Journal of Audiovisual Translation

Embracing the Complexity: A Pilot Study on Interlingual Respeaking

Elena Davitti✉
Centre for Translation Studies, University of Surrey

Annalisa Sandrelli✉
Facoltà di Interpretariato e Traduzione, Università degli Studi Internazionali di Roma- UNINT

Abstract

This paper presents the key findings of the pilot phase of SMART (*Shaping Multilingual Access through Respeaking Technology*), a multidisciplinary international project focusing on interlingual respeaking (IRSP) for real-time speech-to-text. SMART addresses key questions around IRSP feasibility, quality and competences. The pilot project is based on experiments involving 25 postgraduate students who performed two IRSP tasks (English–Italian) after a crash course. The analysis triangulates subtitle accuracy rates with participants' subjective ratings and retrospective self-analysis. The best performers were those with a composite skillset, including interpreting/subtitling and interpreting/subtitling/respeaking. Participants indicated multitasking, time-lag, and monitoring of the speech recognition software output as the main difficulties; together with the great variability in performance, personal traits emerged as likely to affect performance. This pilot lays the conceptual and methodological foundations for a larger project involving professionals, to address a set of urgent questions for the industry.

Key words: interlingual respeaking, speech-to-text, live subtitling, interpreting, human-machine interaction.

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

Editor(s): A. Matamala & J. Pedersen

Received: February 18, 2020

Accepted: July 22, 2020

Published: December 18, 2020

Funding: This work was supported by the Pump Priming Fund of the University of Surrey and the Research Fund of the Università degli Studi Internazionali di Roma - UNINT and served as a basis for an Economic and Social Research Council (UK) funded project called SMART (*Shaping Multilingual Access through Respeaking Technology* – project reference ES/T002530/1, 2020–2022). The project is led by the University of Surrey and relies on a consortium including the University of Roehampton (UK), UNINT in Rome (Italy) and the University of Vigo (Spain), as well as industrial stakeholders.

✉ e.davitti@surrey.ac.uk, https://orcid.org/0000-0002-7156-9275
✉ annalisa.sandrelli@unint.eu, https://orcid.org/0000-0001-6010-4862

1. Introduction[1]

This paper focuses on interlingual respeaking (IRSP) as an innovative method for real-time speech-to-text from one language to another. IRSP is, in its process, a "form of simultaneous interpreting" (Romero-Fresco and Pöchhacker, 2017, p. 157) and, in its product, text "displayed on screen with the shortest possible delay" (Romero-Fresco, 2011, p. 1). This diamesic shift (from spoken to written) is enabled by the interaction between human, i.e. the respeaker(s) producing audio input in the target language, and machine, i.e. the speech recognition (SR) software turning that input into text, which is a fundamental characteristic of respeaking in both its intra- and interlingual variants. IRSP adds a translation component to the already challenging task of respeaking in the same language, which entails listening, speaking, adding punctuation and any additional content labels orally, articulating, controlling prosody to minimise SR errors, monitoring the transcript and editing it for comprehensibility and readability, whenever necessary. As a relatively new practice, there is still a lack of consensus around the terminology used to refer to it (§2), with different expressions highlighting different dimensions of what is ultimately the same technology-enabled hybrid modality of Translation (in its broadest sense), at the crossroads of interpreting and subtitling. Therefore, it might be helpful to spell out some key dimensions and relevant parameters of this practice.

Starting from the communicative dimension, IRSP is emerging as a method to be used in live (or semi-live) programmes or events, where real-time interlingual speech-to-text transfer is required.[2] This covers a range of settings (e.g. TV, digital radio, conferences, workplace, political, educational,…), event types (e.g. breaking news, award speeches, business meetings, parliamentary debates, classroom interaction, MOOCs, museum tours,…) and formats (e.g. monologic vs dialogic vs multi-party interaction).

Further parameters are related to specific features of the (spoken) source language (SL) input and (written) target language (TL) output. In relation to the former, the variables that may affect IRSP performance include topic (general vs highly specialised); speaker's accent (native vs non-native) and speed (original speech rate – OSR); degree of planning (i.e. impromptu vs planned speech and, in the latter case, whether access to the script is granted to respeakers ahead of the event or broadcast); and the presence of visual aids (e.g. slides, graphs).

In relation to the TL output, the variables include display format (what), mode (how), channel (where) and latency (when). The output of the IRSP process can be formatted as subtitles proper (e.g. on TV or during parliamentary debates) or as real-time text (e.g. at conferences or lectures). Depending on the software being used, the output can be displayed as scrolling (i.e. character by character, syllable

[1] Although the paper is the product of a joint effort, Elena Davitti wrote sections §1, 3, 4.2 and 4.4, while Annalisa Sandrelli wrote sections §2, 4.1, 4.3 and 5.

[2] IRSP speed and accuracy for the subtitling of pre-recorded programmes (e.g. films, series, documentaries) is an interesting question that would require further research.

by syllable or word by word) or block text (i.e. in blocks of a few words or even a full line or two lines). The display mode has an impact on readability, with block subtitles being easier to read (e.g. Ofcom, 2013), and on production, with scrolling subtitles being associated with (near-) verbatim subtitling and block subtitles with a degree of editing (Van Waes, Leijten, & Remael, 2013). The display channel may also vary: the IRSP output can be integrated into the TV image, projected onto a screen at a live event or made available on smartphones, tablets or laptops. Latency, i.e. the delay between the source speech and real-time target text, will also vary in relation to the output delivery and degree of editing.

The spatial dimension depends on the number and location of respeaker(s) and users of the IRSP service (audience and speakers), as well as on their relative distribution. Set-ups can vary, which makes direct comparison across settings very complex. Remael, Van Waes, & Leijtenet (2016, pp. 125–126) identify three main configurations, i.e. the Mono, Duo and Multi Live Subtitling (LS) models, where differences lie in the number of professionals involved and distribution of tasks. Configurations can be placed on a continuum from one respeaker handling the whole multitasking IRSP process to a team sharing respeaking, monitoring, editing and broadcasting tasks. Different configurations have implications for task coordination and text synchronisation, with repercussions on latency.

In terms of participants' relative distribution, some terminology borrowed from the field of remote interpreting[3] can provide a broad framework to classify different configurations: respeakers may be located onsite, sharing the same physical space as the participants (event venue or broadcaster station; see Eichmeyer, 2018) or working from its vicinity (proximal IRSP, e.g. a separate room within the venue; see Moores, 2020). There may be hybrid scenarios with one or more remote speakers (e.g. connected to the venue) and/or a remote audience (e.g. a TV or radio interview with audience located virtually anywhere). Respeakers may also provide their service remotely (i.e. by being geographically separated from the main participants) either from an access service provider studio (functioning as a hub) or from home via a platform (distal IRSP). In such cases, the speaker(s) and audience may be entirely or partially co-located (e.g. a live conference streamed online) or entirely distributed (e.g. multipoint, virtual meetings). In the case of a Duo-LS or Multi-LS model, mixed (semi-presence; Eichmeyer, 2018) scenarios include configurations where one respeaker is on-site and one is working remotely. The variety of working set-ups and participants' distribution open up the possibility of further spatial configurations.

Spatial dimension is closely linked to technological dimension, which includes aspects related to the technical solutions (software, hardware, cloud-based, ad hoc) in place to deliver the service across different set-ups. This dimension encompasses aspects of human-machine interaction, such as ease of use, impact of audio and video quality on performance, interface user-friendliness, operability,

[3] The distinction between "proximal" and "distal" configurations for remote interpreting is used, for instance, by the Council of Europe. See https://aiic.ch/press/remote-interpreting-ws1/ and Braun, 2015.

functionality, support to online interaction; it also includes the potential to integrate automatic speech recognition (ASR), machine translation and/or post-editing in the IRSP workflow, which can impact on the IRSP process, tasks, roles and working conditions. These are in turn closely related to a range of socio-economic, cognitive and ergonomic factors, raising interesting questions for future research (e.g. the impact on workflow automation and on how to optimise the role of human factors in systems design, the short-/long-term effects on fatigue and cognitive load, well-being and adaptation, as well as professional recognition and remuneration).

As research on IRSP is still in its early stages, questions around its feasibility, quality and competences need further exploration and empirical grounding. To this end, our recently funded project SMART (*Shaping Multilingual Access through Respeaking Technology*) aims to develop a multi-method, multifactorial design to explore different facets of IRSP. The focus is on its "purest" form, i.e. the Mono-LS model, and on its peri-process phase (Pöchhacker & Remael, 2019; see §2.1). After a brief overview of relevant literature (§2), the paper introduces the conceptual underpinning of the approach that SMART aims to develop (§3). This is followed by the explanation of the selected methodological components tested in the pilot conducted in preparation for the larger project (§3.1); profiling of participants, description of administered tasks and data analysis procedure (§3.2); selected quantitative and qualitative findings, ideas for further research (§4), and concluding remarks (§5).

2. Literature Review

To date, respeaking research has developed along three strands: process and required competences, subtitle quality evaluation, and delivery of respeaking services in various settings. These issues have been investigated using data generated in experimental and real-life settings. We have taken stock of the available literature on the respeaking process and competences (§2.1) and on the accuracy evaluation models, with a special focus on IRSP (§2.2).

