Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4405–4413
Marseille, 20-25 June 2022

A Semi-Automated Live Interlingual Communication Workflow Featuring Intralingual Respeaking: Evaluation and Benchmarking

Tomasz Korybski, Elena Davitti, Constantin Orăsan, Sabine Braun
University of Surrey, Centre for Translation Studies
{t.korybski, e.davitti, c.orasan, s.braun}@surrey.ac.uk

Abstract

In this paper, we present a semi-automated workflow for live interlingual speech-to-text communication which seeks to reduce the shortcomings of existing ASR systems: a human respeaker works with a speaker-dependent speech recognition software (e.g., Dragon Naturally Speaking) to deliver punctuated same-language output of superior quality than obtained using out-of-the-box automatic speech recognition of the original speech. This is fed into a machine translation engine (the EU's eTranslation) to produce live-caption ready text. We benchmark the quality of the output against the output of best-in-class (human) simultaneous interpreters working with the same source speeches from plenary sessions of the European Parliament. To evaluate the accuracy and facilitate the comparison between the two types of output, we use a tailored annotation approach based on the NTR model (Romero-Fresco and Pöchhacker, 2017). We find that the semi-automated workflow combining intralingual respeaking and machine translation is capable of generating outputs that are similar in terms of accuracy and completeness to the outputs produced in the benchmarking workflow, although the small scale of our experiment requires caution in interpreting this result.

Keywords: interpreting, respeaking, Automatic Speech Recognition, Machine Translation

1. Introduction

Traditionally, live interlingual communication has been achieved only with the help of human interpreters. In recent years, advances in Automatic Speech Recognition (ASR) and Machine Translation (MT) have enabled researchers and industrial actors to explore fully automated speech-to-speech (STS) and speech-to-text (STT) workflows without human input in the delivery of live communication across languages (for example Microsoft's Skype translator). However, although such workflows deliver increasingly promising results when tested on specific datasets (such as call centre data or audiobook data), they encounter challenges when faced with real-life spoken source input characterised by large vocabularies, ambient noise, babble, and articulation variability, which may hamper error-free ASR (El Hannani, 2021).

In this paper, we present a semi-automated workflow for live interlingual STT which seeks to reduce the errors introduced in a STT pipeline by the ASR stage in cascaded systems: a professional human respeaker works with speaker-dependent speech recognition software (in our case Dragon Naturally Speaking[1]) to deliver segmented and punctuated same-language output of superior quality than obtained using out-of-the-box ASR of the original speech. This output is later fed into an MT engine to produce live caption-ready text. Because in our experiments we used speeches from the plenary sessions of the European Parliament (EP), we applied the eTranslation engine, an online MT service provided by the European Commission. We benchmark the quality of the output against the output of best-in-class (human) simultaneous interpreters working with the same source speeches. The evaluation approach is based on the NTR model (Romero-Fresco and Pöchhacker, 2017).

The contributions of this paper are as follows:

- Empirical exploration of an innovative and understudied workflow;
- Presentation of a novel method for the analysis and evaluation of interlingual multimodal data.

The remaining part of this paper is structured as follows: we first present an overview of related research on workflows for real-time interlingual communication and then explain the existing challenges for ASR and MT, leading to the rationale for the present study. We then describe the dataset and workflows explored in our project (MATRIC – Machine Translation and Respeaking in Interlingual Communication), and the evaluation procedure we implemented. The paper ends with a discussion of our results and conclusions.

2. Related Work

In this section we give an overview of related research on semi-automated workflows for real-time interlingual communication and highlight some of the challenges for ASR and MT in establishing fully automated systems. By emphasizing the shortcomings of workflows at the automated end of the spectrum, we strengthen our rationale for considering a semi-automated workflow.

2.1 Current Knowledge on Semi-Automated Workflows for Real-Time Interlingual Communication

The semi-automated workflow we are exploring in this paper has a human respeaker as one of its core building blocks. Respeaking is Respeaking is a technique in which a professional listens to the original sound of a (live) programme or event and respeaks it, i.e. repeats or reformulates it in the same language, including punctuation marks and some specific features for the deaf and hard of hearing audience, to a speech recognition software, which

[1] https://www.nuance.com/en-gb/dragon.html

turns the recognized utterances into subtitles displayed on the screen with the shortest possible delay (based on Romero-Fresco 2011). In our experiment we did not require the features for the hard-of-hearing audiences as our focus was not on accessibility.

As a practice, respeaking shares many elements with simultaneous interpreting (Romero-Fresco 2011, Szarkowska, 2016), especially when performed in a way which is not fully verbatim, whereby the respeaker applies many techniques normally used by simultaneous interpreters. These include paraphrasing, compression, elimination of redundancies and other techniques to ensure accuracy of the output while coping with features typical of spoken language (e.g., pace set by the speaker, impromptu delivery, hesitations, self-repairs, accents).

The reason for experimenting with respeaking as a component of a live interlingual STT workflow is the growing need for the provision of cross-lingual multimodal communication. Live interlingual text output is required in addition to audio interpretation to ensure both accessibility and flexibility of communication in many live events. Respeaking is already the most established method used by major broadcasters, such as the BBC, to produce live intralingual subtitles and meet legal obligations linked to accessibility. It is also increasingly used in live events such as conferences and seminars, and the live subtitling industry reports the application of respeaking and similar workflows cross-lingually thanks to the increased accuracy of the output and the ability to maintain it in a range of different scenarios. Industry actors report the application of respeaking workflows in multilingual events such as town hall meetings or online events, predominantly in out-of-English direction (SMART project expert interviews 2021, unpublished)[2]. The recent technological advances in respeaking workflows, along with anecdotal evidence of industry use cases and a lack of research-based evidence have stimulated our interest in exploring the workflow.

