International Journal of Language, Translation and Intercultural Communication

Vol 11 (2026)

The Sustainability of Interpreting as a Profession in the Era of Artificial Intelligence (early access; full release by the end of February 2026)

# Experimenting with Machine Interpreting in the PL-EN Language Pair

*Tomasz Korybski, Wojciech Figiel, Małgorzata Tryuk, Michał Górnik*

doi: 10.12681/ijltic.43557

# Experimenting with Machine Interpreting in the PL-EN Language Pair: Are We (Getting) Close to "Human-Like" Quality?

**Tomasz Korybski**
University of Warsaw
t.korybski@uw.edu.pl
(corresponding author)
**Wojciech Figiel**
University of Warsaw
w.figiel@uw.edu.pl
**Małgorzata Tryuk**
University of Warsaw
m.tryuk@uw.edu.pl
**Michał Górnik**
University of Warsaw
w.gornik@uw.edu.pl

## Abstract

*Recent industry claims that machine interpreting (MI) is nearing human-level quality are still largely unsupported by ecologically valid empirical evidence. To address this gap, an experiment conducted in June 2025 at the Institute of Applied Linguistics, University of Warsaw compared human simultaneous interpreting with two leading MI service providers in the Polish–English language pair*[1]*. The study simulated a real-life conference setting in which a tandem of EU-accredited interpreters, two MA-level interpreting students, and the two MI systems worked in parallel during a live event comprising a Polish introductory speech, a 40-minute English lecture, and a bidirectional Q&A session. Subjective perception data were collected via a non-controlled online survey completed by eleven student observers, while recordings and transcripts were used for a detailed error analysis. This paper reports on the latter, using a simplified error-based approach adapted from Barik's typology and subsequent frameworks such as NER and NTR (Romero-Fresco and Pöchhacker 2017), including their adaptations for interpreting research (Davitti and Sandrelli 2020; Korybski and Davitti 2024). Error counts reveal a clear quality hierarchy: accredited interpreters outperformed students, who in turn outperformed MI systems. Moreover, MI outputs contained a higher proportion of major, meaning-distorting errors, including ASR-related misrecognitions, overly literal translations, redundant punctuation voicing, random language switches, gender bias, and failures linked to limited contextual memory. Overall, despite technological progress, MI performance in this setting remained far from human-like as of mid-2025.*

[1] *This study received ethics approval from the University of Warsaw's Rector's Committee for the Ethics of Research Involving Human Participants, application no.427/2025, granted on May 28th, 2025*

***Keywords***: *interpreting quality, comparison of Machine Interpreting (MI) and Simultaneous Interpreting; error analysis; Speech-to-Speech translation*

## 1 Introduction

This paper is based on an experiment conducted at the University of Warsaw's Institute of Applied Linguistics in June 2025. It is for a reason that we start this article with a clear reference to the time of implementation and data collection: recent advances in speech recognition and speech-to-speech translation have led to rapidly evolving Machine Interpreting (MI) platforms, whose performance can change within months as models and deployment pipelines are updated. In this context, we emphasize that any comparison between the performance of human simultaneous interpreters and MI performance must be explicitly time-stamped to remain interpretable and fair. The experiment reported in this paper and our findings should therefore be read as a snapshot of human–machine performance at that particular technological moment, in the particular language pair (Polish and English).

When designing the experiment, we had two guiding research questions in mind:

1. In a "mock" setting that has all the characteristics of a real-life event, do existing MI solutions offer a service that can compare in terms of overall communicative value (including accuracy) with human interpretation?

2. In a "mock" setting that has all the characteristics of a real-life event, is there a difference in the perception and audience attitudes between the MI products and human interpretations?

This paper focuses on Question 1 and only signals the main findings related to Question 2, which will be explored in a separate paper (and a follow-up controlled study).