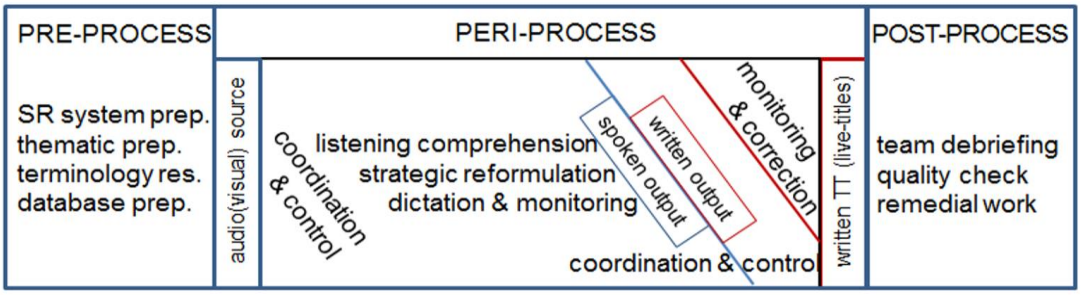
2.1. Respeaking Process and Competences

Early research on intralingual respeaking focused on its similarities with simultaneous interpreting (SI), namely listening and speaking at the same time, multitasking and focused concentration (Romero-Fresco, 2011). However, as respeaking entails a diamesic shift and a technology-mediated, multi-step process, several task-specific skills are also involved. The distinction made by Romero-Fresco (2011) between the skills used before, during and after the process makes it easier to tease apart those shared by respeaking and subtitling/SI and those that are respeaking-specific. Along similar lines, Pöchhacker and Remael (2019, p. 137) have proposed a process model for interlingual live subtitling via respeaking which includes pre-, peri- and post-process phases (Figure 1). The pre-process entails the preparation tasks required to ensure smooth human-machine interaction,

including thematic and terminological preparation, the training of the SR software and use of macros.[4] As regards the peri-process (i.e. the actual respeaking, here renamed transpeaking), one person may produce the spoken output and edit the written output, or a team may share tasks (§1). In the post-process, debriefing, quality checks and remedial work are carried out (e.g. adding new words to the software vocabulary and devising solutions to avoid recognition errors).

Figure 1.

Live Subtitling via Respeaking Process Model



SR= speech recognition; prep. = preparation; res.= research; TT= target text

Source: Pöchhacker and Remael (2019, p. 137)

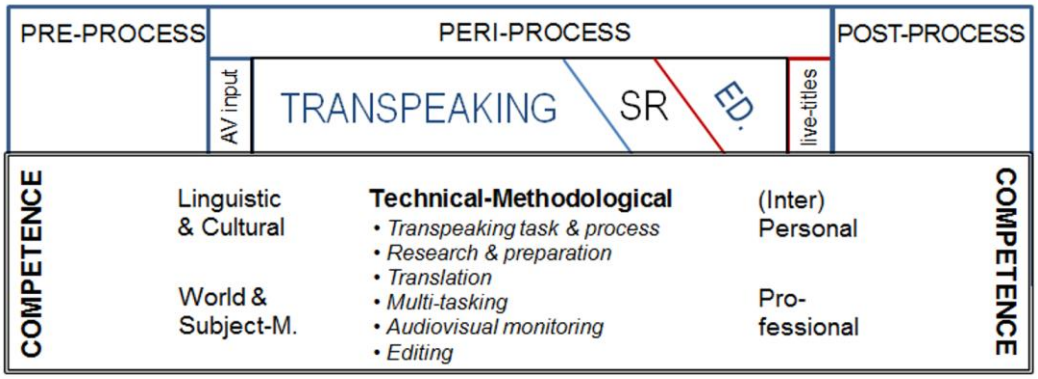
Given the many similarities between SI and IRSP, Pöchhacker and Remael propose an adaptation of Gile's Effort Model (2015) to describe IRSP in terms of the allocation of processing capacity to the various tasks. The original Effort Model of SI identified a Listening and Analysis Effort, a Memory Effort, a Production Effort and a Coordination Effort. Likewise, IRSP requires a SL Listening and Analysis Effort, the involvement of Working Memory, a TL Production Effort and a Coordination and Control Effort. However, the Production Effort includes some IRSP-specific components, since "the transpeaker's output must be geared not to human listeners but to the capabilities and settings of the sofware" (Pöchhacker & Remael 2019, p. 135). Therefore, production entails "strategic reformulation" (to prevent SR errors and to keep up with the original speaker's pace), "dictation" (articulating words clearly and adding oral punctuation) and "coordination and control" (auditory monitoring). The intermediary text thus produced by the respeaker (i.e. "spoken output" in Figure 1) is transcribed by the SR software. The TL transcript (i.e. "written output") needs to be checked and, if necessary, corrected and formatted; if this editing phase is carried out by the same person performing the respeaking, an additional coordination and control effort is required (visual monitoring of the text and manual correction).

[4] Macros are shortcuts, i.e. special voice commands used to transcribe frequent names or phrases or to apply specific house styles to the subtitles, such as speaker labels to help deaf users identify who is speaking.

Building on the above process model, Pöchhacker and Remael (2019, p. 138) propose a competence model, integrating technical-methodological competences, linguistic and cultural competences, world and subject-matter knowledge, interpersonal skills and professional skills (Figure 2).

Figure 2.

IRSP Competences



SR= speech recognition; AV= audiovisual; ED= editing; Subject-M= subject matter

Source: Pöchhacker and Remael (2019, p. 138)

As was mentioned in §1, this area of research is rather new and the available results are not yet conclusive. As Pöchhacker and Remael (2019, p. 141) point out,

> the definition and interrelation of the various competences and sub-competences must remain open to discussion, and much further research will be required to understand how they inform the various stages and components of the transpeaking process and the ILS task as a whole.

A small number of recent experimental studies have focused on some IRSP (sub-)competences, to investigate whether a background in related practices can facilitate their acquisition.

An interesting experimental study was carried out by a research team in Poland on a group of 22 professional and trainee interpreters, a group of 22 professional and trainee translators, and a control group of 12 participants without any aforementioned experience.[5] Participants carried out four intralingual respeaking tasks and one IRSP task, and their performances were compared through eye-tracking, EEG recording and the evaluation of subtitle accuracy. The rich data thus generated were studied from various points of view. Findings highlighted that respeaking difficulties can be triggered by very slow and very fast speech rates, overlapping speakers, figures, proper nouns, and certain linguistic features (complex syntax, word play and others); in relation to IRSP, translation

[5] No professional respeakers were available in Poland at the time, as respeaking was still in its infancy.

difficulties were also found to trigger crisis points (Szarkowska, Krejtz, Dutka, & Pilipczuk, 2016; Szarkowska, Dutka, Pilipczuk, & Krejtz, 2017). Chmiel et al. (2017) focused on ear-voice span (EVS) and pauses in intra- and interlingual respeaking. They found a much longer EVS in IRSP, particularly in scripted TV programmes delivered at a high speed with high information density (the news); moreover, average pause duration was longer in IRSP. They concluded that IRSP "requires more cognitive effort than intralingual respeaking as it combines two complex tasks: respeaking and interpreting" (Chmiel et al., 2017, p. 1222). Interestingly, they found no correlation between the participants' backgrounds and the length of EVS and pause duration; despite being used to simultaneous listening and speaking, the interpreters did not have an advantage over the translators and the control group in terms of pauses and time-lag behind the original speaker. Szarkowska, Krejtz, Dutka, & Pilipczuk (2018) focused on subtitle accuracy, assessed both via the NER model (§2.2) and by three independent raters. A correlation was found between the participants' working memory capacity (measured via the reading span test) and subtitle quality, both in terms of accuracy scores and external raters' evaluation. The interpreters were found to have the biggest memory capacity and to obtain the highest accuracy scores, as well as the lowest text reduction rates. These results were interpreted as evidence of a successful transfer of skills from interpreting to respeaking; however, given the small sample size, further validation is needed.

Dawson (2019) reports the results of a pilot experiment carried out in preparation for a larger experiment within the ILSA (*Interlingual Live Subtitling for Access*) project.[6] The pilot involved 10 participants who performed IRSP (English into Spanish) of three short videos (just over two minutes). The participants' background involved subtitling, intralingual respeaking and interpreting combined in various ways (from subtitling alone to all three together); moreover, some were professionals, while others were postgraduate students. They were given some basic training in IRSP, followed by hands-on tasks (for three hours); their performances were analysed with the NTR model (quantitative analysis, §2.2) and by means of self-evaluation questionnaires (qualitative analysis). The majority of participants indicated that, in their opinion, the best-suited profile for a respeaker is an interpreter; however, the best overall performer was a participant with experience in subtitling and respeaking. Therefore, no clear-cut profile emerged from the pilot.

Finally, Dawson and Romero-Fresco (forthcoming) illustrate the results of a four-week training course delivered online within ILSA and based on the above-mentioned pilot. 44 native Spanish speakers with a training background in interpreting (27) and subtitling (17) participated; over half of the interpreters had some subtitling experience and several subtitlers had some interpreting experience. Participants carried out two intralingual respeaking tasks (Spanish into Spanish) and four IRSP tasks (English into Spanish). Results indicate that IRSP is feasible, with over 40% of participants producing relatively accurate subtitles after a short training period (i.e. achieving or exceeding the 98% score

[6] Erasmus+ Programme (reference number 2017-1-ES01-KA203-037948) led by the University of Vigo, http://ka2-ilsa.webs.uvigo.es/

suggested as the minimum quality threshold in the NTR model). The clear-cut interpreters (with no subtitling experience) obtained the best results, followed by the subtitlers with some interpreter training. The clear-cut interpreters made fewer recognition errors, possibly thanks to their familiarity with simultaneous listening and speaking and their ability to use spoken language; they also made the lowest number of omission and substitution errors.[7] Thus, interpreters may initially have a comparative advantage over trainees with other backgrounds, but their skills are no guarantee that they will perform IRSP better than others.

Our pilot study aims to contribute to this developing strand by testing some research methods. Before that, we will briefly present the key models used to evaluate subtitle accuracy.

2.2. Assessing the Quality of Live Subtitles

The key factors that determine the quality of live subtitles are subtitle latency, speed and accuracy. Live subtitles tend to lag behind the original speech; the delay (latency) depends on the original speaker's speech rate, the respeaker's speech rate and the target audience's reading rate. Respeakers need to understand and reformulate the original speaker's message and to add punctuation orally; therefore, if they produce verbatim subtitles, they must utter literally more words than the original speaker. If the original speaker's speed is up to 180 words per minute (wpm), respeakers tend to lag behind by 0–20 words, and this delay increases at higher speeds (Romero-Fresco, 2009, 2011); therefore, in most cases live subtitles are edited to prevent the respeaker from lagging behind too much. If the original speech rate is very high, verbatim subtitles become virtually unreadable, so some editing is necessary to enable the viewers to keep up (Romero-Fresco, 2009). On top of the human-related time-lag, there is a software-related delay, as the recognition software must process the speech data. Finding the right balance in terms of latency is especially important when subtitling TV programmes or live events in which visual information (graphs, maps, slides, etc.) plays a key role, since the subtitles must be relevant to what is on the screen. Although the visibility of the service (which depends on its set-up) is not directly correlated with quality, it is a significant factor in the perception of quality; open captions on TV are much more subject to criticism because errors may be spotted by thousands, which is why the request for reliable evaluation models originally came from broadcasters.

