivered at a TEDx Stockholm forum in November 2018. It is a short (1364 words), non-scripted speech delivered at a fairly slow pace (124 words per minute) over 11 minutes, with a clear non-native accent and a content featuring a low level of technicality (with the exception of a few specialized terms) and low lexical density and syntactic complexity.

The second source text was the speech "Readers are Leaders", given by Phuong Anh Nguyen Ngoc, a student from Vinschool Times City, at another TEDx event held in Hanoi (Vietnam) in June 2020. This is a slightly longer (2.537 words delivered over 15 minutes), non-scripted speech, delivered at a significantly faster pace (165 wpm), with a fairly clear non-native accent, and a content featuring a low level of technicality and low lexical density and syntactic complexity.

All in all, both texts selected for the study were considered comfortable for simultaneous or live translation and only differed in their length and speed of delivery, which may provide tentative clues as to the extent to which these two factors can impact on ILS performance.

3.3. Participants and set-up

There were two types of participants in the study: humans and machines. The human team was made up of:

-Two professional interlingual respeakers (for Mode 1), who had received specific training in this skill at the STTI course run by Universidade de Vigo. One of them has five years of prior experience as an intralingual respeaker and the other one more than 15 years as a simultaneous interpreter.

-Two simultaneous interpreters (for Mode 2 and Mode 3) with more than 20 years of experience in the professional market.

-Two intralingual respeakers (Spanish to Spanish) (Mode 2), recently graduated at the STTI course run by Universidade de Vigo.

-Two intralingual respeakers (English to English) (Mode 4) with over five years of experience in the professional market.

For the purpose of this study, the participants were only asked to simultaneously interpret or respeak (intra or interlingually) each one of the source texts once, normally in the technique for which they received extensive training.

As far as the machines are concerned, two types of software were used:

-The automatic speech recognition (ASR) software by Apptek, a US-based company specialising in human language technology and a leader in the area of automatic accessibility.

- Google Translate, the popular neural machine translation service developed by Google in 2006. It is currently used by 500 million users, with more than 100 billion words translated daily into 109 languages.

Since the research was conducted in Spring 2021, during the peak of the COVID-19 pandemic, it was carried out entirely online. The source texts were made available to the participants via Google Drive, along with specific instructions about how to complete the task.

3.4. Methodology

As mentioned above, in order to compare the five ILS workflows tested, the variables selected for analysis were accuracy, delay and resources. The accuracy of the output produced by the different methods was analyzed using the NTR model (Romero-Fresco and Pöchhacker, 2017). Created as a development of the NER model (Romero-Fresco and Martínez, 2014), which is currently used by companies and broadcasters all over the world to assess the accuracy of intralingual live subtitles, the NTR model compares source and target texts and distinguishes, first of all, two types of errors: translation errors and recognition errors (see Figure 2). Translation errors are those stemming from the translation process as such, and can relate to content (omissions, additions and substitutions of information) or language (correctness and style). Recognition errors are misrecognitions originated when the target text is dictated to the ASR software. Crucially, both translation and recognition errors are further classified as minor, major or critical, depending on the impact they have on the users' comprehension. Minor errors do not hinder comprehension, major errors cause users to lose information from the source text and critical errors result in misinformation, as they provide users with incorrect information that is credible in the context in which it occurs. The NTR model also accounts for effective editions, that is, deviations from the source text that do not involve a loss of information and that could even enhance the communicative effectiveness of the subtitles. As in the NER model, 98% is set as the threshold that interlingual live subtitles must reach to be acceptable with the NTR model. Since such high percentages are more common in intralingual than in interlingual live subtitling, the NTR model recalculates the accuracy rate to a more standard 10-point scale (with 5/10 as a threshold for acceptable subtitles).