## 2 Comparing quality across human and (partially) automated interpreting outputs

Research on interpreting quality has long demonstrated that evaluation must be multidimensional, encompassing not only accuracy, but also fitness for purpose and user reception. Quality-assurance frameworks in conference interpreting highlight process and product dimensions, including reliability, consistency and communicative effectiveness (Kalina, 2005; Alonso-Bacigalupe, 2023). Functional perspectives stress that outputs must be fit for purpose and may vary depending on user groups and particular communication settings (Kurz, 1993, Kopczyński, 1994), a claim later echoed when technological advances needed to be added to the picture in light of specific use scenarios and risk profiles (Bowker, 2019). Consequently, more recent work proposes reception-based and multi-agent quality assessment models that foreground audience comprehension, trust and perceived usefulness, especially in technology-mediated and hybrid human–machine workflows (Collados Aís, 2018; Han, 2025; Liang & Lu, 2025). The rapid expansion of remote work and the proliferation of technologically mediated interpreting have intensified long-standing questions about how interpreting quality should be assessed across diverse modalities (Davitti et al., 2025). In remote and highly technologised environments, the interpreter's output is shaped not only by professional skill but also by technical factors such as platform design, bandwidth, audio–video fidelity and the interpreter's technological competence. These conditions diverge significantly from on-site interpreting and can hinder access to crucial cues, thereby affecting performance. At the same time, complex multimodal inputs increase cognitive load; processing simultaneous visual and auditory streams can accelerate fatigue and reduce output quality. Addressing these issues requires an interdisciplinary research approach drawing on cognitive science, human-computer interaction (HCI), psychology and cultural studies, while building on established foundations in interpreting research.

Within this broader quality paradigm, the present paper focuses specifically on accuracy and error behaviour in the compared outputs. Building on unit-based accuracy analysis and error-typology research in simultaneous interpreting, we compare professional human interpreters and contemporary MI platforms in terms of the incidence, distribution and severity of content-related errors (Romero-Fresco and Pöchhacker, 2017, Gieshoff, 2022). By narrowing the lens to accuracy while situating it in a wider fitness-for-purpose and reception framework, the study aims to provide a nuanced, time-stamped account of how far, and in what ways, MI approximates or diverges from human performance, as of June 2025 and in the Polish-English language combination.

## 3 Method

### 3.1 Main Speaker and Session Design

The main speaker in the experiment was a bilingual Polish–English academic with recognized expertise in Japanese history and culture. The speaker was instructed to deliver the lecture in a style consistent with a popular science talk, maintaining a natural rhythm and avoiding a scripted delivery from slides. To enable bi-directionality in the interpreting tasks, intervening audience members asked their questions in Polish during the Q&A session. This design ensured multiple language switches (to mimic real-life scenarios) and meant that English into Polish interpreting was required for the lecture segment and Polish into English interpreting was required for the spontaneous interaction and event introduction. By juxtaposing a monologic prepared lecture with a dialogic Q&A, the experiment created conditions that reflect authentic interpreting challenges.

### 3.2 Interpreters

Two groups of human interpreters participated. The first group comprised two professional interpreters accredited by the European Union institutions, both with extensive experience in simultaneous interpreting in institutional and commercial markets. The second group consisted of final-year student interpreters enrolled in the second year of an MA programme in conference interpreting. All student participants had completed a full, three-semester sequence of training in simultaneous interpreting and volunteered to deliver the assignment. Their participation provided a comparative measure of how less experienced but formally trained interpreters performed relative to both professional practitioners and MI delivered through the two platforms.

### 3.3 Audience

The audience comprised 11 (eleven) MA students of interpreting with prior training in both consecutive and simultaneous modalities. Ten audience members attended on-site, experiencing both booth-based human interpreting and machine interpreting in comparable conditions. One student joined remotely, accessing only the machine interpreting output. During the experiment, audience members completed an online survey via Google Forms. This enabled them to capture their reactions live, indicating the respective sources they were listening to. The survey captured qualitative impressions of intelligibility, accuracy, fluency, and delivery, alongside open-ended comments. The students were also asked to produce lists of adjectives evoked by different interpretations.

### 3.4 Machine Interpreting Providers

Two Machine Interpreting providers participated in the experiment. Before the experiment, we approached four providers who advertised machine interpreting services in the required language pair on their website and, to the best of the Authors' knowledge, were technologically advanced due to their extensive knowledge of the field and previous experience: all of the approached platforms had operated as remote simultaneous interpreting (RSI) platforms before adding automated interpreting to their offer. Two platforms responded positively and expressed their interest in taking part in the experiment. In this paper, following the requests from both platforms not to reveal their trade names, we refer to them as Provider 1 and Provider 2. Provider 1 agreed to provide their services and pre-experiment support for free, whereas Provider 2 charged their going rate for 2 hours of machine interpretation (but offered free-of-charge pre-experiment training and live support on the day of the experiment).