As live subtitles are the product of human-machine interaction, an accuracy evaluation model should account for errors made by both. The models used to evaluate ASR output, such as WER (Word Error Rate), are not suitable, as they classify any discrepancy between the original speech and the

[7] The authors also stress that IRSP training should take into account not only the skills that need to be acquired, but also those which individuals must "unlearn". For example, interpreters are trained to use prosodic devices to convey information, but intonation in respeaking must be as flat as possible (almost "robotic") in order to ensure good recognition.

transcript as an error; by contrast, in respeaking some editing is often advisable, so an omission or a paraphrase is not necessarily an error.

To date, the most widely used accuracy evaluation standard for live subtitling is the NER model (Romero-Fresco & Martínez, 2015, p. 32), which distinguishes the respeaker's edition errors from software-related recognition errors. A score is attributed to each error depending on its severity (minor, standard, serious), in terms of potential impact on viewers' comprehension. A minor error can be recognised or understood, a standard error causes confusion or loss of information, and a serious error introduces false or misleading information. Strategic editing that does not cause information loss or distortion is referred to as a correct edition (CE). CEs are not scored numerically but can be considered in analysis and highlight the respeaker's strengths. The formula in Figure 3 is applied to calculate the accuracy rate; 98% is considered the threshold for intralingual live subtitles (Romero-Fresco, 2011). Several studies have validated the NER model in professional settings (e.g. Ofcom, 2015a, 2015b), while in training it is a useful diagnostic tool to identify recurrent errors.

Figure 3.

The NER Model

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

CE (correct editions):

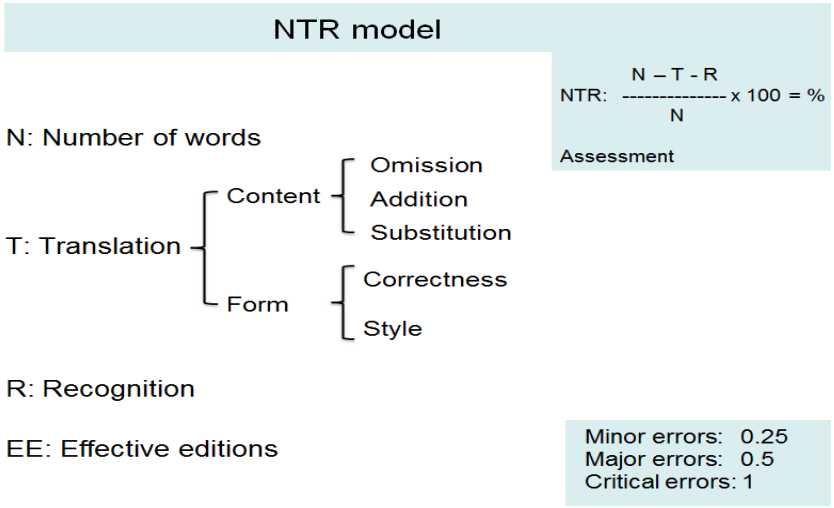
N = number of words
E = edition errors
R = recognition errors

Source: Romero-Fresco & Martínez (2015, p. 32).

IRSP adds a further layer of complexity because of the translation element involved. Romero-Fresco and Pöchhacker (2017, p. 159) have developed the NTR model, a NER-based formula (Figure 4). It distinguishes between recognition errors and translation errors, which include both content-related ones (omissions, additions and substitutions) and form-related ones (grammatical correctness and style). Errors are attributed a different score depending on their severity: minor errors (penalised with a -0.25 point deduction) do not hamper comprehension; major errors (-0.50) cause confusion or loss of information; finally, critical errors (-1) introduce false or misleading information. Like CEs in the NER model, Effective Editions (EEs) in the NTR model account for editions that do not cause loss of information and may improve the text.

Figure 4.

The NTR Model



Source: Romero-Fresco and Pöchhacker (2017, p. 159).

Both models assess accuracy from the viewers' point of view; however, this is based on raters' assessment and it does not involve the actual end-users (e.g. via a reception study). There can be borderline cases in which it is difficult to distinguish between an error and a positive strategy; in addition, error rating is also partially subjective. Therefore, when applying the model, it is advisable to implement a second-marking process. On the basis of the experimental studies mentioned in §2.1, a quality benchmark of 98% has been suggested for IRSP too; however, this has not been validated in professional settings yet, as IRSP itself is not widely practised.

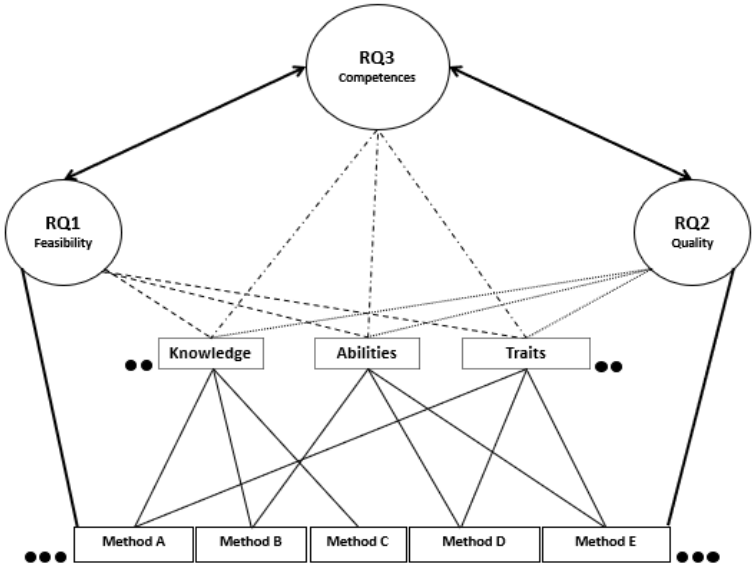
3. SMART Conceptual and Methodological Framework

The hybrid nature of IRSP calls for an interdisciplinary, multi-method and multifactorial approach that correlates findings about process and product to gain insights into different dimensions of this practice. To this end, contributions from Interpreting Studies, Audiovisual Translation, Multimodality, Human–Machine Interaction and Cognitive and Behavioural Sciences are needed. This section presents the design of the conceptual framework under development in SMART and the specific methodological components and procedures tested in the pilot.

The pentagon framework (Figure 5) is inspired by Structural Equation Modelling (SEM), a framework used in the Social Sciences to establish relationships between multifaceted constructs that cannot be observed directly (latent variables) but require triangulation of measurable indicators (dependent variables).

Figure 5.

SMART Pentagon Framework



SMART's research questions (RQs) revolve around three key latent variables for IRSP, namely the feasibility of its process, the quality of the product (analysed in terms of target text accuracy), and the competences moderating IRSP performance. These latent constructs require measurement and triangulation of different variables. SMART aims to bring together all critical endogenous variables (i.e. pertaining to the individual performing the task) at play in human-led IRSP and explore correlations among them, with a view to gaining empirically-grounded insights into this complex practice. These can be grouped into knowledge (declarative and procedural), (cognitive) abilities and (interpersonal) traits. For example, measures of working memory, processing speed, multitasking and self-monitoring abilities can all be grouped under the cognitive dimension; traits such as self-efficacy, goal orientation, and willingness to engage in cognitively complex tasks have yet to be correlated with actual IRSP performance.

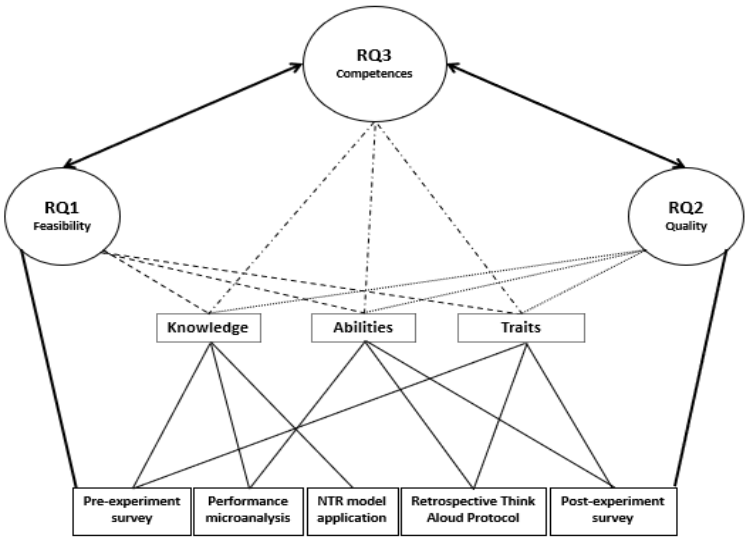
As each variable can be measured by different indicators and via different methods, triangulation of quantitative and/or qualitative methods is needed for a rounded understanding of this practice. This framework enables the integration of multivariate techniques from different disciplines in a dynamic and flexible environment that can be expanded as the project develops, i.e. to encompass further factors, indicators and data collection methods.

3.1. Methods Tested in the Pilot Project

The pilot project brought together the methodological tools illustrated in Figure 6.

Figure 6.

Methods Applied in the SMART Pilot Project



A selected range of indicators was measured using quantitative and qualitative methods, presented in relation to whether they were implemented before, during or after each IRSP task. Consent from all participants was obtained prior to data collection.