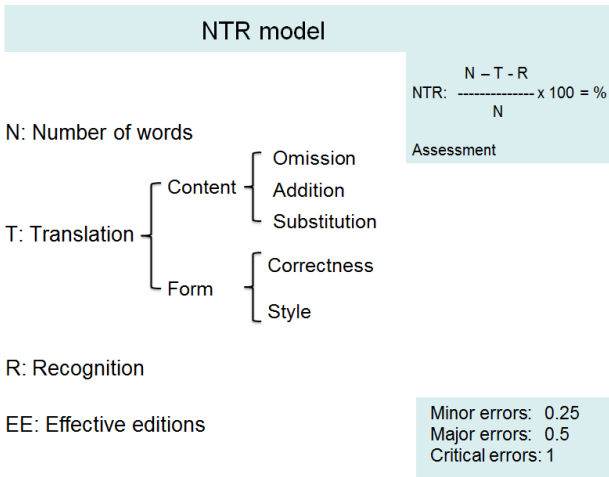


**Figure 2: The NTR Model (Romero-Fresco and Pöchhacker, 2017)**

The delay of the ILS output produced by the five ILS workflows tested was measured following an adaptation of the formula specified in the official Spanish guidelines on subtitling for the deaf and hard of hearing (Norma UNE 153010). One sentence ending was identified every minute of each source text (11 instances in the first speech and 15 instances in the second speech –the same for all methods) in order to calculate the lag between the moment in which a specific oral utterance was spoken and the moment in which it was displayed as subtitles on screen.
Finally, the resources required for each ILS method were calculated on the basis of the length of the meeting. For the purpose of this research, a 30m-1h meeting was considered as a standard scenario that is representative of the current professional reality in this area.

3.5. Results
Table 2 shows the results obtained as far as accuracy is concerned. It includes, for every method, NTR scores in percentage and in a 1/10 scale for both source texts, a total average accuracy rate and the place it occupies in the ranking alongside the other methods.

| **ACCURACY** | | | | |
|---|---|---|---|---|
| **Modes** | **ST1** | **ST2** | **Total** | **Rank** |
| 1. Interlingual Respeaking | 98.2% (5.5/10) | 98.6% (6.5/10) | 98.4% (6/10) | 3 |
| 2. Simultaneous Interpreting + Intralingual Respeaking | 99% (7.5/10) | 98.7% (7/10) | 98.8% (7/10) | 1/2 |
| 3. Simultaneous Interpreting + Automatic SpeechRecognition | 97.6% (4/10) | 97.3% (3/10) | 97.4% (3.5/10) | 4 |
| 4. Intralingual Respeaking + Machine Translation | 99.1% (8/10) | 98.6% (6.5/10) | 98.8% (7/10) | 1/2 |
| 5. Automatic Speech Recognition + Machine Translation | 97.4% (3.5/10) | 97.1% (3/10) | 97.2% (3/10) | 5 |

**Table 2: Results: accuracy**

The results show that the most accurate modes are 2 and 4, which obtain a very good average accuracy rate for both texts (98.8%, 7/10) and even manage to hit the 99% mark for source text 1. Although not quite as accurate, Mode 1 also manages to obtain good results, especially for source text 2. All three modes reach the NTR threshold for both texts. In contrast, modes 3 and 5, the only ones including ASR with no respeaking, fall short off the mark for both texts, obtaining an average score of 3.5/10 and 3/10, respectively.

| DELAY | | | | |
|---|---|---|---|---|
| **Modes** | **ST1** | **ST2** | **Total** | **Rank** |
| 1. Interlingual Respeaking | 4.6s | 4.6s | 4.6s | 2 |
| 2. Simultaneous Interpreting + Intralingual Respeaking | 8.1s | 8.9s | 8.5s | 3 |
| 3. Simultaneous Interpreting + Automatic Speech Recognition | 5.1s | 4.1s | 4.6s | 2 |
| 4. Intralingual Respeaking + Machine Translation | 4.5s | 4.8s | 4.6s | 2 |
| 5. Automatic Speech Recognition + Machine Translation | 3.2s | 3.2s | 3.2s | 1 |

**Table 3: Results: delay**

Finally, Table 4 addresses the resources required by each of the 5 methods for events under and over 30 minutes long. Since resources are directly related to cost, the different methods have also been (speculatively) ranked from the most expensive to the most affordable one.

| RESOURCES | | | |
|---|---|---|---|
| **Modes** | **<30 minutes** | **>30 minutes** | **Rank** |
| 1. Interlingual Respeaking | 1 human | 2 humans | 4 |
| 2. Simultaneous Interpreting + Intralingual Respeaking | 2 humans | 4 humans | 5 |
| 3. Simultaneous Interpreting + Automatic Speech Recognition | 1 human + machine | 2 humans + machine | 3 |
| 4. Intralingual Respeaking + Machine Translation | 1 human + machine | 1 human + machine | 2 |
| 5. Automatic Speech Recognition + Machine Translation | 2 machines | 2 machines | 1 |

**Table 4: Results: human and machine resources**

Partially in line with the results obtained by Eugeni (2020), it seems that the more automatic the method (see for instance Mode 5), the lower the accuracy rate, the shorter the delay and the lower the cost. Conversely, the only method involving two humans (Mode 2) yields the highest accuracy rate, the longest delay and the highest cost. However, the aim here is not necessarily to ascertain which is the most effective method, but rather to identify strengths, weaknesses and factors to consider if any of them is to be used professionally. To this end, the next section looks at each method more closely, from most to least efficient.