### 3.5 Venue and Procedure

The experiment took place in the interpreting lab of the Institute of Applied Linguistics, University of Warsaw (see fig. 1), at the beginning of June 2025. The lab is equipped with six booths. Each booth has a single Braehler console featuring two condenser microphones. The output sound from booths, together with floor sound from two dynamic microphones, was relayed through Behringer UMC 1820 audio interface. The entire audio output was captured and recorded on a MacBook laptop using Audio Hijack software. This software was also employed to reroute the floor signal to two machine interpreting (MI) platforms (each launched on a separate Web browser, accessing sound via a dedicated virtual audio cable). The configuration had been pre-tested in collaboration with MI providers in May 2025. Each testing session lasted approximately two hours. Notably, apart from ongoing support in the run-up to the experiment, both MI providers offered live support and troubleshooting during the event. A full technical check ensured synchronisation before the experiment commenced. The MI platforms operated continuously with only one brief interruption on one of the platforms most likely caused by a software bug (unrelated to any of the platforms) and audio stream redirection through the main computer microphone. However, the ensuing pause (approximately 120 seconds) did not affect the remaining part of the experimental data in any way, and the paused fragment was disregarded when selecting the samples for analysis.

### 3.6 Data Collection

Data were collected from three sources: the original speech, human interpreting output, and machine interpreting output. Both MI providers also supplied backup audio recordings and transcripts. Human and original speech tracks were transcribed using OpenAI's ASR system, with manual checking by a reviewer prior to analysis. Particular care was taken to ensure that any elements excluded by the ASR (such as repeated words, self-corrections, etc.) were present in the transcripts used for evaluation. When selecting the fragments for analysis, we took care to include both directionalities, so the selected samples were in the PL-EN, EN-PL, and mixed directionality (the Q&A session).

### 3.7 Data Analysis

Two team members with expertise in interpreting evaluation analysed the data using widely-applied Barik's (1971) taxonomy of omissions, additions, and substitutions. This framework has been successfully applied in interpreting and respeaking studies (Romero-Fresco and Pöchhacker, 2017; Davitti, 2019; Korybski et al., 2022, Alonso Bacigalupe 2023). The aim was to identify how many instances of the three categories of errors and form (grammar) errors there were in the data while adopting a relatively simple and quick error spotting approach: the five texts (source, transcripts of two automatic outputs, and transcripts of two human outputs) were first split into four two-column files: source and human output (accredited interpreters, source and human output (novice interpreters), source and machine interpreting output 1, source and machine interpreting output 2. Subsequently, the files were analysed one by one by the two evaluators, including a final reconciliation stage to resolve cases where there was no immediate clarity on error status and/or identification. Following this step, the two files were compiled to create one large five-table file, enabling the research team to access all sample transcripts in one document for final analyses and pattern tracking. Qualitative survey data from the audience were analysed independently by the assigned research team member, focusing on user perceptions of intelligibility, accuracy, fluency, and overall impressions. As indicated earlier, this part of the experiment was not controlled and had been designed as an observation study to offer insight into perceptions of an audience with a linguistic background, with a view to designing a follow-up study focusing solely on reception. This part of the study will be subject to a separate analysis and is not presented or discussed in this paper.

## 4 Findings and Discussion

### 4.1 Error Instances in the Data Set

This section presents error counts and examples with additional explanations. Let us start with a global picture of the accuracy of the analysed samples. We counted error instances across all the three samples taken out of our data. We took a global approach and the error count was not meant to reveal error severity, but rather to afford a global picture of the existing accuracy divergence. However, we did recognize different types of errors and their impact on the output both at micro (sentence) and macro (text) levels. A detailed presentation of the predominant error types follows (in 4.2 below). The counts are presented in Table 1 and Fig. 1 below: Table 1 presents the breakdown of samples including sample sizes and directionalities, while Fig. 1 shows the total breakdown of error instances across the three samples and the four outputs.