Pre-performance data via a survey collected information on:

1. Demographics;
2. Education, including a breakdown of the training hours received in interpreting (consecutive/dialogue and simultaneous), subtitling and intralingual respeaking;
3. Language proficiency in participants' working language(s) under the Common European Framework of Reference for Languages;[8]
4. Self-perception of own competence in interpreting, respeaking and subtitling;
5. Familiarity with and use of intra- and interlingual subtitles from a user perspective;
6. Performance expectations in the intra- and inter-lingual respeaking tasks.

Responses from [1], [2], [5] contribute to participants' profiling; [3] provides an indicator of existing language skills, which can be correlated with actual IRSP performance; [4] provides an indicator of

[8] Basic user A1–A2, independent user B1–B2; proficient user C1–C2 (available at https://rm.coe.int/1680459f97).

metacognition; [6] provides subjective information about expected outcome, to be correlated with performance and post-experiment self-assessment.

Performance data: participants' IRSP performances were captured via screencast technology to enable moment-by-moment access to the process and scrutiny of challenges in IRSP. A grid was developed to enable (qualitative) performance microanalysis adjacent to the application of the NTR model, for a quantitative assessment of accuracy (§3.2.3).

Post-performance data: include retrospective think-aloud protocol (rTAP) sessions and a post-experiment survey. rTAP requires participants to verbalise what went on in their minds just after completing each IRSP task, thus providing insights into decision-making, metacognitive awareness and interpersonal factors leading to a given choice, irrespective of whether it was successful or not. Participants could watch the recordings of their own performances as they saw fit. Their rTAP comments were transcribed orthographically in the analysis grids used to collect performance data (§3.2.3) and translated into English by the authors.

The post-experiment survey required participants to self-rate different elements of difficulty in relation to the task (Figure 13) and the perception of own performance.

3.2. Experimental Set-up and Procedure

25 participants were recruited for the pilot from three different sites (Universities of Surrey and Roehampton in the UK, and UNINT in Italy). They were all postgraduate students with Italian as their mother tongue and a training background ranging from interpreting to subtitling and/or intralingual respeaking, or a combination of all these (§3.2.2 presents their profiles and qualifications in depth). All participants took part in a face-to-face 8-hour crash course which included a brief theoretical introduction about the differences between intra-/inter-lingual respeaking and similarities with cognate disciplines, and practice in intra- (Italian–Italian) and interlingual respeaking (English into Italian) (§3.2.1). One of the goals was to test IRSP feasibility after minimal training. The length of training imparted before testing is important in view of one of SMART's main goals, i.e. designing a training course for language professionals' upskilling.

A Mono-LS configuration was adopted, with students working from individual workstations equipped with one laptop and one headset/microphone (Figure 7). To minimise potential disruption across datasets collected in different sites, all participants used the same laptops equipped with Dragon NaturallySpeaking (v14), i.e. a proprietary and speaker-dependent SR software, and Screen-O-Matic, i.e. a screen recording software.

Figure 7.

IRSP Set-ups in the Experimental Sites

3.2.1. Tasks and Materials

In the first half of the course participants created voice profiles in Dragon and learned the basics of the SR software. The intralingual training phase was essential for those with no respeaking experience and a useful refresher for those with some experience. A dictation exercise was followed by three intralingual respeaking activities. The second half focused on IRSP practice with three videorecorded speeches. Opportunities for debriefing and self-reflection were provided after each task (§3.1).

The videos for all respeaking tasks were selected from the SCIC Speech Repository of the European Commission,[9] a speech bank of interpreter training materials. The materials selected for the pilot were classified as suitable for beginner or intermediate interpreters. Table 1 provides an overview of the main features of the two speeches selected for the IRSP tasks.

[9] Available at https://webgate.ec.europa.eu/sr/

Table 1.

Experimental Speeches for IRSP

Title	Level	Duration	Number of words	Words per minute (wpm)	Lexical density
Gender inequality	beginner	8 mins 5 secs	1,083	134	46.3%
Mobile phones	intermediate	9 mins 26 secs	1,576	167	46.6%

In the SI literature, a delivery rate of 100–130 wpm is considered comfortable for interpreting, while 135–180 wpm is considered fast, although these parameters vary depending on speech type and source language (Riccardi, 2015). The two selected speeches are progressively more challenging in terms of speed but have a similar lexical density.

Participants watched each videoclip once just before the respeaking task, as this would not significantly alter the results but would contribute to reducing their anxiety (which could have impaired performance). Participants opened the video in a media player and launched DragonPad (i.e. the window displaying the SR-recognised output); no instructions were given on how to position the two windows on their screens, which resulted in a variety of configurations (Figure 8). Although this was not measured in the pilot, the influence of interface layout on (visual) monitoring and performance deserves to be investigated. Moreover, participants were not required to segment and/or edit their rendition, but merely to produce scrolling text. Some participants, nevertheless, took the initiative to chunk their target text into subtitle-like units (Image a, Figure 8).

Figure 8.

Interface Layouts During Experiments

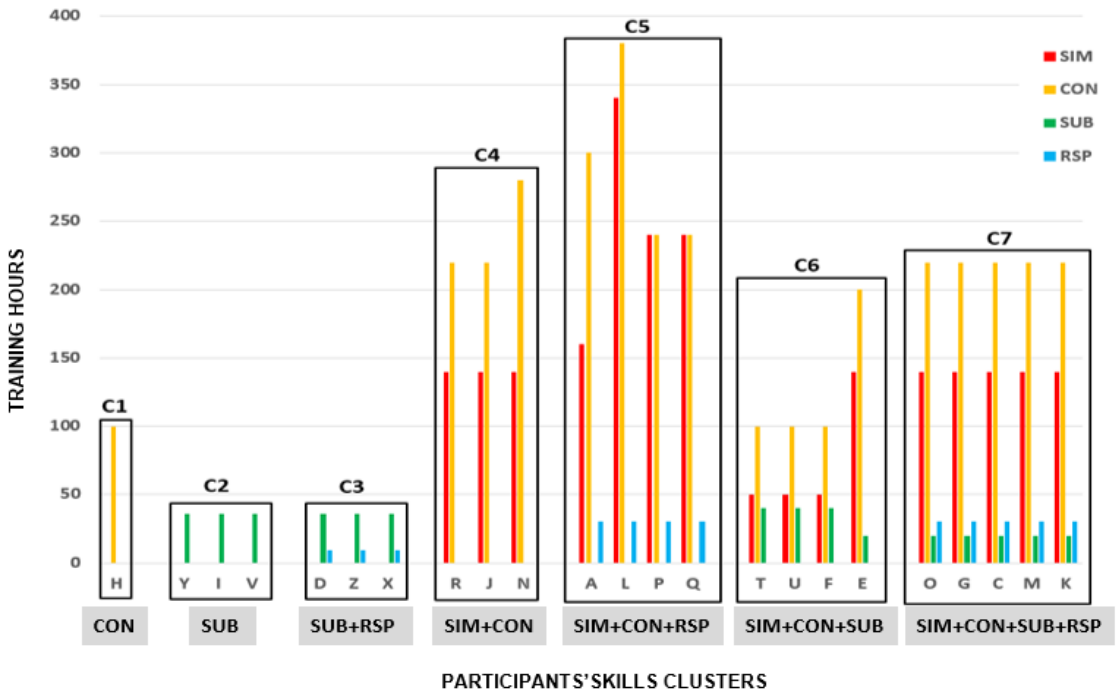
Image a Image b Image c

3.2.2. Participants' Profiles

Of the 25 postgraduate students, 22 were female (vs 3 male) and were an average of 25 years old. Analysis was conducted on 23 participants, owing to technical problems during the performance that invalidated two datasets. Figure 9 provides a visual representation of each participant's training background, showing that no clear-cut profile emerged from our investigation. Their profiles were thus grouped into seven clusters of skills placed in incremental order, from participant H having received training only in consecutive interpreting to participants O, G, C, M, K having received training in simultaneous, consecutive, subtitling and intralingual respeaking.

Figure 9.

Skill Clusters



The large discrepancy in the amount of training hours is explained by (1) the different duration of postgraduate programmes in different countries (one vs two years); (2) the number of language pairs required on the programme; and (3) the optional nature of some interpreting modules. This varied picture reflects the population of language professionals who may be offering IRSP services in the future, as they are likely to come from different walks of life and might therefore benefit from targeted training in the competences they lack.

3.2.3. Procedure for Data Analysis

Data from performance microanalysis, NTR model application and rTAPs were collected in a grid adapted from the Canadian NER score spreadsheet[10] to ensure consistency in the analytical process for different evaluators, to allow for easy comparison of qualitative commments and for automatic calculation of quantitative data.

Figure 10.

Analysis Grid

NTR SCORING																	Speech01: Gender inequality		TAP transcript
NTR deductions	EE	MinT(Cont-OMISS)	MajT(Cont-OMISS)	CritT(Cont-OMISS)	MinT(Cont-ADD)	MajT(Cont-ADD)	CritT(Cont-ADD)	MinT(Cont-SUBS)	MajT(Cont-SUBS)	CritT(Cont-SUBS)	MinT(Form-STYLE)	MajT(Form-STYLE)	MinT(Form-CORR)	MajT(Cont-CORR)	MinR	MajR	CritR	Verbatim transcript	Dragon rendition	@1= TAP1 @2= TAP2
-0.25															1			**if** we want to create a healthy happy and more equal society then this is a challenge we have to face up to	se si vuole una società felice e più bilanciata questa **[94]** una sfida che dobbiamo affrontare. **@1**	**@1:** Allora, un'osservazione che posso fare a video concluso è che il controllo sull'intonazione diminuisce ogni qualvolta la frustazione per la difficoltà di comunicare nei tempi e tradurre nei tempi aumenta. Quindi, più percepivo la difficoltà e più involontariamente, inconsciamente cercavo di compensare con l'intonazione, perché quando si parla in maniera naturale l'intonazione ci aiuta a comunicare il significato. Quindi credo che, ,involontariamente, questo mi abbia spinto a usare maggiormente il tono della voce, nonostante questo non funzioni con Dragon.
	13	25	16	1	0	1	0	5	0	1	11	0	11	0	18	3	1	VERBATIM SCORING		

														TRANSLATION - CONTENT (omission, addition, substitution)			TRANSLATION - FORM (style, correctness)		RECOGNITION		
TOT	EE	OMISS		ADD		SUBS		STYLE	CORR	R	Effective editions			MinT(Cont-errors)	MajT(Cont-errors)	CritT(Cont-errors)	MinT(Form-errors)	MajT(Form-errors)	MinR	MajR	CritR
-30.50	13	15.25		0.5		2.25		2.75	2.75	7	**[81]** *well it means that* not necessary to render (decision in line with previous one to get rid of the question)			**[82]omiss:** *in various ways...*			**[82bis]style** *piuttosto che*				

The verbatim transcript column includes the full transcript of the source speech chunked into independent idea units. The Dragon rendition column features the chunks respoken by participants, aligned with the source chunks; where source segments had been condensed or omitted, cells were merged or left blank, respectively. The greyed-out column accommodates the transcripts from rTAP sessions (see post-performance data in §3.1). The NTR scoring part on the left is devoted to quantitative analysis: each error type is identified by a colour, different severity levels for each error type are identified by different colour gradients (from faded to intense indicating from least to most severe error). The grid automatically calculates the total occurrences for each error type and severity, and the related point deduction, which speeded up the calculation of NTR scores. Finally, the

[10] Website: https://nertrial.com/

qualitative analysis part of the grid was used to provide a succinct comment on the errors. Two raters filled in each participant's grid independently; any rating discrepancy was discussed to come to an agreement.