3.6. Discussion

- Mode 4 (Intralingual respeaking + MT): very good accuracy, short delay, low cost

As was the case in Eugeni's experiment (2020), the combination of an intralingual respeaker and MT, which is not very common in the industry, obtained very good results in terms of accuracy. Unlike in Mode 1, the respeaker is not concerned here with translation and can therefore focus on "taming" the original speech and turning into coherent written English subtitles, devoid of the hesitations, false starts and errors made by the speakers. This facilitates the job of the MT, allowing it to obtain much better results than in Mode 5. Since the performance of MT technology is gradually improving, this method could be a promising one. However, questions remain as to whether this method may work with more specialized source texts and whether it may be too risky for important events, since the last part of the process is MT. In modes 1 and 2, a human acts as a final gatekeeper to decide what text is displayed on screen as subtitles. In modes 3, 4 and 5, unless human revision is added at the end of the process (which would also increase delay), it is the machines that have the final say, which could be problematic.

- Mode 2 (SI + Intralingual respeaking): very good accuracy, long delay, high cost

Mode 2, which is currently being used in the professional industry, obtained the highest accuracy results. This may be explained by the fact that this workflow includes one human in charge of translation and another one concerned with turning the translated audio into written subtitles, with control of the final output, as both professionals can detect and correct errors. Additionally, this method offers simultaneous interpreting output for audiences who wish to receive a translation through audio and subtitles for those who wish to receive a written text. On the downside, the cost is higher than that of other methods (as it involves two humans in events under 30 minutes and four humans in longer events) and the delay is almost twice as long as with the other methods. When working in this mode, interpreters should be advised to keep décalage to a minimum, but it is hard to see how the final delay can be lower than 6-7 seconds (2s-3s for interpreting and 3-4s for intralingual respeaking).

- Mode 1. (Interlingual respeaking): good accuracy, short delay, low cost

Mode 1, currently used in the industry and the focus of the EU-funded project ILSA, offers and interesting compromise regarding the variables assessed in this paper. It ranks third in accuracy with a good result (6/10), second in delay with 4.6s and it is more affordable than the other "fully human" method (Mode 2). More importantly, it is the only method that allows complete human control over the final output. In Mode 2, the mistakes of the SI cannot normally be corrected by the respeaker, who does not have access to the source text. In Mode 1, the respeaker has full control, as there is only one professional involved. The main drawback is that interlingual respeaking is cognitively challenging and it is still to be determined whether or not it will be effective in specialized settings.

- Mode 3. (SI + ASR): good accuracy, short delay, low cost

Mode 3 has been identified as a promising workflow for ILS. As was the case with Mode 2, it offers interpreting output for audiences who wish to receive audio and subtitles for audiences who wish to receive text. It combines human translation with a technology that is constantly being improved, which should lead to good accuracy, short delay and affordable cost. However, none of the four texts analyzed here (two texts translated by two interpreters and then recognized by the ASR software) reached

the 98% threshold, mainly because the ASR software struggled to recognize the interpreters' utterances. For this mode to work, it may be necessary to train interpreters so that they enunciate clearly and avoid self-corrections and hesitations. Some thought may also need to be given to the context in which this mode is used, as there is no human control over the final output.

- Mode 5. (ASR + MT): insufficient accuracy, very short delay, very low cost

This is a very attractive mode for many stakeholders in the industry, as it is the most affordable way to provide ILS and produces the shortest delay. Accuracy, however, remains an issue, as none of the texts reached the 98% threshold. As in the case of Mode 3, the lack of human control over the final output (which becomes more significant here as the errors of the ASR software are added to those of the MT) often results in unintelligible output. It is hard to see how this mode can be used with confidence in important or specialized settings, but its accuracy should increase as ASR and MT technologies continue to improve and in any case it remains as a viable option for contexts in which hiring humans is not an option.

**4. Final thoughts: Horses for courses**

The study presented here is, to the best of our knowledge, the first one to test five different speech-recognition-based ILS methods using the same source texts. This means that it has a pioneering contribution to make, but also that it is only a first attempt to assess the efficiency of these ILS methods. Any conclusions and claims resulting from this study must bear this limitation in mind, as well as the fact that the sample analyzed was small and limited to the language combination (EN-ES). Further studies could test larger samples, including different genres (for instance spontaneous, unscripted interactions involving several speakers, such as TV talks shows, political or social debates, online meetings, etc) and other language combinations. Given that the degree of maturity achieved by speech recognition and machine translation technology varies across languages, a replication of this study with a different language combination may well yield different results.