Table 1: Error count distribution across the three samples

| Sample No. and size (source word count) | Directionality | Error Count: accredited interpreters | Error count: novice interpreters | Error count: MI Provider 1 | Error count: MI Provider 2 |
|---|---|---|---|---|---|
| 1 / 899 | PL > EN | 5 | 8 | 20 | 19 |
| 2 / 667 | EN > PL | 4 | 14 | 18 | 17 |
| 3 / 507 | PL > EN | 2 | 5 | 15 | 15 |

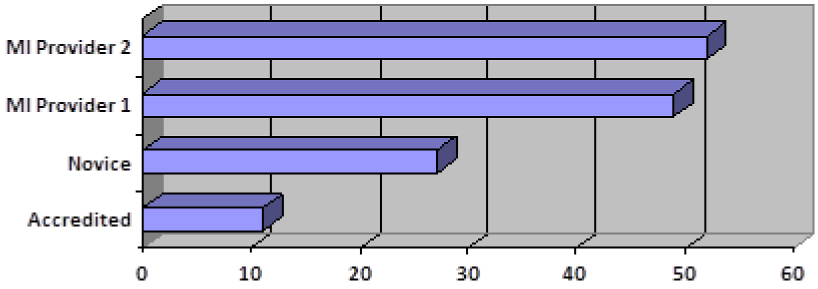

Figure 1: Visualisation of total error counts for all the samples

Across all three samples, the error counts reveal a consistent performance hierarchy, with accredited interpreters producing the fewest errors, followed by novice interpreters, and with both Machine Interpreting providers displaying markedly higher error rates. In the two PL>EN samples, accredited interpreters produced 5 and 2 errors respectively, compared with 8 and 5 for novices, while the machine systems registered substantially higher figures (20 and 19 in Sample 1; 15 and 15 in Sample 3). A similar pattern is evident in the EN>PL sample: accredited interpreters again showed the strongest performance (4 errors), novices produced more than triple that number (14), and the machine outputs contained the most errors (18 and 17). Although variation across language directions is noticeable—novices exhibited a particularly large increase in errors in EN>PL—the relative ranking of the four groups remains stable. The similarity between the two machine systems, whose error counts are consistently close, further suggests a comparable level of maturity or shared underlying technological constraints. Overall, the data indicate that human expertise remains a decisive factor in ensuring accuracy, with accredited interpreters outperforming both novices and machine systems by a considerable margin, and with machine interpreting still demonstrating significantly higher error rates across both directions.

### 4.2 Error Examples and Discussion

When comparing human interpreting output with machine-generated output, simple error counts offer only a partial view of performance. As argued in research on interpreting quality assessment, a meaningful comparison requires qualitative examination of the errors themselves, including their types, recurrence and communicative impact. Pöchhacker (2016) and Romero-Fresco and Pöchhacker (2017) emphasise that accuracy assessment must move beyond numerical tallies to consider how specific errors affect meaning transfer and user comprehension. Similarly, Alonso-Bacigalupe (2023) stresses that severity weighting—distinguishing minor lapses from serious distortions—is essential, as different error types carry different risks for the communicative event. This layered approach is particularly important when contrasting human and machine outputs, since machine systems may exhibit systematic error patterns (e.g., lexical mistranslations or segmentation problems) that differ fundamentally from those typically produced by interpreters. The following section therefore presents the most prevalent and consequential error types observed in the dataset, highlighting patterns that numerical counts alone cannot capture. To offer fair comparison and facilitate comparison of output quality, in the examples below we provide both the transcripts of the source fragment (with backtranslation into English, if required) and transcripts of MI output.

*Example 1*:

| **Source and backtranslation** | **Accredited Interpreters** | **Novice Interpreters** | **MI Provider 1** | **Mi Provider 2** |
|---|---|---|---|---|
| Odbierają Japonce dziecko, wracają do Ameryki, a ona z żalu, z rozpaczy, popełnia samobójstwo. Dramat.<br>BT: *They take away the child from the Japanese woman, they return to America, and she, out of grief, commits suicide. What a tragedy.* | The baby's taken away from the Japanese woman, they go back to America and she, from despair, commits suicide. It's a disaster. | They are getting, they are taking away her child, and out of despair, she commits suicide. | They receive the Japanese, the kids are coming back to America, and she, out of grief, out of desperation, he commits suicide.<br>Drama. | Dot. They pick up the Japanese, child and, and she commits suicide out of grief, out of despair.<br>Drama. |