4. Selected Findings and Avenues for Future Research

This section reports on some key results of our pilot study, namely the NTR scores in relation to participants' training profiles (§4.1) and the most frequent errors (§4.2), illustrated with insights from both quantitative and qualitative methods. Given the relatively small sample and the purpose of the pilot (i.e. testing methods rather than achieving conclusive results), no inferential statistics were used; the results and observations are based on a descriptive approach, and are used as a springboard to identify interesting points for further investigation in the main study.

4.1 . Subtitle Accuracy: NTR Scores

Table 2 presents the NTR scores obtained by our participants in Speech01 (S1) and Speech02 (S2). None of them managed to achieve 98% accuracy in either speech; the mean value was 93.07% on S1 and 91.07% on S2. The difference in mean value between the two speeches can be interpreted either as an empirical indication that S2 was intrinsically more taxing (on account of its higher speed and longer duration; Table 1), or that it was perceived as such owing to fatigue, or both. Correlations with indicators of cognitive load and self-perception comments will be established on a larger sample in the main project to gain a better understanding of the cause(s). Although no participant produced subtitles of acceptable quality by professional standards, the highest scores were encouraging (96.62 on S1 and 95.47 on S2), if we consider that our participants were students who had received only 8 hours of training and attempted three IRSP tasks one after the other in the same session (one for practice and two analysed in the experiment). Three participants (D, K, O) produced the most consistent performances, scoring over 95% in both speeches. The mean values are considerably lowered by a group of 10 students who produced very poor performances (with two outliers, Z and Y, scoring below 90% in both speeches); by contrast, the other 13 (highlighted in bold) form a group of "best performers" who scored considerably above the mean values in both speeches, namely 94.96 on S1 and 93.38 on S2.

Table 2.

Participants' NTR Scores and Mean Values

	S1	S2
A	92.1	84.8
C	**94.51**	**91.10**
D	**95.04**	**95.47**
E	**95.48**	**92.19**
F	**96.62**	**92.56**
G	92.92	91.36
H	92.92	90.51
I	**93.69**	**93.04**
J	90.92	89.68
K	**95.55**	**95.34**
L	**93.74**	**92.43**
M	**95.06**	**93.33**
N	**94.72**	**93.58**
O	**95.63**	**95.15**
P	**96**	**93.90**
Q	92.62	88.96
R	90.59	89.09
T	**95.12**	**94.13**
U	**93.80**	**91.79**
V	90.58	85.63
Z	88.76	85.26
X	92.42	91.98
Y	82.37	83.43
MEAN	**93.07**	**91.07**

If the NTR scores are correlated with the students' training backgrounds, interesting patterns emerge. Table 3 presents the NTR scores grouped by skill cluster (§3.2.2).

Table 3.

NTR Scores by Skill Cluster

Cluster	Composition	Participant	NTR S1 (avg. 93.07)	NTR S2 (avg. 91.07)	Average NTR (92)
C1	CON	H	92.92	90.51	91.70
C2	SUB	Y	82.37	83.43	88.10
		I	**93.69**	**93.04**	
		V	90.58	85.64	
C3	SUB+RSP	**D**	**95.04**	**95.47**	91.50
		Z	88.76	85.26	
		X	92.42	91.98	
C4	SIM+CON	R	90.59	89.09	91.10
		J	90.92	89.68	
		N	**94.27**	**93.58**	
C5	SIM+CON+RSP	A	92.10	84.82	92
		L	**93.74**	**92.43**	
		P	**96**	**93.94**	
		Q	92.62	88.96	
C6	SIM+CON+SUB	**T**	**95.12**	**94.13**	94
		U	**93.80**	**91.79**	
		F	**96.62**	**92.56**	
		E	**95.48**	**92.19**	
C7	SIM+CON+SUB+RSP	**O**	**95.63**	**95.15**	94
		G	92.92	91.36	
		C	**94.51**	**91.10**	
		M	**95.06**	**93.33**	
		K	**95.55**	**95.34**	

The first pattern that can be observed is that 11 of 13 "best performers" (highlighted in bold) received training in both simultaneous and consecutive interpreting. An interpreter training background may confer an advantage in the initial stages of IRSP training; other aspects (such as SL proficiency) may also have played a role, but it is an interesting result in line with findings from previous research on interpreting and respeaking (§2.1). However, as some interpreting students performed below average (A, G, J, R, Q,), an interpreting background per se does not ensure good performance. In addition, 2 "best performers" (I, D) have no interpreting background: I is a "pure" subtitler, while D has been trained in both subtitling and intralingual respeaking. This can be interpreted as evidence of the need to correlate cognitive and (inter)personal variables as key factors in IRSP performance,

as previous training alone does not suffice; these will be investigated in more depth in the larger SMART project.

The second emerging pattern is that most of the "best performers" (9 of 13) have a wider training background; the best combinations seem to be interpreting (simultaneous and consecutive) with subtitling (C6) and interpreting with subtitling and intralingual respeaking (C7). By contrast, a pure subtitling background or a subtitling/respeaking background do not seem to facilitate the acquisition of IRSP skills as much as the other clusters. This can be interpreted as empirical evidence that IRSP is indeed more challenging than interpreting or subtitling and that it mobilises a very composite skillset.

4.2. Distribution of Problems Across Speeches and Clusters

In line with the ILSA results (Dawson and Romero-Fresco, forthcoming), the errors that caused the biggest point deductions from the overall score in our data were omissions, followed by substitutions and misrecognitions (Figure 11, Table 4). Form-related errors (i.e. style and correctness) had a much lower weight in the NTR score calculation, and content-related additions were the least common problem, which points to the difficulty of expanding content while performing IRSP.

Table 1 showed that S2 has a similar lexical density, but is longer and faster than S1. A higher point deduction is therefore entirely predictable, in line with Szarkowska et al.'s (2016, 2017) evidence that a fast speech rate is one of the factors that can drain performance. In our dataset, the deduction score was around 1.5 times higher in S2.

Figure 11.

Main Problems Across Speeches

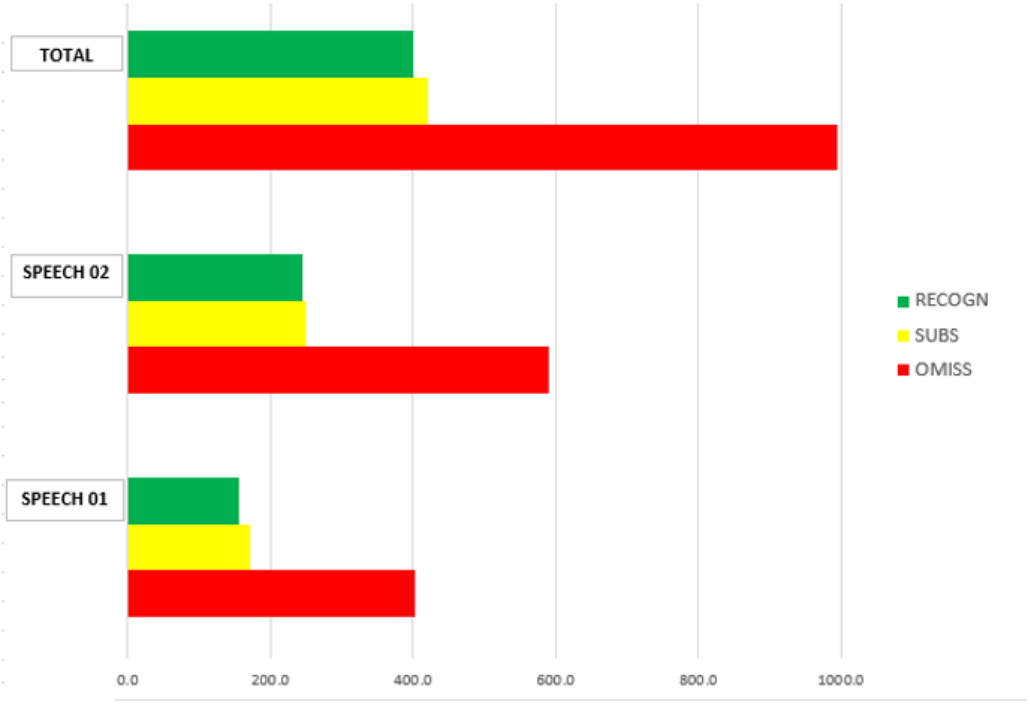


Table 4.