Be that as it may, the results obtained in this study and those presented at recent conferences on this topic (Dawson, 2021; Pagano, 2021; Davitti, 2021) are beginning to point to some interesting trends, as well as to a scenario of horses for courses, where different methods may be more or less useful or feasible depending on the context, the difficulty of the source text, the resources available, etc.

The methods that seem to be more commonly used in the industry at present, Mode 1 and Mode 2, have fared well. In line with what has been found by Dawson (2019), interlingual respeaking seems to yield good results both in terms of accuracy and (as shown in this study) delay. The cost is lower than that of Mode 2 and similar to that of the other methods that are not fully automatic, which makes interlingual respeaking a promising method to provide ILS.

Some companies seem to be choosing Mode 2 (a combination of a simultaneous interpreter and an intralingual respeaker) for specific contexts, so as to ensure a quality output by having one human in charge of translation and another one concerned with turning this oral translation into written subtitles. This seems a good option as long as delay and cost are not an issue. Now that the EU is planning to provide accessibility to live events at its institutions, hiring intralingual respeakers who can turn the simultaneous interpreters' audio into subtitles can ensure that the translations provided are accessible to people with hearing loss and to anyone who may need to access a written output.

As for Mode 4, the good results obtained in this study confirm what was already found by Eugeni (2020) –MT can be very accurate if the original speech is initially

converted into coherent written text. Interestingly, these results may shed some much-needed hopeful light on the future of intralingual respeaking. As the quality of ASR improves, more and more broadcasters are beginning to roll out fully automatic intralingual subtitles, thus replacing intralingual respeakers (Fresno and Romero, 2021). In an interlingual context, the role of the intralingual respeaker in Mode 2 may also end up being replaced by an ASR engine (Mode 3) when its accuracy improves. Against this backdrop, the use of intralingual respeaking in ILS as a preliminary stage to MT may be an unexpected and promising line of work for intralingual respeakers. Still, as discussed above, employers choosing this mode need to bear in mind that the final output is unsupervised.

Finally, the relatively poor results obtained by modes 3 and 5 seem to suggest that, for all its improvements, ASR still has some way to go until it can deliver quality ILS output. As found by Fresno and Romero-Fresco (2021) in the case of intralingual live subtitles, the accuracy of ASR increases in relatively controlled conditions (one speaker, pre-prepared speech, little or no noise, etc.). In order to transcribe the words of a simultaneous interpreter accurately, the interpreter would need to be trained accordingly. This does not mean the interpreter must be trained as a respeaker, but they would need to enunciate clearly, avoid hesitations and, in general, be aware that their translation is only a first step towards a final written output.

This leads to a final reflection about training. The industry is testing different methods for ILS, technologies such as ASR and MT are constantly improving and the demand for written/accessible translations of live events is increasing. Yet, most translation and interpreting programmes at higher education are still designed around a prototypical conception of translation and interpreting as two separate entities. It is our hope that the present study, as well as others that are currently being conducted, can help to spur a revision of translation and interpreting training that can prepare students and trainees for the complex, multicultural, technological and diverse world we are living in.

**Acknowledgements**

We would like to thank Yota Georgakopoulou and AppTek for allowing us to use their ASR software and especially the respeakers and interpreters who have taken part in the study for their time and generosity.

This research has been conducted within the framework and with the support of the Galician-government-funded project "Proxecto de Excelencia 2017 - Observatorio Galego de Accesibilidade aos Medios" and of the Spanish-government-funded project "The Quality of Live Subtitling (QuaLiSub): A regional, national and international study" (PID2020-117738RB-I00).

**Bibliographic references**

Davitti, E. & Sandrelli, A. (2020). Embracing the Complexity: A Pilot Study on Interlingual Respeaking. Journal of Audiovisual Translation, 3(2), 103-139.

Davitti, E. & Korybski, T. (2021). Automating the Interlingual Respeaking Workflow: Hype or Reality? Insights from an Empirical Pilot Study. Paper presented on 24 September 2021 at the Languages and the Media Conference (Berlin, Germany).

Dawson, H. (2019). Feasibility, quality and assessment of interlingual live subtitling: A pilot study. Journal of Audiovisual Translation, 2(2), 36-56.