In this example, the outputs from the two MI providers show a few areas where MI seems to remain prone to errors. First of all, literality of translation: the Polish verb *"odbierać"* may have several different meanings, depending on the context (to perceive, to receive, to claim back, to take away). In human interpreting, picking the right equivalent is dependent on understanding the entire string of sentences – hence the successful choices in both human outputs (although the novice interpreter output is not problem-free as it contains a self-correction and an omission, it still successfully conveys the main message). In the output from MI Provider 1, the verb is wrong, which distorts the message, and the sentence is further affected by the mistranslated "the Japanese" (instead of "the Japanese woman"), forming what would be branded as a "critical error" in the NTR error evaluation method (i.e., an error that introduces a plausible meaning that is entirely different from the meaning conveyed by the source). The word "drama" copied from the Polish original is stylistically questionable, too. A different problem area for MI interpreting is signalled by the output from MI Provider 2: the sentence begins with a punctuation mark, "dot", read out by the synthetic voice (and produced in the tool's live captions, too), which is confusing for the recipient. The object in the output is also wrong, followed by what appears to be quite frequent in MI output: repetition of the linking word "and" for no apparent reason. Another critical error in the output of both MI Provider 1 in this fragment is caused by gender bias: due to failed contextual understanding, the automatic interpretation is "he commits suicide", confusing the listener even further. All in all, the whole fragment is confusing or plain wrong in both MI outputs, whereas the human outputs are either sufficient (novice interpreters) or very good (accredited interpreters) in terms of their informativeness.

Example 2:

| **Source and backtranslation** | **Accredited Interpreters** | **Novice Interpreters** | **MI Provider 1** | **Mi Provider 2** |
|---|---|---|---|---|
| Drugim przykładem, wydaje mi się, równie znanym, chociaż może mniej dramatycznym, jest film *Sayonara* z Marlonem Brando w roli głównej.<br>BT: *A second example that I think is equally well known, although perhaps less dramatic, is the film Sayonara, starring Marlon Brando.* | And the second example, equally known, I think, but maybe less dramatic, is the film entitled Sayonara with Marlon Brando as the main hero. | Another example is maybe a less, maybe less traumatic, Sayonara, a movie with Marlon Brando. | A second example, it seems equally familiar to me, although perhaps less dramatic, is a Sayonara movie starring Marlon Brando. | A second example, it seems to me equally familiar, although perhaps less dramatic, is the movie Saisonara starring Marlone Branda. |

In Example 2, different problem areas emerge in both MI outputs. In MI Provider 1, the indefinite article "a" is wrong, as if the title of the movie had been misunderstood by the MI system and interpreted as an attribute of the word "film". This shows that MI does not (yet?) have the contextual understanding to facilitate an effective transition between two grammatically different languages. The case of PL > EN language pair is a good testing ground for such systems, as Polish does not use definite and indefinite articles and depends on the recipient's contextual interpretation or indicative pronouns to convey the intended meaning. When this is interpreted into English, human interpreters typically do not have to think twice about the choice of articles in English as they are aware of the context, both at micro (sentence) level, and macro (entire text) level. The fact that the analysed MI outputs struggle with articles on a number of occasions proves that for such a system, the process of interpreting is (as of now) far from a natural learning process, and that they do not build their output based on a coherent understanding of the previous fragments. We have found this shortcoming to be one of the predominant sources of confusion in both MI outputs in this study. Furthermore, the output from MI Provider 2 features an example of another problem area, i.e., proper names. Although proper names can be fed into MI systems ahead of the assignment, it is never possible to rule out that speakers will use *other* proper names during their delivery, thus confusing the system and leading to errors like "Saisonara" (wrong spelling of a widely recognized movie title), and "Marlone Branda". The latter seems to be an effect of the system being confused by the Polish original, as in Polish, nouns and many proper names, including foreign ones, are inflected according to case. This "system confusion" led to errors in the output that, depending on the setting, could be considered major or critical, or even interpreted as disrespectful in worst-case scenarios. There is an additional error caused by literal machine translation of the phrase "wydaje mi się, równie znanym", mistranslated by both Providers as "equally familiar to me", whereas in fact the original meaning was "equally familiar to many, I think". This shows a high dependency of the MI systems on the ASR-derived textual layer, which is apparently machine-translated before speech synthesis, with no deeper contextual understanding of the full fragment (in fact, the interjection in question may have been entirely avoided, with minimum detriment to the meaning - as in the output from the novice interpreters).