Breakdown of Problems Across Speeches

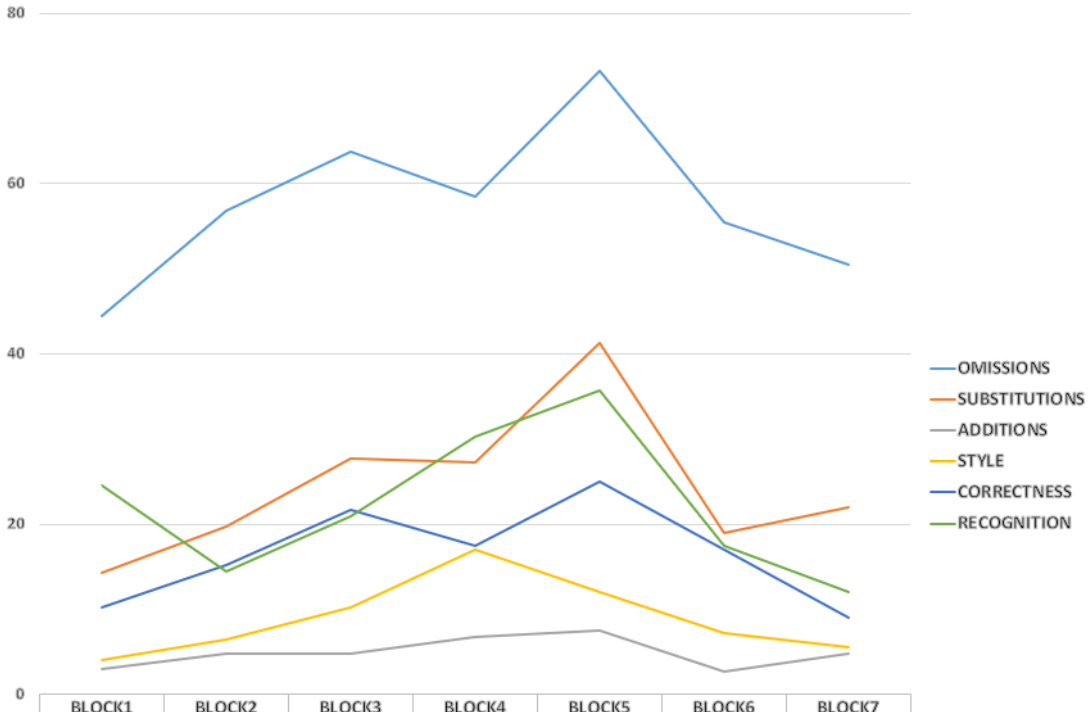
	S1	S2	Total
CONTENT			
Omissions	403.8	590.3	994
Substitutions	171.3	249.8	421
Additions	35	45.3	80.3
FORM			
Correctness	115.8	159.8	275.6
Style	59.8	57.8	117.6
RECOGNITION	155.5	244	399.5

Figure 11, however, does not provide any insights into how problems are distributed *within* each speech, i.e. whether they occur together or in isolation and whether they are evenly distributed or not. Figure 12 shows a possible analytical approach to obtain more granularity, already used to identify the distribution of problems on a timeline in face-to-face vs remote interpreting research (Braun, 2013). The approach is here applied to S1 as a case in point.

As was explained in §3.2.3, the source input was chunked into independent idea units for analytical purposes. Such idea units were then grouped into "idea blocks", i.e. a coherent set of ideas elaborating upon a specific part of the argument. Taking S1 on gender inequalities in the workplace as an example, the semantic boundaries between blocks are marked by questions introducing a new theme (e.g. "Now what do I mean by differences in values?"), markers signalling different aspects of the argument (e.g. "On the first hand it can affect women's career") or markers that the speech is coming to an end (e.g. "Now finally there's a third aspect of gender inequality"). These are key predictors of how the argument will develop and important elements to monitor during performance. S1 was divided into 7 main blocks of similar length, given the well-structured nature of the speech (Figure 12); the occurrence of different problems was then mapped onto each block of ideas. In this particular case, the graph shows that the majority of problems occur approximately one third into S1, when the argument becomes more nuanced and therefore more complex. This snapshot also provides evidence that omissions and substitutions tend to go hand-in-hand throughout S1. In addition, recognition errors initially seem to follow the opposite (downward) trend, but then pick up alongside omissions and substitutions. While the graph reveals some potentially interesting trends, only triangulation with qualitative data (e.g. rTAP, cognitive indicators of overload, stress, self-monitoring) can provide more depth into the cause(s) behind them.

Figure 12.

Distribution of Problems Within S1



	BLOCK1	BLOCK2	BLOCK3	BLOCK4	BLOCK5	BLOCK6	BLOCK7
OMISSIONS	44.5	56.75	63.75	58.5	73.25	55.5	50.5
SUBSTITUTIONS	14.25	19.75	27.75	27.25	41.25	19	22
ADDITIONS	3	4.75	4.75	6.75	7.5	2.75	4.75
STYLE	4	6.5	10.25	17	12	7.25	5.5
CORRECTNESS	10.25	15.25	21.75	17.5	25	17	9
RECOGNITION	24.50	14.5	21	30.25	35.75	17.5	12

Looking at the three main error types identified in relation to skill clusters, the emerging picture is varied. Based on weighted average results, Table 5 shows that C6 and C7 are the only two clusters that seem consistent at producing below-average errors: the background of C6 and C7 participants combines simultaneous interpreting and subtitling, and they were also among the best performing clusters overall (§4.1). Omissions seem to prevail in clusters presenting "single" training backgrounds, such as pure interpreting (C1, C4) and pure subtitling (C2). These were also amongst the worst-performing clusters (<92% NTR), which suggests loss of content as one of the main reasons for poor performance. Furthermore, participants with interpreting in their profile seem to make more substitutions than participants with only subtitling and/or respeaking. However, when combined, subtitling and respeaking seem to support a better performance in relation to the same error types. Single-skilled groups also had the highest deduction scores for recognition errors; notably, though, the group with interpreting and respeaking skills (C5) struggled the most with recognition problems.

Table 5.

Main Problems Across Skill Clusters

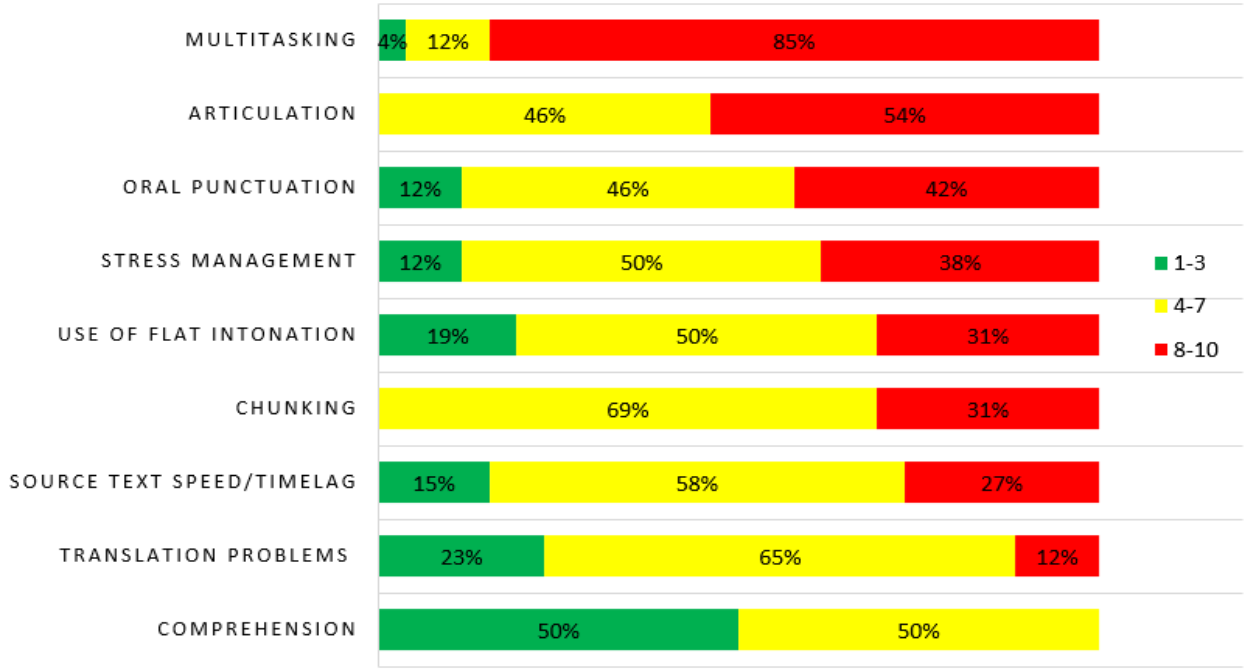
	Omissions	Substitutions	Recognition
C1 (CON)	15.3	5.1	7.9
C2	17.5	4.0	8.4

(SUB)			
C3	14.7	4.5	5.5
(SUB+RSP)			
C4	17.1	7.7	3.7
(SIM+CON)			
C5	14.6	8.8	8.6
(SIM+CON+RSP)			
C6	12.5	5.7	5.0
(SIM+CON+RSP)			
C7	13.7	5.6	3.7
(SIM+CON+SUB+RSP)			
MEAN	15.1	5.9	6.1

The post-experiment survey and rTAPs shed light on the main problem source(s) identified by the students. Figure 13 shows that source speech comprehension was rated as the least problematic factor, followed by translation problems. By contrast, multitasking was rated as the main difficulty during the test (85%), followed by articulation, adding oral punctuation, stress management and use of a flat intonation (§2.1). Thus, technical-methodological and (inter)personal skills are perceived as main issues in IRSP. The next two sections zoom into the two main content-related errors found across the IRSP performances, namely omissions and substitutions.

Figure 13.

Students' Rating of Main Difficulties



4.3. A Focus on Omissions

Omission is a distinguishing feature of subtitling in general (Díaz Cintas and Remael, 2007) and, even more so, of intralingual respeaking (Romero-Fresco, 2011), on account of its real-time translation constraints that make it essential to be able to identify redundant or secondary items in the SL input. The NTR model distinguishes between minor, major and critical omissions, depending on the potential impact of information loss on the audience (§2.2). In this sense, the amount of information conveyed by participants with a similar NTR score and point deduction through omissions can vary if the error distribution across the three omission categories is different. For example, Table 6 shows that participants K, O and D obtained a similar NTR score (just over 95%) and lost a similar amount of points through omissions (between 34 and 39.75). However, although K lost the highest number of points through omissions, these were mostly minor; there were no critical omissions and fewer points were lost through major omissions than D. The latter had a lower overall omission-related point deduction, but a higher number of major omissions and one critical omission. In other words, K's output arguably provided more information than D's subtitles.