The predominant (speech-recognition based) workflow options available for real-time interlingual STT communication can be placed on a continuum from human-led to semi- and fully automated, as mapped out by Davitti et al. (2020):

(1) Interlingual Respeaking (incl. speaker-dependent speech recognition of target language output);
(2) Simultaneous Interpreting + Intralingual Respeaking;
(3) Simultaneous Interpreting + Automatic Speech Recognition;
(4) Intralingual Respeaking + Machine Translation;
(5) Automatic Speech Recognition + Machine Translation

There is a growing need to validate these workflows empirically to identify the conditions under which they perform best. Workflow (1) has been subject to recent in-depth analysis (e.g. SMART project[3], ILSA project[4]). Romero-Fresco and Baciagalupe (2021) compared workflows (1) to (5). In a small-scale experiment (based on two speeches, two interlingual respeakers, two interpreters and four intralingual respeakers) they used Apptek[5] and Google Translate[6] as their experimental ASR and MT solutions, respectively. Their preliminary results suggest that, in terms of output accuracy, the semi-automated workflow (4) is almost on a par with workflow (2), in which human simultaneous interpreter's output is respoken. They also highlight that workflow (4) appears to be attractive in terms of cost and delay. An important caveat is, however, that the experiment used two speeches delivered impromptu (not read out), with clear articulation and at slow to medium speed, i.e., 100-120 words per minute, which is deemed interpreter-friendly (Seleskovitch, 1978).

Multimodal cross-linguistic communication in workflow (4) with Plain English was also explored by Eugeni (2020) based on speeches of a one-day conference with a similar set of conclusions, claiming that intralingually respoken and machine-translated output can provide sufficient accuracy of output at relatively low service costs.

Fantinuoli and Prandi (2021), in turn, examined fully automated STS, i.e. workflow (5), and sought to propose a communicative and user-centric method of evaluation of its output, benchmarking it against human performance. They show a better performance by human interpreters in terms of intelligibility and a slightly better performance by the fully automated workflow with MT in terms of accuracy.

The MATRIC project we report here focuses on workflow (4) and attempts to compare it with a more traditional (non-automated) practice, namely simultaneous interpreting. Furthermore, we evaluate a more realistic scenario than Romero-Fresco and Baciagalupe (2021), by using authentic speeches from the European Parliament, which have the potential of bringing to the surface more challenges than easier and slower speeches used by the two researchers.

At this stage we are not adding a fully automated workflow to the comparison as preliminary attempts at using fully automated speech-to-text with our source speeches produced output burdened with many more linguistic challenges than in the case of the two workflows under comparison (see 2.2.3). However, we recognise the need to add that evaluation to the set as a follow-up of this study.

2.2 Current Challenges for ASR and MT, and the Rationale for the Study

2.2.1 Challenges for ASR

Despite significant advances in ASR fuelled by involvement of technology giants such as Google or Microsoft, recent research suggests that 'even state-of-the-art speech recognition systems, some of which deliver impressive benchmark results, struggle to generalize across use cases' (Aksënova et al., 2021). Furthermore, to

[2] SMART (*Shaping Multilingual Access through Respeaking Technology*, ES/T002530/1, Economic and Social Research Council UK, 2020–2022).

[3] https://smartproject.surrey.ac.uk

[4] ILSA (*Interlingual Live Subtitling for Access*, 2017-1-ES01-KA203-037948, Erasmus+, European Commission, 2017-2020). http://ka2-ilsa.webs.uvigo.es/

[5] https://www.apptek.com/

[6] https://translate.google.com/

improve ASR output there is a need to collect and annotate more audio data which better represents the current wide spectrum of ASR applications (Szymański et al., 2020). Another challenge for ASR systems is noise robustness: in real-world applications, 'noise source and characteristics may change rapidly' (Sharma and Atkins, 2014: 232), and work is ongoing to improve ASR performance through consideration of generalised classes of noise rather than individual types (ibid.) What is more, ASR systems are trained on 'standard voices', whereas real-life speakers often have accents that do not match the training data, which leads to a lower level of accuracy.

2.2.2 Challenges for MT

MT, the second component of the semi-automated workflow we are exploring, is also facing some challenges despite a very dynamic growth following the transition from Statistical Machine Translation (SMT) systems to Neural Machine Translation (NMT) systems. Among the challenges that remain, Zhang and Zong (2020) list multimodal NMT and emphasize the difficulties posed by simultaneous speech translation as well as the need to disambiguate verbal messages through access to images.

Another issue reported by the authors is balancing quality and latency, which is also a crucial question for real-life application of the workflow explored here. Additionally, the concept of *document-level* MT (docMT) and context-awareness (e.g. Lopes et al., 2020) is also relevant for MT of live spoken output, as overly local MT of fragments and lack of consistency can disrupt communication not just for readers of documents, but also for readers of interlingual live captions. This then becomes the challenge of MT tailored to the needs of a specific real-time communication event (e.g. a conference or a lecture), where consistency needs to be ensured through both pre-event MT training and throughout the event, via the MT's capacity to learn live from the linguistic data as the event rolls out Koehn and Knowles (2017) report further challenges for NMT that are relevant for the context of this paper: poor out-of-domain performance, problems with rare word translation and long sentence translation, issues resulting from sub-word level translation, and failure to copy certain words into the target text (such as proper names or numbers).