Example 3:

| **Source and backtranslation** | **Accredited Interpreters** | **Novice Interpreters** | **MI Provider 1** | **Mi Provider 2** |
|---|---|---|---|---|
| First of all, United States claimed that, quote, "America is...", sorry, "The Pacific is our ocean." | Po pierwsze, Stany Zjednoczone twierdziły, że, cytat: „Pacyfik, Ocean Spokojny jest naszym oceanem”.<br>BT: *First of all, the US* | USA uważała, że „Pacyfik to nasz ocean”. Cytując<br>BT: *The US claimed that "The Pacific is our ocean". Quoting* | Jeden<br>BT: *one* | Przede wszystkim Stany Zjednoczone. Proszę przytoczyć ten cytat, Ameryce jest przykro. Proszę minąć Pacyfik, to nasz ocean. |

| | | | | |
|---|---|---|---|---|
| | *claimed that, and I quote: "The Pacific, the Pacific Ocean is our ocean."* | | | BT: *United States first. Please cite this quote, America is sorry. Please circumnavigate the Pacific, it is our ocean.* |

Example 3 features a self-correction of the original speaker in the source transcript. Self-correction and false starts / stuttering at the beginning of utterances are all a natural part of improvised spoken language and can occur even in seasoned public speakers. Human interpreters usually use such minor 'hiccups' to their advantage by slightly extending their decalage locally to produce a final version with no hesitation or false start. That is what the accredited interpreters did in their rendering: the message in the output is complete and even better stylistically than the original. This fragment is therefore an example of what Romero-Fresco and Pöchhacker (2017) branded as an "effective edition". In our samples, effective editions from the two "human" booths outnumber such interventions in the MI output 10 to 1, highlighting the uniqueness of "on the fly editing" which seems to be available to humans only – and widely applied not just by interpreters, but also by respeakers (see Korybski and Davitti, 2024). The fragment turned out to be more challenging for novice interpreters, but they, too, managed to convey the main message from the sentence. In turn, in the output from MI Provider 1 we only have one word – incidentally, the word "one". Such unexpected major omissions of whole sentences or large parts of sentences may happen in MI output, and the obvious consequence is missing information – sometimes crucial for the development of the speech and its comprehension by the audience. In the output from MI Provider 2 we see how confusing a self-correction may turn out to be for an MI system: the cryptic and unintentionally funny (as well as self-contradictory and absurd) fragment "*United Sates first. Please cite this quote, America is sorry. Please circumnavigate the Pacific, it is our ocean"* is both misleading and confusing. The inability of the MI systems investigated in this study to work with the context is a major source of issues in the output.

## Conclusion

The analysis presented in this study demonstrates clear performance differentials between accredited interpreters, novice interpreters and two contemporary machine interpreting (MI) systems. In this paper we have provided a comprehensive description of three examples heavily impacted by errors, presenting the most prevalent error types in MI contrasted with strategic approaches by human interpreters. Across all samples and language directions, human interpreters—particularly accredited professionals—consistently produced fewer and less severe errors than MI systems, confirming that, at the time of implementation of the experiment (June 2025) and in the investigated language pair, human expertise appears to remain essential for ensuring accuracy and communicative reliability. Importantly, the study shows that effective benchmarking of MI performance does not require the full application of complex and time-consuming multidimensional models such as the NTR model; instead, rapid, targeted error analysis based on tried-and-tested taxonomies can yield meaningful insights into the strengths and limitations of MI output when compared with human performance. The examples examined illustrate recurring weaknesses in MI systems, including excessive literality, failure to disambiguate polysemous verbs, inappropriate article selection, difficulties in handling proper names, and an inability to process false starts or self-corrections—issues which often result in distortions of meaning, omissions, or confusing formulations. Many of these shortcomings stem from MI's limited contextual awareness, both at micro- and macro-levels, and from its inability to perform effective on-the-fly editing, a skill that human interpreters (and respeakers) employ naturally and frequently.

