The Interpreters' Newsletter Special Issue 2025 (30), 37-56
DOI: 10.13137/2421-714X/37795 ISSN 1591-4127 (print) ISSN 2421-714X (online)
https://www.openstarts.units.it/dspace/handle/10077/2119

The digital boothmate in an educational setting: students' experience with SmarTerp

Mariachiara Russo
Ira Torresi
Amalia Amato
University of Bologna - Department of Interpreting and Translation (DIT), Forlì Campus

Abstract

This study investigates the impact of the ASR-AI Computer-Assisted Interpreting (CAI) tool SmarTerp on simultaneous interpreting (SI) performance in an educational setting. Its novelty lies in a multi-measure approach, whereby subjects were tested for CAI-enhanced performance, CAI-induced stress, and tool usability and satisfaction. Following a review of cognitive challenges in SI and problem triggers, we describe the experiment, which involved 24 interpreting students at the University of Bologna, tested with and without SmarTerp support, after specific training with this CAI tool. Performance was assessed at the trigger (e.g. numbers, terms, named entities) and sentence levels. Statistically significant improvements at trigger level with the CAI-tool included increase in correct renditions ($p < 0.0001$) and decrease in omissions ($p=0.002$) and semantic errors ($p < 0.0001$) already from the first test session. The same trend was detected at sentence level. Students reported general satisfaction with the tool's usability and support but emphasised the need for specific training and improvements in tool latency. Although overreliance and occasional distraction were noted, the findings confirm that CAI tools like SmarTerp can enhance SI performance when properly integrated into interpreter training. The study advocates for more empirical research and tailored training to optimise digital support in interpreter education.

Keywords

Simultaneous interpreting, SmarTerp, CAI (computer-assisted interpreting), ASR (automatic speech recognition), digital boothmate, AI (artificial intelligence).

Introduction

Interpreting and technology are an indissoluble duo. It is, indeed, thanks to technology that simultaneous interpreting became a possibility, at first experimentally in the 1920s (Delisle/Woodsworth 1995), and then during the Nuremberg Trials (Baigorri-Jalón 2014). Today, computer-assisted interpreting (CAI) tools support the work of interpreters from preparatory documentation to in-booth assistance (Fantinuoli 2018; Fantinuoli/Prandi 2021; Rodríguez *et al.* 2022a). Integrating the use of these technologies in interpreters' workflow calls for reflections ranging from human-computer interaction to ergonomic design (Frittella 2023), to cognitive load assessment (Defrancq *et al*. 2024), to a new trend in interpreter language production studies (Chmiel 2024). All this requires a new perspective on traditional interpreter education that calls for empirical research, which is only beginning to appear (see Defrancq/Fantinuoli 2021; Rodríguez *et al*. 2022b).

To contribute to filling this gap, this article presents the results of the experimental testing of the ASR-AI CAI tool SmarTerp at the University of Bologna. Its novelty lies in a multi-measure approach as subjects were tested for CAI-enhanced performance in rendering well-known problem triggers in simultaneous interpreting: i.e. numbers, terms, and named entities based on seven evaluative criteria; CAI-induced physiological stress, measured by heart rate and heart-rate variability; and tool usability and satisfaction elicited through questionnaires. We will first offer a literature review on problem triggers and experimental studies with ASR-AI CAI tools (section 1), followed by a methodology section (section 2), results (section 3), their discussion (section 4) and conclusions (section 5).

1. Literature review

Simultaneous interpreting (SI) is a cognitively demanding task (Moser-Mercer *et al.* 1997; Gile 2009). It involves several functions including attention, working memory, long-term memory, language processing, comprehension, language transfer and production. All these functions require information processing and memory resources that are limited (Miller 1956), with different concurrent tasks potentially interfering with each other (Fagot/Pashler 1992; Pashler/Johnston 1998; Seeber 2011). This can be further aggravated by so-called problem triggers, "associated with increased processing capacity requirements which may exceed available capacity or cause attention management problems, or with vulnerability to a momentary lapse of attention of speech segments with certain features" (Gile 2009: 171). According to Gile, triggers can be lexical (proper names, numbers, enumerations, terminology, speaking-related triggers), speaking-related (such as pronunciation and speed of delivery, or overlapping talk), environmental/technical (for instance poor sound quality, background noise, poor visibility of the speaker and slides), and multimodal (like combining both audio and visual channels in simultaneous with text and/or with slides).