Dawson, H. (2021). Exploring the quality of different live subtitling methods: a Spanish to English follow up case study. Paper presented on 17 September 2021 at the 7th IATIS conference held at Pompeu Fabra University (Barcelona, Spain).

De Korte, T. (2006). Live inter-lingual subtitling in the Netherlands: Historical background and current practice. inTRAlinea Special issue: Respeaking. Retrieved from http://www.intralinea.org/specials/article/Live_inter-lingual_subtitling_in_the_Netherlands

Den Boer, C. M. (2001). Live interlingual subtitling. In Y. Gambier & H. Gottlieb (Eds.), (Multi)media translation: Concepts, practices, and research, 167-172. Amsterdam: John Benjamins. doi:10.1075/btl.34.20boe
Eichmeyer-Hell, D. (2021). Speech recognition (Respeaking) vs. the conventional method (Keyboard): A quality-oriented comparison of speech-to-text interpreting techniques and addressee preferences. In Susanne J. Jekat, Steffen Puhl, Luisa Carrer, Alexa Lintner (Eds.) Proceedings of the 3rd Swiss Conference on Barrier-free Communication (BfC 2020)
Eugeni, C. (2020). Human-Computer Interaction in Diamesic Translation. Multilingual Live Subtitling. In: Dejica, D, Eugeni, C and Dejica-Carţiş, A, (eds.) Translation Studies and Information Technology - New Pathways for Researchers, Teachers and Professionals. Translation Studies Series. Editura Politehnica, Politehnica University Timişoara, 19-31.
Fantinuoli, C. & Prandi,, B. (2021). Towards the evaluation of automatic simultaneous speech translation from a communicative perspective. IWSLT.
Fresno, N. & Romero-Fresco, P. (2021). Automatic Live Subtitling in English: Taking the Bad with the Good. Paper presented on 24 September 2021 at the Languages and the Media Conference (Berlin, Germany).
Gottlieb, H. (2017). "Semiotics and Translation". In K. Malmkjær (ed.) The Routledge Handbook of Translation Studies and Linguistics. London & New York: Routledge, 45-63.
Hewett, T. T., Baecker, R., Card, S., Carey, T., Gasen, J., Mantei, M., Perlman, G., Strong, G. & Verplank, W. (1992). ACM SIGCHI curricula for human-computer interaction, Broadway: ACM, available at https://www.researchgate.net/publication/234823126_ACM_SIGCHI_curricula_for_human-computer_interaction [accessed December 2019].
Kurz, I., & Katschinka, L. (1988). Live subtitling: A first experiment on Austrian TV. In P. Nekeman (Ed.), Translation, our future: Proceedings of the XIth World Congress of FIT (479-483). Maastricht: Euroterm.
Pagano, A. (2021). Interlingual respeaking vs other live (sub)titling methods: an EN>IT follow-up case study. Paper presented on 17 September 2021 at the 7th IATIS conference held at Pompeu Fabra University (Barcelona, Spain).
Pochhacker, F. (2019). Moving boundaries in interpreting. In H. V. Dam, M. N. Brøgger, & K. K. Zethsen (Eds.), Moving boundaries in translation studies, London: Routledge. 45-63.
Pochhacker, F. & Remael, A. (2019). New efforts?: A competence-oriented task analysis of interlingual live subtitling. Linguistica Antverpiensia, New Series: Themes in Translation Studies, 18, 130-143.
Romero-Fresco, P. (2011). Subtitling through Speech Recognition: Respeaking. Routledge.
Romero-Fresco, P. & Martinez, J. (2015). Accuracy Rate in Live Subtitling: The NER Model. In J. Díaz-Cintas & R. Baños Piñero (Eds.), Audiovisual Translation in a Global Context. Mapping an Ever-changing Landscape, 28-50. Palgrave.
Romero-Fresco, P. & Pochhacker, 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. & Eugeni, C. (2020). Live subtitling through respeaking. In Ł. Bogucki & M. Deckert (Eds.), Handbook of Audiovisual Translation and Media Accessibility, Palgrave, 269-297
Stinson, M. S. (2015). Speech-to-text interpreting. In F. Pöchhacker (Ed.), Routledge encyclopedia of interpreting studies, London: Routledge, 399-400

*Words: 6517*
*Characters: 42 488 (23,60 standard pages)*

Prof. Dr. Pablo Romero-Fresco
Prof. Dr. Luis Alonso-Bacigalupe
Universidade de Vigo
36310 Pontevedra
Spain
promero@uvigo.es
lalonso@uvigo.es