Importantly, the patterns identified in the error analysis align with the findings of the observational component of the experiment which will be described separately and which clearly indicates that the question of delivery rhythm, pausing pattern and intonation are vital for MI output to be of acceptable quality for human audiences. There clearly is a gap in MI performance in this are in the language pair in question. Nevertheless, our claims of alignment between our reception data and error analysis results remains preliminary and will be explored further in a follow-up study conducted under controlled conditions and with a more diverse participant pool than the sample of students of applied linguistics we worked with during the experiment described here. Finally, given the dynamic development of MI technologies, the methodological approach outlined in this paper—rapid, iterative, and replicable—should be applied to additional language pairs. Doing so will enable continuous, real-time tracking of MI performance as systems evolve, ensuring that benchmarking remains current, meaningful and responsive to technological change. This is especially important at the moment of writing (end of 2025), with big and medium-size technological companies having announced a number of automated interpreting services as ready for release and deployment in 2026.

## Acknowledgements

This project was funded by the University of Warsaw's IDUB research funding initiative. The authors are very grateful to Karolina Fajfer for her meticulous data alignment and support in analysis, to Agata Balińska for her lecture delivery, and to two experts who advised on the analysis and dissemination process: Professor Agnieszka Chmiel and Professor Karolina Broś.

## References

Alonso-Bacigalupe, L. (2023) Joining forces for quality assessment in simultaneous interpreting: the NTR model, *Sendebar: Revista de la Facultad de Traducción e Interpretación*, 34, pp. 198–216.

Barik, H.C. (1971). A Description of Various Types of Omissions, Additions and Errors of Translation Encountered in Simultaneous Interpretation. *Meta* 16 (4), pp. 199-210. URL https://doi.org/10.7202/001972ar

Barik, H.C. (1975). Simultaneous Interpretation: Qualitative and Linguistic Data. *Language and Speech* 18 (3), 272-297

Bowker, L. (2019) 'Fit-for-purpose translation', in M. O'Hagan (ed.) *The Routledge Handbook of Translation and Technology*. New York: Routledge.

Collados Aís, Á. (2018) 'Quality assessment and intonation in simultaneous interpreting', *MonTI: Monografías de Traducción e Interpretación*, Special Issue 3.

Davitti, E., Sandrelli, A. (2020). Embracing the Complexity: A Pilot Study on Interlingual Respeaking. Journal of Audiovisual Translation 3(2), pp. 103-139. URL https://doi.org/10.47476/jat.v3i2.2020.135

Gieshoff, A.C. (2022) ‘Interpreting accuracy revisited: a refined approach to interpreting performance analysis’, *Perspectives*, 32(2), pp. 210–228.

Han, C. (2025). Quality assessment in multilingual, multimodal, and multiagent translation and interpreting (QAM3 T&I): Proposing a unifying framework for research. *Interpreting and Society: An Interdisciplinary Journal*, *5*(1), 27-55. https://doi.org/10.1177/27523810251322645.

Kalina, S. (2005) ‘Quality assurance for interpreting processes’, *Meta*, 50(2), pp. 768–784.

Korybski, T., Davitti, E., Orasan, C, and Sabine Braun (2022). A Semi-Automated Live Interlingual Communication Workflow Featuring Intralingual Respeaking: Evaluation and Benchmarking. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 4405–4413, Marseille, France. European Language Resources Association.

Kopczyński, A. (1994). Quality in Conference Interpreting: Some Pragmatic Problems. In: Lambert, S., Moser-Mercer, B. (eds) *Bridging the Gap: Empirical Research on Simultaneous Interpretation.* John Benjamins, Amsterdam, pp. 87-99.

Korybski, T., Davitti, E. (2024). Human Agency in Live Subtitling through Respeaking: Towards a Taxonomy of Effective Editing. *Journal of Audiovisual Translation*, *7*(2), 1–22. https://doi.org/10.47476/jat.v7i2.2024.302

Kurz, I

Liang, L., & Lu, S. (2025). The evaluation and reception of the translation quality of three translation modalities in live-streaming contexts: computer-assisted simultaneous interpreting, machine translation (MT) with human revision and raw MT. *The Translator*, 1–19. https://doi.org/10.1080/13556509.2025.2494566Way, A. (2018) ‘Quality expectations of machine translation’, in J. Moorkens et al. (eds) *Translation Quality Assessment*. Cham: Springer International Publishing.

Pöchhacker, F. (2016) *Introducing Interpreting Studies*. 2nd edition. London: Routledge.

Romero-Fresco, P. and Pöchhacker, F. (2017) ‘Quality assessment in interlingual live subtitling: The NTR Model’, *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 15, pp. 149–167.