Mankauskienė (2016) proposes a different classification of triggers: speaker-related (e.g. accent, non-native speaker, rate of delivery); message-related (including lexical, syntactic, and semiological elements); interpreter-related (such as experience, competence, knowledge, fatigue); and technical (bad equipment, poor sound, noise, etc.). His study confirms that interpreters find numbers and proper names difficult, but technical terms and language-specific lexical items may be problematic as well.

Numbers seem to be the most investigated trigger, and all studies point out the high rate of errors in rendering them. Possible explanations are lack of conceptual representation, unpredictability, lack of redundancy, informativity, and differences in syntax between source and target languages (Braun/Clarici 1996; Mazza 2001; Gile 2009; Pinochi 2009; Timarová 2012; Seeber 2015). Numbers were studied both in isolation and combined with other triggers, for example delivery rate (Korpal/Stachowiak 2013, 2020).

As for multimodal triggers, there seems to be a substantial difference between cases where the audio and visual inputs are complementary and must be both processed to ensure full comprehension of the message, and cases where the two inputs are identical and redundant, which seem to be less cognitively demanding. Lamberger-Felber's (2001) and Lambert's (2004) studies showed that when SI is performed with a written text, the error rate in rendering names and numbers dropped when compared to SI without text. Seeber's (2017) eye-tracking study demonstrated that when given both auditory and visual input for numbers, interpreters tend to rely on the visual input for large numbers, supposedly because they require more processing capacity than small numbers and the visual input appears to be more immediate. Based on this kind of evidence, CAI tools visually repeat aural triggers, with the aim of reducing errors and omissions.

The literature about CAI tool effectiveness starts with Desmet *et al*. (2018), who used a slideshow to simulate CAI tools that did not exist at the time, and found that number accuracy increased from 45-50% in the control group to over 80% in the group that was aided by slides, while errors decreased by 69% (*Ibid.*: 22).

More recent studies have experimented with CAI tools proper, including those that employ automatic speech recognition (ASR) and AI for trigger detection, visualisation and translation. Such innovations are already shaping the market of interpreting aids as we write, as predicted by Fantinuoli (2018, 2023). For the purposes of this paper, we will divide studies on CAI tools conducted in educational settings (section 1.1) and in professional settings (section 1.2).

1.1 Educational settings

In this section, we review some of the main experimental studies which empirically measured students' SI performances while using real-time ASR/AI CAI tools, with reference to problem triggers (named entities, acronyms, specialised terms, and figures).

One of the first studies of this kind (Defrancq/Fantinuoli 2021) tested InterpretBank ASR (developed by Fantinuoli 2017) with six students interpreting from English into Dutch at the University of Ghent. The tool provided a running transcription of the ASR output where numbers were highlighted. Results indicated high tool precision (96%) and an increase of accuracy in number rendition from 67.7% to 90.2%, with omissions decreasing from 15.8% to 3.5%. The study also provides evidence of the psychological reliance on ASR, perceived as "a safety net" (*Ibid.*: 15) and confirms previous positive results by Desmet *et al*. (2018). However, it also raises an interesting point about human-machine interaction: although not all participants used the CAI tool input all the time, the authors noted that "the mere availability of ASR already improved the participants' accuracy, irrespective of whether it was used" (*Ibid.*: 26).

Using the same InterpretBank ASR only to visualise numbers without a full ASR transcription, Pisani/Fantinuoli (2021) compared the performance of 20 interpreting trainees of the University of Trieste with and without the tool. The results showed a significant reduction in errors when using InterpretBank. As to usability, the students reported feeling distracted by the additional visual input, which the authors felt could have been mitigated by specific training. Russello/Carbutto (2023) tested the same tool at UNINT in Rome, where 16 trainees were divided into a study group using InterpretBank ASR for number transcription, and a control group where each trainee was prompted by a peer transcribing the numbers for them. The results indicate a slightly higher error rate with the human (55.6%) *vs* the digital (50%) boothmate. The tool latency was an issue: the mean latency stood at 4.2 seconds (in a 1-18.75 range), while interpreters can only delay their ear-voice-span up to 3 seconds without disrupting their performance (Fantinuoli/Montecchio 2023). Finally, students displayed overreliance on the tool and asked for preliminary training.

Prandi (2023) shifts the focus on another potential trigger in SI: technical terminology. Prandi compared ASR-assisted SI, SI with glossary queries via InterpretBank and SI with a PDF glossary. The highest accuracy was achieved with ASR. The use of InterpretBank led to slightly worse performances, followed by the PDF, but "overall, the terminological accuracy achieved with digital support was quite high, also in the PDF condition" (*Ibid.*: 176).

Rodriguez *et al*. (2022b) conducted a study at the University of La Laguna on the usability of SmarTerp by students. The results were presented at the 3rd HKBU international conference on Interpreting and Technology, but unfortunately never published.