In addition, the output of ASR is not always a grammatical sentence due to hesitations and self-corrections of the speakers. In such cases, the quality of MT is lower than in the case of grammatical sentences. While many of these problems are likely to remain in the proposed workflow, some issues (such as sentence length) can be at least partially tackled by human intervention (intralingual respeaking) in the initial phase of our experimental workflow.

2.2.3 Rationale for this study

The challenges for ASR listed above were confirmed by our preliminary tests, performed in late 2020, with leading ASR systems (including Google's Speech-To-Text service and Microsoft's Speech-To-Text, part of the Azure suite) on the speeches we later used in the experiment. The output included misrecognitions, omissions, punctuation issues and substitutions. If such output were to be used as input for MT (as in (5)), all the errors would then be propagated, thus revealing the shortcomings of existing cascaded automated speech-to-text workflows. Importantly, even recently introduced promising end-to-end STT solutions still deliver output with a Word Error Rate (WER, a commonly used performance metric in ASR and MT) in the range of, at best, 7-8% (Park et al., 2019). In practical terms, this means still too many errors to facilitate live communication to the standard expected of professional interpreters. We therefore concluded that there was a good rationale for experimenting with a human-in-the-loop workflow involving professional respeakers, predominantly seeking to reduce ASR challenges, and combining the intralingual respeaking step with MT to produce a multilingual output. The details of the workflow and the data are explained in the following section.

3. The MATRIC Dataset and Workflows

In this section we provide more information on the data we used in the experiment and the two workflows we investigated: the experimental workflow and the benchmark workflow. The experimental workflow is semi-automated (see section 2.1, workflow (4)). As mentioned in 2.1, it involves a professional human respeaker who works with Dragon Naturally Speaking speaker-dependent speech recognition software to deliver a punctuated transcript of the English source speech (the original speech). This transcript is later fed into a MT engine (the European Commission's eTranslation).

The benchmark workflow is simultaneous interpreting, the most human-centric workflow in STS communication, which is commonly used for interlingual communication at live events. As indicated above, our intention was to work with authentic data for which a human-generated benchmark is available. We used the transcripts of interpretations delivered by EU accredited interpreters who work for the EP for benchmarking purposes.

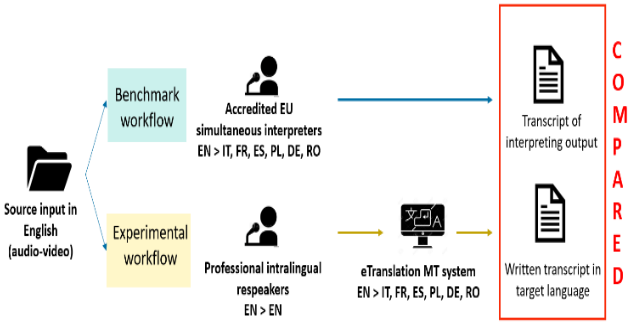

Benchmark workflow
Accredited EU simultaneous interpreters EN > IT, FR, ES, PL, DE, RO
Transcript of interpreting output
Source input in English (audio-video)
Experimental workflow
Professional intralingual respeakers EN > EN
eTranslation MT system EN > IT, FR, ES, PL, DE, RO
Written transcript in target language
COMPARED

Figure 1. MATRIC workflow diagram

Our dataset consists of EU speeches, namely recordings from authentic EP events, interpreted from English into six target languages (IT, DE, ES, PL, RO, FR). The reason for using this pool of data was the fact that it offers a possibility to compare two outputs: the transcribed speech output from accredited EU interpreters (made available by the EP's multimedia resources website) and our experimental semi-automated workflow output, also in text format.

Our first step in the methodological design was the careful selection of the source speeches. For the purposes of our experiment, we settled on very specific source data: monologic, prepared, and scripted speeches in English with low redundancy, as we anticipated they would present fewer dysfluencies than impromptu and multiparty interactions (such as discussion panels, etc). EP interventions tend to be delivered at a fast pace, and fragments may be read out rather than improvised. This mode of speech delivery is also frequent in other conference and media setting, so it is justified to investigate it.

We selected three sample speeches of a duration ranging between 2 minutes and a half and just over 11 minutes, which is representative of speech duration during EP plenary sessions and committee meetings. The speakers were both native and non-native speakers of English with different accents. All in all, we strived to select speeches that represented the variety and difficulty level of speeches delivered at the EP. Table 1 below presents detailed information on the speeches.

	Duration (sec)	Speed (wpm)	Features
Speech 1	2m15s	155	Native speaker (Irish accent), impromptu, fast intervention during a plenary session. Topic: gender pay gap.
Speech 2	4m29s	133	Non-native (Slavic accent). Partially read out, medium pace. Topic: bushfires in Australia.
Speech 3	11m30s	140	Non-native (Greek accent). Mostly read out, with some improvised fragments. Topic: EU health reform.

Table 1. Information on the experiment’s speeches.