Table 6.

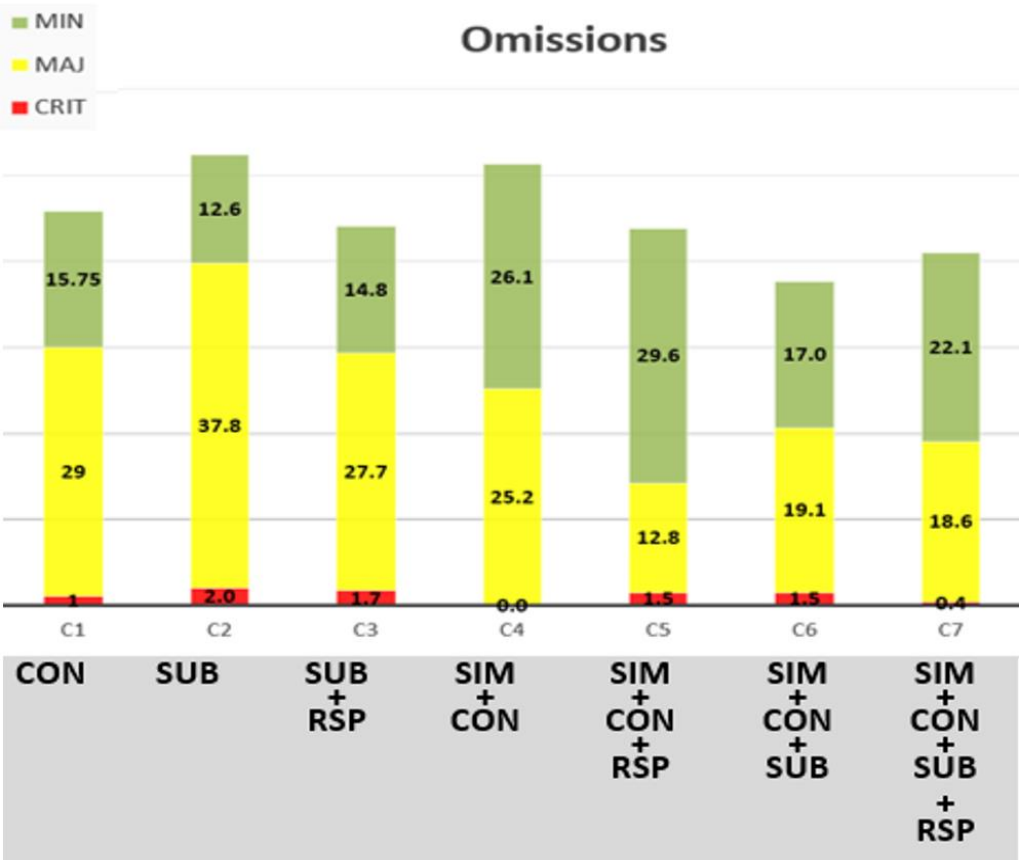
Point Deductions Across Omission Types

Participant	Skill cluster	Average NTR score	Minor omission	Major omission	Critical omission	Point deduction
K	C7 (SIM+CONS+SUB+RSP)	95.44	-24.25	-15.5	0	-39.75
O	C7 (SIM+CONS+SUB+RSP)	95.39	-25	-9.5	-2	-36.5
D	C3 (SIM+CONS)	95.25	-13	-20	-1	-34

Figure 14 shows the relative weight of minor, major and critical omissions in terms of point deductions in the two speeches combined; the data have been arranged by skill clusters. As was shown in §3.2.2, our best performers in terms of NTR scores (C5–C7) are also the clusters with less content loss overall and with the lowest point deduction through major omissions.

Figure 14.

Distribution of Omission Types Across Skill Clusters



All the rTAP comments on omissions were selected and classified by topic, i.e. the self-reported cause of error. Most comments mentioned the inability to deal with multitasking and stress (often associated with time-lag). In turn, stress was often associated with the pressures of split attention; adding oral punctuation and monitoring the output of the SR software were considered taxing tasks which distracted from the translation activity.

The prevalence of rTAP comments on specific phenomena does not necessarily indicate their overall prominence, but merely their higher post-task recall. As it is not possible to provide a comprehensive summary of rTAP comments, some examples were selected from the performances of three participants, namely K, O and D. While O and K have an interpreting background, D was a subtitling student with some intralingual respeaking training and was dealing with real-time interlingual translation for the first time.

Example 1 is an rTAP comment produced by K in relation to a major omission produced at the beginning of S1. In the source text, the speaker asked two questions (why most business executives and politicians are men and why men tend to be paid more than women). K rendered the first one, but omitted the translation of the second question, thus producing a major omission. The comment below reveals that this was intentional and motivated by the need to add punctuation orally.

(1) rTAP comment on a major omission

> Here I was facing difficulties due to the fact of having to interpret simultaneously and think of punctuation... this threw me a little, so I skipped this bit here because I thought its omission would not significantly affect the final meaning of the sentence.

In Example 2, after the introduction, the speaker explained that he would talk about the differences in values and priorities between men and women, followed by a rhetorical question. Participant O did not translate the latter, thus producing a major omission, as the accumulated time-lag was becoming excessive.

(2) rTAP comment on a major omission

> Here I lost the rhetorical question before the speaker's answer, I think because I was lagging behind considerably with my rendition.

Finally, in Example 3, stress and lack of concentration resulted in a critical omission. The speaker was giving reasons for men's greater drive to success and listed several top-level jobs that men often aspire to. Participant D began to translate the concept, but missed the list and left the sentence unfinished, thus producing a critical omission.

(3) rTAP comment on a critical omission

> The items in the list here were missed, owing to the many preambles the speaker made. If he had made a simple list, such as bankers, CEOs, maybe I would have been able to transcribe, dictate all the items. Instead, I got lost, I also froze. Indeed, I even left my sentence incomplete.

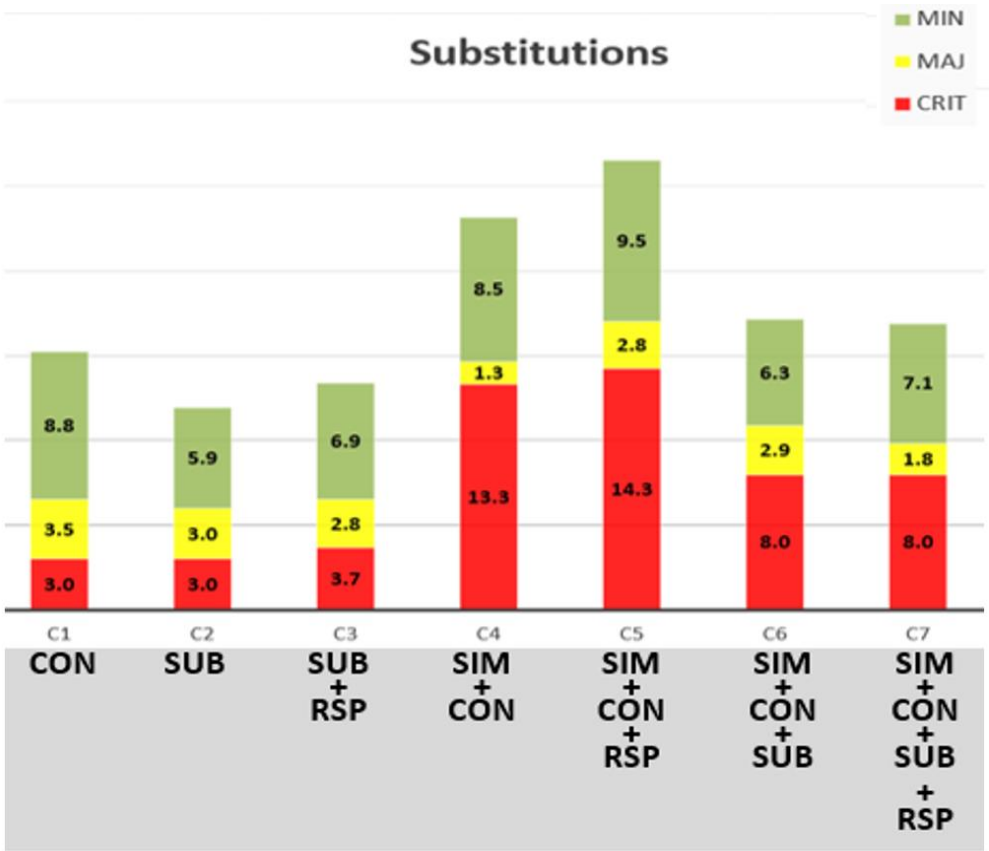
The above examples only provide a snapshot of the data generated by rTAP; in the full-scale SMART project on professionals, rTAP analysis will be triangulated with performance data, participants' profile data and data on cognitive abilities and (inter)personal traits to produce more accurate and nuanced results.

4.4. A Focus on Substitutions

Substitutions are translation-related problems leading to distortion of meaning. Figure 15 shows that the worst performing clusters (C1–C3) produced fewer critical substitutions than best performing ones (C5–C7), but the former have more major omissions than the latter; this is logical, since the less content conveyed, the fewer meaning distortions produced. In terms of skill clusters, the profiles including interpreting plus respeaking and/or subtitling produced the highest number of substitutions; these are also the profiles with the lowest point deduction for major omissions.

Figure 15.

Distribution of Omissions and Substitutions Across Skill Clusters



The rTAP quotes below highlight awareness of the meaning distortions, and provide insights into different reasons for them. In relation to S1, Example 4 from participant O and Example 5 from

participant K (whose rTAP were analysed in relation to omissions in 4.3) highlight lack of time and awareness of software-related constraints (e.g. the need to avoid anglicisms), respectively, as the main reasons for the substitutions.