Giudice (2024) tested 11 trainees of the University of Bologna at Forlì interpreting from Spanish into Italian in an inter-subject and intra-subject design, comparing SmarTerp, which only visualised named entities, numbers, specialised terms and acronyms, to Google Meet, which visualised the full ASR transcript. Participants were trained for both tools, and their eye movements recorded with an eye-tracker. Students' renditions were more accurate with SmarTerp, especially in syntactically complex sentences, which shows that the tool leaves greater leeway in source text reformulation. As to the eye-tracking data, students did not look much at the input but when they did, the results were better with SmarTerp. With Google Meet, students looked at the screen much longer before starting their translation, indicating that they relied more on sight translation, which is cognitively less demanding, but more conducive to calques. Students' questionnaires revealed that they preferred SmarTerp because the input was less intrusive, even though it took more time to familiarise with, and its latency was too long.

Finally, Defrancq *et al.* (2024) analysed the cognitive load induced by InterpretBank ASR by measuring the mean Fundamental Frequency (F0) of the voices of the same 6 students of Defrancq/Fantinuoli (2021) and their errors with numerals. ASR seemed to have little effect on the overall quality of the students' performances, as the number of errors with or without the tool was very similar (338 *vs* 348). Based on the results of their previous study and the fact that accuracy appeared to be slightly better, the authors concluded that students benefited from ASR support. As to the mean F0, there seemed to be little evidence of the association between increased cognitive load and the challenge of interpreting numbers, and no evidence of an association between ASR support and increased F0.

In addition to studies carried out in Western educational settings, mention should be made of those appearing in China, where the use of CAI tools was introduced later, but is on the rise. According to Guo *et al*. (2023, 2024), Chinese interpreting research is characterised by coverage of both consecutive interpreting (CI) and SI: the tendency to adopt a more holistic perspective and not just the rendition of problem triggers; a broader evaluation of CAI tools' impact considering the overall performance and coherence of interpreting; and very large samples (e.g. Xiao/Wang 2020, with 177 respondents).

1.2 Professional settings

One of the first experimental studies with professionals was carried out by Frittella (2021a, 2023) who used a mock-up of SmarTerp and testing materials based on her theoretical framework of incremental number processing difficulties (Frittella 2017, 2023). Frittella carried out a pilot test with 5 simultaneous interpreters working from English into Italian. Their performance and usability results contributed to refining a new tool version, which was employed in the main study with 10 simultaneous interpreters who used SmarTerp after specific training. As argued by Frittella (2022a, 2022b), most studies in computer-assisted interpreting focus their quality assessment solely on trigger rendition and do not include semantic aspects, such as how the whole utterance was rendered. The author recommends that the context be taken into account to have better data reliability, since a trigger might be rendered with a paraphrase (e.g. the date 2025 with 'this year') or might be omitted when it is a repetition. Frittella recommends that the trigger be analysed in context, and defines 'context' for numerals as a "numerical information unit" containing: the numeral and referent(s); unit of measurement (for instance currency or size); relative value (increase or decrease); time and geographical reference (Frittella 2022b: 38). This approach had previously been adopted by Korpal and Stachowiak (2018) who found that in SI with no CAI tool there was a positive correlation between number accuracy and context accuracy in interpreters' rendition.

Consistently, in her 2023 study Frittella evaluated participants' deliveries considering all levels in her model, namely: the numeral; the numerical information unit; the textual coherence of numerical information; its extralinguistic plausibility; and functional equivalence to the original message. Predictably, interpreters achieved a higher success rate in triggers of lesser difficulty (isolated acronyms, isolated name entities, numeral and complex referent, list of three unknown terms), with isolated numbers being an unexpected exception (*Ibid*.: 116). As to users' perception, the usability qualities with the highest rating were ease of use and effectiveness, while ease of learning, timeliness and dependability scored significantly lower (*Ibid*.: 122). These results informed further development of SmarTerp concerning user interface design, especially trigger visualisation, and technical specifications, confirming the good practice of tool developers collaborating closely with an interpreting scholar to obtain quality and thorough empirical usability data. In our study we adopted Frittella's methodology and taxonomy of interpreting errors and strategies (section 2).