Each speech was carefully transcribed and verified by two proficient speakers of English. We did not use the official transcripts available on the EP’s website as these can be heavily edited before publication, while our experiment required detailed transcripts of the speeches as they were delivered. For example, EP editors tend to eliminate any redundancies and change sentence structure, merging short sentences into longer ones, and taking liberties with punctuation to produce a reader-friendly text. Consequently, we needed to produce our own transcripts which are the text output for comparison from the benchmark workflow.

For the experimental semi-automated workflow, we asked professional respeakers to respeak the three English source speeches in English. For each speech, this was preceded by a short warm-up respeaking session featuring the same speaker on a different topic. We recruited four professionals working full time as respeakers for media outlets, who used speaker-dependent and industry-standard speech recognition software (Dragon Naturally Speaking), specifically trained to recognise the respeaker’s voice. Importantly, the respeakers were asked to treat the recording sessions as a regular job whose output is ready for broadcast in the form of captions and were informed about the nature of our experiment (i.e., about the complete workflow where the respeaker’s output is fed into a machine translation engine). We then selected the best respoken output set out of the four respeaker outputs using the NER model (Romero-Fresco, 2011), an established method for evaluating the accuracy of live subtitles produced through intralingual respeaking in media or live event broadcasts. The model’s three letters represent the total number of words in the live subtitles (N), edition errors (E) and recognition errors (R). The percentage of accurate content in the subtitles is determined as shown in Fig. 2 below.

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

Figure 2. The NER model (Romero-Fresco and Martinez 2015)

All four respeakers in our experiment produced output which met the industry-accepted accuracy criteria (NER at 98% or above). The output of the best performer (at 99.2%) was then used as input for the eTranslation MT system. The output from eTranslation constituted our set of experimental texts for comparison against the benchmark workflow output.

To create the benchmark workflow dataset, we asked transcribers (native speakers of the target languages) to transcribe and punctuate the (oral) interpreted output to facilitate later comparison. Each speech was transcribed by two persons who compared their transcriptions and resolved any discrepancies.

Upon completion of this phase, we acquired all texts for empirical analysis. The transcripts of the English source speeches were aligned with the corresponding outputs from the experimental workflow (MT output) and the benchmark workflow (transcribed human interpretations) to facilitate NTR scoring and accuracy evaluation (see section 4.2).

The comparison and evaluation of the text outputs from the different workflows was one of the project’s main challenges, as it involved comparing a written output (from the experimental workflow) with an originally spoken output (from the benchmark workflow). The latter was transcribed for comparative purposes, but in its original form includes the use of intonation and other features of spoken language that could not be captured in writing, and that interpreters traditionally rely on to convey meaning. Considering the many differences between spoken and written language, and the complexity of any comprehensive comparisons, we decided to focus on accuracy and completeness of the message (content) conveyed by the different text outputs (transcripts of interpreter outputs and MT of the best respeaker’s output).

To control for unintended differences in task performance we carefully selected the participants and components in the workflows: we used data from ‘best-in-class’ professional interpreters (EU-accredited professionals) and professional respeakers.

Furthermore, methodological rigour was ensured by our evaluation model, which builds on the NTR model (discussed in 4.2) to capture key error types impacting the accuracy and completeness of the message delivered.

Other factors that contributed to the experiment’s methodological rigour included working as a cross-disciplinary team with expertise in interpreting, respeaking, translation and computational linguistics as well as special training for the evaluators involved in the project, all of whom had a linguistic background in translation, interpreting and/or respeaking.

4. MATRIC Data Evaluation

For the purpose of the evaluation, all data were collected in a bespoke spreadsheet to ensure consistency for the evaluators involved in the analytical process and to facilitate comparison of the results. The analysis grid used was adapted from the NER score spreadsheet used by Canadian media companies for the evaluation of intralingual respeaking data (Davitti and Sandrelli, 2020). The adapted version is attached to this paper in Appendix 1.

4.1 The MATRIC Scoring Sheet

The grid is divided into three main parts: NTR scoring, aligned source and target columns, and verbatim scoring – as shown in Table 2 below.

NTR SCORING	SPEECH (aligned)		VERBATIM SCORING
Error deductions (different error types)	Source segment	Target segment	Qualitative notes with examples and comments on the type of errors and strategies used.
Minor omission error (-0.25), major substitution error (-0.50)	And I want this to be something that will I hope you will have a chance to discuss this afternoon	E questo pomeriggio ne discuteremo. (back-translation into EN: *And this afternoon we will discuss it.)*	[O] I want this to be something that will I hope ; [S] you will have = (noi) discuteremo

Table 2. The main parts of the scoring sheet featuring an example (EN-IT, simplified view)

In the following sections we discuss the quantitative and qualitative parts of the scoring sheet (NTR Scoring and Verbatim Scoring, respectively), as well as the alignment of data, which is represented in the central part of the scoring sheet.

4.2 The NTR Model

The NTR model (Romero-Fresco and Pöchhacker, 2017) was introduced to tackle the challenge of accuracy assessment in interlingual respeaking. It builds on the NER model (see Section 3) for live/real-time subtitle quality evaluation, first introduced by Romero-Fresco (2011) and later expanded by Romero-Fresco and Martínez (2015) to include degrees of error severity alongside the categories of edition and recognition errors.

In the NTR model, edition errors (E) are replaced with translation errors (T) to reflect the language transfer process. The content category includes omission, addition, and substitution errors, while the form category comprises correctness and style. The errors can come in three levels of severity: minor (0.25 penalty point), major (0.5 penalty point), and critical (1 penalty point).