(4) rTAP comment on a critical substitution

> Here I contradicted myself saying “less ambitious” [instead of “more”] and I realised that, but there was no time to go back so I did not stop all.

(5) rTAP comment on a major substitution

> Here as well I tried to paraphrase “get to the top” to avoid using the English word [in Italian] but my choice has not been very successful. The sentence is misleading.

Unlike K and O, J has a background in interpreting only (simultaneous and consecutive); in Example 6, self-monitoring and multitasking can be inferred as the reasons for meaning distortion.

(6) rTAP comment on a major substitution

> Here I summarised too much... the main difficulty was to focus on listening and production, as well as making sure that the text would continue to appear on screen.

This finding points in the direction of prioritising the training of technical-methodological competences (Pöchhacker and Remael, 2019), particularly multitasking (interacting with the software for optimal recognition through articulation and use of flat intonation and self-monitoring). Stress management, which falls under the broader category of (inter)personal competence (Pöchhacker and Remael, 2019), also ranks high among the difficulty factors in IRSP. This whole dimension is key to IRSP performance and will be tested in the larger SMART project through psychometric and personality measures.

5. Conclusions

The pilot presented in this study proved useful in three main respects. Firstly, to identify a conceptual framework that allows for correlation of variables collected via different tools. Secondly, to test the suitability of some methodological components and analytical procedures that may feed into the final project design; the larger SMART project (started in July 2020) extends the methodology to test and correlate a broader range of factors on a sample of language professionals. Thirdly, to collect data on IRSP from a student population to be compared with existing empirical studies. As ILSA (§2.1) is the only other current project specifically focused on IRSP, we compared, wherever possible, our results with those illustrated in Dawson and Romero-Fresco (forthcoming). The comparison can only be in broad terms, given the different duration of training undertaken by participants (eight hours vs. one month); it comes as no surprise that over 40% of the ILSA participants managed to achieve or exceed 98% accuracy, while none of our participants did. Nevertheless, the NTR scores in our pilot can be considered encouraging, with top scores of 96.62% in S1 and 95.47% in S2; our NTR analysis

provides further confirmation that IRSP is indeed feasible and that that a minimum quality threshold of 98% may be ambitious, but not unattainable. However, to what extent IRSP is feasible and the optimal amount of training remain open questions.

Our participants painted a more varied picture than the ILSA participants, as their profiles included various combinations of interpreting, subtitling and respeaking, which is why their backgrounds were described by skill clusters (Figure 9). In addition, while in ILSA prior training was quantified by course duration (i.e. number of years and/or months), in the SMART pilot it was calculated using actual class hours, to reflect the differences in the three MA programmes included in the study; therefore, it is impossible to compare the skills of the two sets of participants accurately. Despite these differences, common trends emerge in the results of the two studies. Interpreting students were found to have a relative advantage over others in the acquisition of IRSP skills, but some participants with other backgrounds were also found to perform well. In ILSA, the "pure" interpreters were the best performers, followed by the subtitlers with some interpreting experience; our best performers (with NTR scores over the mean values in both speeches) had a training background that included interpreting + subtitling (C6) and interpreting + subtitling + respeaking (C7). These results seem to indicate that, while a training background in interpreting may initially facilitate the acquisition of IRSP skills, it is not sufficient, and a composite skillset can support the learning process more effectively.

As regards error analysis, Dawson and Romero-Fresco (forthcoming) calculated the average number of errors per participant group (e.g. average number of omissions made by the interpreters), whereas the data presented in Figure 11 account for the points lost to each category of error in the calculation of the NTR score. Despite this difference in approach, in both studies the most common errors were omissions and substitutions, followed by recognition errors. In ILSA the clear-cut interpreters produced the lowest number of omissions, substitutions and recognition errors; by contrast, in our study the students who lost the least amount of points had a composite skill-set (C5–C7), while the clear-cut interpreters (C1 and C4) and the clear-cut subtitlers (C2) performed relatively poorly. Once again, our data seem to indicate that students with a mixed training background can cope better with the multiple demands of IRSP, but also that training can be targeted to address its specific challenges. For example, our focus on omissions and substitutions demonstrates that students must be taught to select information strategically, so that when the IRSP task gets too demanding, secondary items can be sacrificed and translation difficulties can be overcome (for example through paraphrasing). In addition, the rTAP comments indicate that in many cases omissions and substitutions were caused by information overload. This is hardly surprising, given the cognitive complexity of IRSP; however, it was interesting to find empirical evidence of the fact that stress management skills affected performance considerably, as several participants indicated that the main sources of difficulty were the pressures of multitasking, including trying to control their time-lag and the need to visually monitor the output of the SR software. Together with the great variability in performance that was found, this seems to indicate that personal traits are likely to play a significant role in IRSP performance.

The larger SMART project will investigate the above issues much more thoroughly on a population of professionals, to address a set of urgent questions for the industry, including how to optimise upskilling to meet the needs of a rapidly growing practice.

Acknowledgements

We would like to thank all the participants for their time, patience, resilience and insights. We are especially grateful to Sabrina Toscani and Aniekan Bassey from the University of Surrey (UK) and Giulia Cremonese from UNINT (Italy), for their invaluable help in rating some of the analysis grids.

References

Braun, S. (2013). Keep your distance? Remote interpreting in legal proceedings. *Interpreting*, *15*(2), 220–228.

Braun, S. (2015). Remote interpreting. In H. Mikkelson & R. Jourdenais (Eds.), *Routledge handbook of interpreting* (pp. 352–367). London: Routledge.

Chmiel, A., Szarkowska, A., Koržinek, D., Ljiewska, A., Dutka, Ł., Brocki, Ł., & Marasek, K. (2017). Ear-voice span and pauses in intra- and interlingual respeaking: An exploratory study into temporal aspects of the respeaking process. *Applied Psycholinguistics*, *38*(5), 1201–1227.

Dawson, H. (2019). Feasibility, quality and assessment of interlingual live subtitling: A pilot study. *Journal of Audiovisual Translation*, *2*(2), 36–56. Retrieved from http://www.jatjournal.org/index.php/jat/article/view/72/24

Dawson, H., & Romero-Fresco, P. (forthcoming). Towards research-informed training in interlingual respeaking: An empirical approach. *The Interpreter and Translator Trainer*.

Díaz-Cintas, J., and Remael, A. (2007). *Audiovisual translation. Subtitling*. Manchester, St. Jerome.

Eichmeyer, D. (2018). Speech-to-text interpreting: Barrier-free access to universities for the hearing impaired. In S. J. Jekat & G. Massey (Eds.), Barrier-free communication: Methods and products. *Proceedings of the 1st Swiss conference on barrier-free communication* (6–15). Winterthur: ZHAW.

Gile, D. (2015). Effort models. In F. Pöchhacker (Ed.), *Routledge encyclopedia of interpreting studies* (135–137). London: Routledge.

Moores, Z. (2020). Fostering access for all through respeaking at live events. *The Journal of Specialised Translation*, *33*, 207–226. Retrieved from https://jostrans.org/issue33/art_moores.php

Ofcom. (2013, 26 July). *The quality of live subtitling* [Statement]. Retrieved from https://www.ofcom.org.uk/consultations-and-statements/category-1/subtitling

Ofcom. (2015a, 13 May). *Ofcom's code on television access services.* Retrieved from https://www.ofcom.org.uk/__data/assets/pdf_file/0016/40273/tv-access-services-2015.pdf

Ofcom. (2015b, 27 November). *Measuring live subtitling quality. Results from the fourth sampling exercise*. Retrieved from https://www.ofcom.org.uk/__data/assets/pdf_file/0011/41114/qos_4th_report.pdf

Pöchhacker, F., & Remael, A. (2019). New efforts? A competence-oriented task analysis of interlingual live subtitling. *Linguistica Antverpiensia, 18*, 130–143. Retrieved from https://lans-tts.uantwerpen.be/index.php/LANS-TTS/article/view/515/471

Riccardi, A. (2015). Speech rate. In F. Pöchhacker (Ed.), *Routledge encyclopedia of interpreting studies*. London: Routledge.

Remael, A., Van Waes, L., & Leijtenet M. (2016). Live subtitling with speech recognition – How to pinpoint the challenges? In D. Abend-David (Ed.), *Media and translation. An interdisciplinary approach* (pp. 120–147). London: Bloomsbury.

Romero-Fresco, P. (2009). More haste less speed: Edited vs. verbatim respeaking. *Vigo International Journal of Applied Linguistics* (VIAL), *6*, 109–133. Retrieved from http://vialjournal.webs.uvigo.es/pdf/Vial-2009-Article6.pdf

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

Romero-Fresco, P., & Martínez, 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* (pp. 28–50). London: Palgrave.

Romero-Fresco, P., & Pöchhacker, F. (2017). Quality assessment in interlingual live subtitling: The NTR Model. *Linguistica Antverpiensia, 16*, 149–167. Retrieved from https://lans-tts.uantwerpen.be/index.php/LANS-TTS/article/view/438

SCIC Speech Repository of the European Commission. Retrieved from https://webgate.ec.europa.eu/sr/

Szarkowska, A., Dutka, Ł., Pilipczuk, O., & Krejtz, K. (2017). Respeaking crisis points. An exploratory study into critical moments in the respeaking process. In M. Deckert (Ed.), *Audiovisual translation – research and use* (179–201). Frankfurt am Main: Peter Lang.

Szarkowska, A., Krejtz, K., Dutka, Ł. & Pilipczuk, O. (2018). Are interpreters better respeakers? *The Interpreter and Translator Trainer, 12*(2), 207–226.

Szarkowska, A., Krejtz, K., Dutka, Ł., & Pilipczuk, O. (2016). Cognitive load in intralingual and interlingual respeaking – a preliminary study. *Poznań Studies in Contemporary Linguistics*, *52*(2), 209–233.

Van Waes, L., Leijten, M., & Remael, A. (2013). Live subtitling with speech recognition. Causes and consequences of text reduction. *Across Languages and Cultures, 14*(1), 15–46.