Defrancq/Pleșca (2024) investigated users' experience and interaction with the ASR/AI tool developed at Ghent University, the Artificial BoothMate (ABM), displaying terms, numbers, and a running transcript. The study involved 47 interpreters of the 24 booths of the

European Parliament, the majority of whom had never experienced an in-booth support before. Only one respondent stated that the ABM improved their performance, while several negative experiences were reported, mainly concerning the extra cognitive load caused by the unusual visual input (e.g. locating the term, jittering transcript). Impressions were more positive concerning the usefulness of the tool for numbers, terms, accents, dense information, fast delivery and as a safety net in case of uncertainty. Terminology support was perceived as fast enough but only sometimes useful. Only half of the respondents reported improved accuracy, and one third of them reported improved speed. Most respondents found that number transcription was often useful, and the running transcript was sometimes useful but difficult to manage due to lack of experience. Finally, most found the ABM reliable but less trustworthy than a human boothmate, and reported a willingness to use it in the future.

One further application of CAI tools appears in Gieshoff *et al.*'s study (2024) which involved nine conference interpreters using augmented reality (AR) glasses rather than on-screen support. Triggers (acronyms, proper names, numerals and technical terminology) appeared as written words, both in the source and target languages, displayed next to the speaker's image. On average, participants used 81.2% of the suggested renditions (*Ibid.*: 295), but all participants except one expressed discomfort in wearing AR glasses, and the majority reported feeling visually hindered.

Finally, the efficacy of using ASR providing complete running transcripts was investigated by two further studies. Li/Chmiel (2024) tested the use of ASR-generated subtitles in SI and measured the cognitive load caused by subtitles through eye-tracking, EEG and self-perceived mental workload. For the 23 professional interpreters involved, when the level of precision in subtitles was 90% or above, accuracy in interpreting grew proportionally. Conversely, with lower ASR precision, interpreting accuracy decreased. Furthermore, the workload decreased with higher ASR precision and increased with lower ASR precision, probably due to the need to solve auditory-visual inconsistencies. In his doctoral thesis, Rodríguez González (2024) investigated the impact of speed and lexical density on ASR-aided SI to assess interpreting performance and user experience. The study was conducted with 16 professional interpreters in four different settings: fast speech without ASR; fast speech with ASR; lexically dense speech without ASR; and lexically dense speech with ASR. In general, participants rated positively the support of ASR in SI, but with high delivery rates the system produced transcription errors "which in turn will delay the interpreter's processing, in particular when there is speed to tackle at the same time" (*Ibid.*: 198).

2. Methodology

In 2021 the EIT Digital-funded Project "SmarTerp – SMARTER INTERPRETING: Seamless Management and Automation of Resources and Tools for an Efficient Remote Simultaneous Interpreting" developed SmarTerp, an ASR and AI-based CAI tool. Its initial release was tested in an interpreter education setting – the DIT of the University of Bologna, a project partner – to explore the impact of its use on trainees' SI performance and to collect participants' perceived usefulness. SmarTerp[1] combines CAI features with

1 SmarTerp features were further integrated into the SmarTerp-Educational platform to remotely train interpreters with state-of-the-art technologies (Arriaga *et al*. 2023).

speech recognition and translation by displaying problem triggers (terms, named entities, numbers, acronyms) alongside their translation on the interpreter's PC screen.

Two hypotheses were formulated. The first was that students' rendition of problem triggers would be more accurate and complete with SmarTerp. The second hypothesis was that students would feel supported by SmarTerp in their SI effort and there would therefore be a reduction in stress. Furthermore, we assessed SmarTerp usability as perceived by students.

An experimental study was carried out following an intra-subject approach to reduce the number of uncontrolled variables. The same group of interpreting students was tested under two different conditions: with SmarTerp (study condition) and without SmarTerp (control condition). We also adopted an inter-subject approach and compared the two groups. The test comprised three interpreting sessions per participant, two of which (sessions 1 and 3, see section 2.2. below) were used for data collection and analysis and were followed by surveys. The results of the performance evaluations (hypothesis one) and students' perceptions will be presented here, whereas for the results of students' level of stress (hypothesis two), readers are referred to Olalla *et al*. (2023).

2.1 Participants

The study recruited 24 volunteers among second-year students of the Master in Interpreting of the Department of Interpreting and Translation (DIT) of the University of Bologna, Forlì campus. All students (21 females, 3 males; mean age 24.8, median 24, min. 22, max 43) were Italian L1 speakers and were divided according to the following language combinations, each comprising 6 students: Italian > Spanish, Spanish > Italian, Italian > English, English > Italian. All the participants signed Informed Consent forms approved by the University of Bologna's Bioethical Committee.

2.2 Procedure

Before the actual test, participants received a link to a set of training materials developed by Frittella (2021b), which they could use asynchronously. Three testing sessions were then held across three weeks.