Although the NTR is an error-based model, it leaves scope for capturing the so-called Effective Editions, i.e., successful interventions by the interpreter/respeaker leading to rephrased or lexically changed utterances that still communicate the full meaning of the source utterance/s. Fig. 3 presents the NTR formula and its components.

$$\text{NTR} = \frac{N - T - R}{N} \times 100 = \%$$

N: Number of words

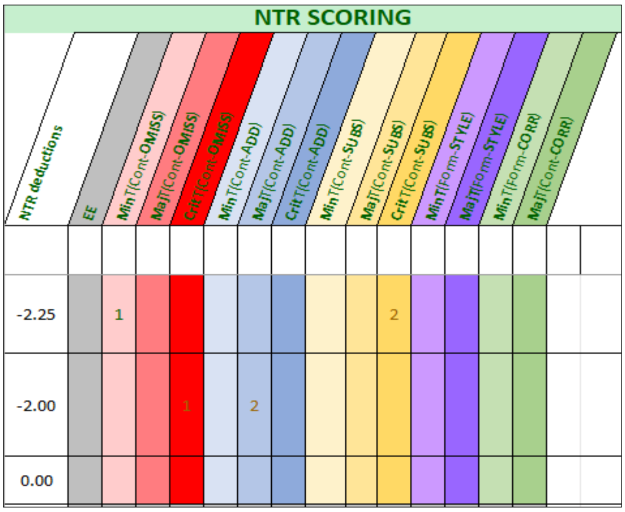
T: Translation errors (content errors: addition, omission, substitutions; form errors: correctness and style)

R: Recognition errors

EE: Effective Editions (a successful intervention by the interpreter/respeaker, resulting in a rephrased or lexically altered utterance that still successfully conveys the meaning of the original)

Figure 3. The NTR model (based on Romero-Fresco and Pöchhacker 2017 :163)

Like the NER, the NTR shows the weight of error deductions in relation to output length. Importantly, the 'full' NTR formula also includes recognition errors (R). Although our data makes it possible to capture recognition errors in respeakers' output, we did not consider this category in our experiment. In our semi-automated workflow, R-type errors only apply to the intralingual respeaking process, which is an interim stage in the process. They do not apply to the simultaneous interpreting (benchmark) workflow, which is why we considered the final output without R-type errors. Figure 4 below shows a close-up view of the NTR scoring part of the scoring sheet, featuring all error types and Effective Editions.

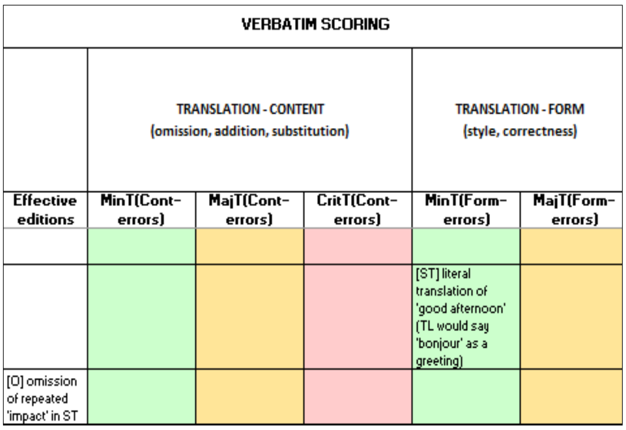


NTR SCORING

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)
-2.25		1								2				
-2.00				1		2								
0.00														

Figure 4. Close-up view of the NTR Scoring part of the scoring sheet, with abbreviated error labels

NTR scoring is facilitated by the more qualitatively oriented Verbatim Scoring part of the scoring sheet which provides space for evaluators to include comments and examples if necessary. A close-up view of this part of the scoring sheet is reproduced in Figure 5.



VERBATIM SCORING					
	TRANSLATION - CONTENT (omission, addition, substitution)			TRANSLATION - FORM (style, correctness)	
Effective editions	MinT(Cont-errors)	MajT(Cont-errors)	CritT(Cont-errors)	MinT(Form-errors)	MajT(Form-errors)
				[ST] literal translation of 'good afternoon' (TL would say 'bonjour' as a greeting)	
[O] omission of repeated 'impact' in ST					

Figure 5: Verbatim Scoring part of the NTR scoring sheet

4.3 Bilingual Data Alignment

A crucial phase in the preparation of the data for the evaluation was the alignment of the source speech transcripts with the output from the two workflows. To facilitate this, the source speech transcripts were first segmented into units of meaning, each presenting a fully formed idea, usually encapsulated within one utterance. In the NTR spreadsheet, each cell in the source speech column represents one unit of meaning. In cases where there was no one-to-one correspondence between a source segment and a target segment, multiple target segments could represent the content of one source segment, or vice versa. An example of this is shown in Table 3. As human interpreters employ a range of strategies when relaying the original message in the target language, automatic alignment will not produce satisfactory results, making manual alignment preferable. For example, a human interpreter will often intentionally eliminate redundant elements at word and sentence level and/or merge several units of meaning to produce concise output. We therefore opted for manual alignment. We also logged the interpreting strategies and/or errors that led to challenges in the alignment process, with a view to further research into automatic alignment of large datasets. Similar challenges occurred in the alignment of the respeaker-based MT output, although to a lesser degree, because respeakers normally work at sentence level, mirroring the structure of the original speech.

Source (EN)	Target (ES)
The pandemic has highlighted many areas which we need to strengthen.	La pandemia nos ha demostrado que es así. (*The pandemic has shown us that this is the case.)*
Among them I would stand that there are resilience to crisis and our mechanisms for prevention and preparedness, our supply weaknesses, the needs to strengthen it these in particular in the areas of pharmaceuticals and the supply chain management	La resiliencia a la crisis es importante. (*Resilience to crisis is important.)*
	Es importante estar preparados de antemano y reforzar el sector farmacéutico y la cadena de abastecimiento. *(It is important to be prepared in advance and strengthen the pharmaceutical sector and the supply chain.)*

Table 3. Example of alignment of source speech units of meaning with (transcribed) interpreter output; back-translation into EN in italics

The example shows several strategies employed by interpreters including condensation, omission, addition, reordering of information. The NTR scoring sheet we used makes it possible to capture such interventions as well as verbatim comments of the evaluators, supporting a realistic assessment of the informativeness and accuracy of the target output.

4.4 The Evaluation Procedure

The evaluators compared the source and target segments, and subsequently scored the accuracy and completeness of the meaning transfer for each segment in the relevant part of the spreadsheet, taking into account both content- and form-related errors. To ensure evaluation consistency, the research team organised a training session on NTR evaluation, which preceded the evaluation and featured common challenges and practical examples of scoring. During the training, the workflows to be assessed were explained to the evaluators. Each output from the different workflows was assessed and scored using the NTR scoring sheet by two evaluators who performed the analysis. The evaluators were aware of which workflow they were assessing. Each evaluator assessed six sets of aligned source and target data (three source-target comparisons with interpreter output, and three source-target comparisons with MT output). Four target languages were evaluated to date, each involving two evaluators. Discrepancies between evaluators were discussed with a view to finding an agreement. Subsequently the evaluators delivered one set of agreed final scores per language. This mitigated the risk of individual evaluator bias affecting the scores. The evaluation process (including training and alignment) took about 120 person hours in total across the four language pairs.

5. MATRIC Data Analysis

We will now present our findings based on the four language pairs evaluated to date (EN-IT, EN-ES, EN-FR, EN-PL). Bearing in mind the origin of the evaluation approach we used, we computed NTR scores for each of the 24 target language outputs (i.e., the benchmark output for the three source speeches in four languages and the semi-automated experimental output for the same three speeches in four languages). The NTR model shows the weight of error deductions in relation to output length. Table 4 shows the average NTR scores achieved in each language pair across all speeches as well as the NTR score difference between the two workflows analysed.

Languages	**NTR average** across speeches	**NTR average** across speeches	**% Change when using the EXPERIMENTAL workflow**
	Benchmark workflow	**Experimental workflow**	
Spanish	98.3%	**98.9%**	0.6% in favour of EXPERIMENTAL
Italian	**98.9%**	98.5%	0.4% in favour of BENCHMARK
French	**99.5%**	99.2%	0.3% in favour of BENCHMARK
Polish	**98.7%**	98.6%	0.1% in favour of BENCHMARK

Table 4: NTR scores and NTR difference computed across the experiment's outputs

Importantly, all NTR scores are above the 98% threshold which is regarded as an acceptable level of accuracy in intralingual respeaking. There is no validated accuracy threshold for interlingual practices yet to date. The high scores are plausible considering the design we adopted where we opted for best-in-class performers, both in the case of interpreters and respeakers. The scores also suggest comparable levels of accuracy in the messages delivered through both workflows. In order to offer a more detailed comparison of the workflows, we also compiled an NTR score breakdown per each workflow and speech (Table 5).

	SPEECH 1		**SPEECH 2**		**SPEECH 3**	
	BENCHM.	EXPERIM.	BENCHM.	EXPERIM.	BENCHM.	EXPERIM.
Spanish	98.6%	98.8%	98.2%	98.7%	98.2%	99.2%
Italian	98.0%	98.8%	99.3%	98.6%	99.4%	98.2%
French	99.5%	99.0%	99.3%	99.3%	99.7%	99.3%
Polish	99.0%	98.7%	98.7%	98.5%	98.4%	98.7%

Table 5: NTR score breakdown per workflow and speech

Overall, the NTR results are similar for many paired outputs, with the French benchmark and experimental outputs for speech 2 being on a par (99.3%), and the Italian outputs for speech 3 showing the largest discrepancy. In most pairs, the difference is less than one percentage point, which, due to the nature of the NTR formula, may, however, still indicate some stark differences in terms of error type and severity. One notable finding is that the semi-automated experimental output was evaluated as more accurate and complete in 5 of 12 output pairs.

In the case of Speech 3, the semi-automated workflow scored better, by a large margin, in Spanish (1%) and by a much smaller margin in Polish (0.3%). One reason could be that parts of this speech were read out at a fast pace with self-repairs, and the speaker was not a native speaker of English. These features are typical of the EP setting but are also prevalent in many other multilingual conference settings where English is chosen by some non-native speakers of English as the language in which they deliver their speech.

However, for the same speech (Speech 3), the human interpreter output was evaluated as better by a large margin in Italian and French. French was also the only language for which the interpreter output scored fewer errors across all three speeches. Based on the data we can furthermore notice that there tends to be more variation in the NTR scores for human interpreter output than across all intralingual respeaking + MT outputs. This may mean that although there is a standardized accreditation process for EU interpreters, individual differences still play an important role in shaping the accuracy of interpreter output. It may also suggest that the variation can be even larger in freelance settings, where interpreters do not undergo a stringent recruitment and testing procedure prior to service delivery.

Furthermore, although the difference in scores across all workflows and speeches does not seem large, it does reflect the differences in the type and severity of errors 'hidden' in the NTR formula. One case in point is the NTR score for French, where an NTR score of 99.3% has been achieved in both workflows. It is therefore crucial to examine the actual error types and their severity. By considering those two elements we can obtain a clearer picture of the impact of the workflow on content completeness and accuracy, which is our focus in this study.

We therefore collected another pool of data that can help evaluate accuracy, namely the total deductions scores for each output. Table 6 below shows that overall, across all language pairs, the point deduction is larger for the outputs from the semi-automated experimental workflow than for the benchmark output, which suggests that the interpreter output has generally been evaluated as more accurate (in majority of the speeches).

	SPEECH 1		**SPEECH 2**		**SPEECH 3**	
	BENCHM.	EXPERIM.	BENCHM.	EXPERIM.	BENCHM.	EXPERIM.
Spanish	-5.00	-4.25	-8.25	-6.00	-20.25	-11.50
Italian	-5.50	-3.75	-4.50	-8.00	-7.75	-24.00
French	-1.75	-3.50	-3.50	-4.25	-4.50	-10.25
Polish	-2.50	-4.00	-5.50	-6.75	-17.25	-15.75
TOT	**-14.75**	**-15.50**	**-21.75**	**-25.00**	**-49.75**	**-61.50**

Table 6. Error deductions per speech and deduction totals

As indicated earlier, it is interesting to explore not only the percentage scores but also the underlying errors. In terms of error types, we found a larger number of (non-strategic) omissions in interpreter outputs. At the same time, interpreter outputs also contained more instances of the positive category of Effective Editions (see section 4.1). A couple of prototypical examples from the data is shown in Table 7 below.

Source (EN)	Target (FR)	EE
Right so dear Chair dear Pascal Honourable members ladies and gentlemen.	Cher Pascal, chers députés, mesdames et Messieurs (*Dear Pascal, dear members, ladies and gentlemen.)*	Omission of oral spoken marker 'right so'
Source (EN)	**Target (ES)**	**EE**
And I believe that both **Farm to Fork** in their own ways and of course **EU for Health** gives us this opportunity	Yo creo que estas dos estrategias nos dan la oportunidad para ello. (*I believe that these two strategies give us the opportunity for it.)*	Proper names replaced and summarised by 'these two strategies' which is clear on the basis of context.

Table 7. Effective Edition examples from the data

In the semi-automated workflow, we found consistently more style and correctness errors, where STYLE errors include literal translations, calques, proper nouns, idiomatic expressions, and CORRECTNESS error include pronouns and gender agreement. Most of these errors were evaluated as minor and were caused by MT, providing an indication of an area of improvement for the MT engine used in this workflow. With better tailoring to spoken text, however, the MT output could possibly further improve.

Rather unsurprisingly, we found a very large difference in the number of Effective Editions in the two outputs, with the benchmark workflow output featuring many more such editions than the experimental semi-automated output. This reflects the fact that effective editing seems to be a strategy that is regularly and very consistently implemented by human interpreters, regardless of the type of text.

Another finding that our evaluation method and the NTR scoring sheet have revealed is the difference in the number of omissions, which appears to be the key error type in relation to accuracy: across all investigated speeches and language pairs, human interpreters in the benchmark workflow produced more omissions than the semi-automated workflow output. As interpreters have been shown to use omissions strategically (e.g., Napier, 2005), it will be crucial for follow-up studies on larger data samples to investigate further not only the total scoring, but also the nature and severity of omissions and their possible impact on the accuracy of the output.

6. Conclusions

This paper has presented a comparison of a semi-automated workflow for live interlingual STT which combines human respeakers and an MT component, with the typical workflow involving human interpreters. The advantage of the semi-automated workflow is that it requires only one language professional, an intralingual respeaker, for the source language (English in our case) and an MT engine to perform the transfer to the target languages. Our evaluation suggests that this workflow is capable of generating outputs that are similar in terms of accuracy and completeness to the outputs produced in the benchmarking workflow. However, this result can currently not be generalized. Our experiment has used source texts from a very specific environment, and the scale of the experiment is small. Further research is needed to explore whether our findings can be replicated on a larger and more diverse body of bilingual data. The variation we find in our small dataset certainly warrants further analysis on a much larger dataset.

A number of errors identified in our data are due to the MT component. Even small changes in the MT component may therefore yield different results for the semi-automated workflow as a whole. It would thus be interesting to compare the outputs of other MT engines or pre-trained MT solutions.

This study focuses on proving the viability of the workflow from the point of view of output content. It does not investigate the question of technical feasibility and latency, which require a separate research initiative featuring a prototyping component. Further research could also seek to find out if the mode of delivery of MT output in this study (offline) has any impact on the accuracy of the output and error types in comparison with the online mode.

In future research, we furthermore plan to investigate the impact of different methods for evaluating multimodal and interlingual data, including, Fantinuoli & Prandi's (2021) approach or using Carroll's scales as revisited by Tiselius based on the notions of intelligibility and informativeness to 'allow for grading of interpreter performance by non-experts in interpreting' (Tiselius, 2009: 95). Another assessment option for comparison is the Direct Assessment method (Graham et al., 2018) or the application of METEOR and BertScore to gain insight in the usefuleness of BLEU-derived methods for the assessment of multimodal data. However, a more immediate task will be to further explore the trends and dependencies in the identified types of error and transfer problems in each of the two workflows to gain a more comprehensive understanding of how they differ and where they converge. Yet another direction for expanding this research is to implement a project with a prototyping component to gauge the latency of the workflow and compare it with the variable latency of human interpreters as well as the latency of a fully automated on-line workflow.

7. Acknowledgements

The Authors would like to acknowledge the expert support from Zoe Moores, who helped with the NER and NTR analyses and computations, as well as the linguists who volunteered to support the project team in data transcription, alignment, and evaluation. The volunteers were (in alphabetical order): Eimee Brown, Michał Górnik, Eloy Rodriguez Gonzalez, Maria Perez Garcia, Megan Stockwell, and Francesco Saina.

The project was implemented at the Centre for Translation Studies, University of Surrey, with financial support from Research England's 'Expanding Excellence in England (E3)' programme (2019-2023).

8. Bibliographical References

Aksënova, A., van Esch, D., Flynn, J., Golik, P. (2021) How Might We Create Better Benchmarks for Speech Recognition? In : Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future, 2021 Association for Computational Linguistics, https://doi.org/10.18653/v1/2021.bppf-1.4 22-34

Davitti, E., Moores, Z., Dawson, H. and L. Fryer (2020). Real-time interlingual speech-to-text via speech recognition: mapping the field of a surging accessibility service. Paper presented at 7th Live Subtitling and Accessibility Symposium, 5-6 November 2020, Universitat Autònoma de Barcelona, https://ddd.uab.cat/pub/poncom/2020/234951/ISLSA_panel2_p4_a2020.pdf

El Hannani, A., Errattahi, R., Salmam, F.Z. et al. (2021). Evaluation of the effectiveness and efficiency of state-of-the-art features and models for automatic speech recognition error detection. Journal of Big Data 8, 5 (2021). https://doi.org/10.1186/s40537-020-00391-w

Eugeni, C. (2020). Human-Computer Interaction in Diamesic Translation. Multilingual Live Subtitling. In: Translation Studies and Information Technology - New Pathways for Researchers, Teachers and Professionals / Daniel Dejica, Carlo Eugeni, Anca Dejica-Carţiş (eds.), pp. 19-31. Timişoara: Editura Politechnica

Fantinuoli C., Prandi B. (2021). Towards the evaluation of simultaneous speech translation from a communicative perspective. arxiv.org/abs/2103.08364

Graham Y., Awad G., Smeaton A. (2018). Evaluation of automatic video captioning using direct assessment. PLoS ONE 13(9): e0202789. https://doi.org/10.1371/journal.pone.0202789

Koehn, P., Knowles, R. (2017). Six Challenges for Neural Machine Translation, https://arxiv.org/abs/1706.03872

Lopes, A.V., Farajian. A., Bawden, R. Zhang, M., Martins, A.F. (2020) Dcument-level Neural MT: A Systematic Comparison, 2020.eamt-1.24.pdf (aclanthology.org)

Napier, J. (2005) Interpreting omissions. A new perspective. Interpreting Vol. 6:2 (2004), pp. 117-142

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint. https://arXiv.org/1904.08779

Romero-Fresco, P., Pöchhacker, F. (2017). Quality assessment in interlingual live subtitling: The NTR Model. Linguistica Antverpiensia, pp. 149-167. Doi 10.52034/lanstts.v16i0.438

Romero Fresco, P. (2011). Subtitling through Speech Recognition: Respeaking. London : St Jerome.

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

Romero-Fresco, P., Baciagalupe, L. (2021). Testing the efficiency of different live (sub)titling methods: a practical experience. Paper presented at 7th IATIS Conference, Barcelona, September 2021

Seleskovitch, D. (1978). Interpreting for international conferences. Washington, D. C.: Pen&Booth.

Sharma, D., Atkins, J. (2014). Automatic speech recognition systems: Challenges and recent implementation trends. In: International Journal of Signal and Imaging Systems Engineering (7), pp. 220-234.

Szarkowska, A., Krejtz, K., Dutka, Ł., Pilipczuk, O. (2018). Are interpreters better respeakers? In: The Interpreter and Translator Trainer, 12:2, pp. 207-226

Szymański, P., Żelasko, P., Morzy, M., Szymczak, A., Żyła-Hoppe, M., Banaszczak, J., Augustyniak, J., Mizgajski, L., Carmiel, J., Yishay, J. (2020). WER we are and WER we think we are. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 3290-3295.

Tiselius, E. (2009). Revisiting Carroll's scales. In: Angelelli, A., Jacobson, H. (eds.).Testing and Assessment in Translation and Interpreting Studies. ATA Monograph Series, pp. 95-1221. Amsterdam: Benjamins.

Zhang, J., Zong, Ch. (2020), Neural Machine Translation: Challenges, Progress and Future. arXiv:2004.05809v1

Appendix 1:

The MATRIC analysis grid adapted from the Canadian NER score spreadsheet used for the evaluation of intralingual respeaking data (Davitti and Sandrelli 2020): https://bit.ly/3I47FH4 .