Session 1 (October 4-9, 2021) started with a warm-up video to practice with SmarTerp: an English training video developed by Frittella (2021b) and its Spanish and Italian versions. After a ten-minute break, participants were asked to interpret simultaneously a video without SmarTerp, followed by another ten-minute break, and another video with SmarTerp. The session was followed by an online usability questionnaire.

Session 2 (October 11-16, 2021) consisted of one video in each language of the project, to be interpreted with SmarTerp. This session was intended only to maintain participants' familiarity with SmarTerp.

Session 3 (October 18-25, 2021) replicated session 1, but without warm-up.

Each student completed each session on the same day. The 12 students who were available for onsite testing were tested in the interpreting booths of the Department equipped with computers and headsets. This allowed the team to collect physiological stress measurements through Empatica wristbands (Olalla *et al.* 2023). The other 12 students participated online via Zoom, at the time that best suited them, to maximise participation.

2.3 Materials

Fifteen videos were produced (5 for each language involved), plus three videos (1 for each language) for the warm-up practice. Italian and Spanish speeches were translated from the English original. All the speeches were video-recorded by native speakers and scripted by the researchers, except for the English training video and one English speech originally created by Frittella for her early testing in July 2021 (Frittella 2021a, 2022a), used in session 1. These two videos served as prototypes for the other speeches' length and duration (mean and median duration=10.3 minutes, at 110 words per minute) and speech features. In particular, speeches were scripted so as to contain instances of all the problem triggers identified by Frittella in her unpublished early testing project report (2021a) (Table 1).

code	*trigger* + definition
AC	One acronym in a *simple sentence* (20- 30 words, with simple syntax and logic and no further problem trigger).
NE	One named entity in a simple sentence.
NU	One complex numeral (3 digits, order of magnitude = *trillion*) in a simple sentence.
NR	One complex numeral (4 digits, order of magnitude = *billion*) and a complex referent (an acronym/ named entity/ specialised term/ a numerical value) in a simple sentence.
NIU	A complex numerical information unit (NIU), approx. 30 words, constituted of: 4. a complex referent 5. a complex unit of measurement (an acronym/ named entity/ specialised term/ a numerical value) 6. five consecutive numerals, as in the following structure: (referent) increased/decreased by (X%) from Y in (time1) to Z in (time2).
NCR	A number-dense speech passage with the following characteristics: 5. constituted of three subsequent NIUs, approx.. 10-15 words each; 6. the time and place references remain unvaried and are repeated in each NIU; 7. the unit of measurement and the referent remain unvaried, but the referent is expressed with a different synonym in each NIU; 8. the numeral changes in each NIU.
NCN	A number-dense speech passage with the following characteristics: 4. constituted of three subsequent NIUs, approx. 10-15 words each; 5. time, place, referent, unit of measurement and numeral change in each NIU; 6. either the referent or the unit of measurement is complex.
TE	A specialised term in a simple sentence.
TL	A list of three specialised terms in a simple sentence.
TS	A series of three terms in a semantically complex sentence (i.e., characterised by implicit logical passages where comprehension requires relevant background knowledge and inference).
SO	Address to the participants with three unknown given names and surnames and two unknown acronyms.
CP	Announcement of the conference programme.

Table 1: Categorisation of problem triggers (Frittella 2021a, unpublished report)

The speeches were adapted from the web or researchers' professional experience. The first session included a welcome address at a High-Level Meeting on Japan's economic recovery after Covid-19 (warm-up video), a welcome address at the Summit on the future of the United Arab Emirates and a welcome address at an African Union Conference. The second session had one speech on the blueberry market. The third session included a speech by the President of the Republic of Uzbekistan at a Summit of the Economic Cooperation Organisation and a welcome address at a ceremony of the Asia Cooperation Dialogue.

Readers were instructed to speak at 110 words per minute and to pause briefly (under 3 seconds) at the end of each concept, and to pause for a full 3 seconds at the end of each paragraph. The scripts incorporated ‘stage directions’ on how to pronounce each acronym (letter by letter) and number.

2.4 Data analysis and statistics

Each student produced 5 performances across the three sessions: three with SmarTerp and two without it. Only students’ performances of sessions 1 and 3 were analysed thus totalling 96 performances. Students’ performances were recorded and transcribed with a speech recognition system developed by the Project partner FBK.[2] These automatic transcriptions were subsequently checked against the corresponding audios by the University of Bologna research team.

Students’ performances were jointly assessed by two pairs of researchers, one for each language combination: EN<>IT (n. = 48) and ES<>IT (n. = 48). The same six analytical categories suggested by Frittella (2022b: 45) were applied, as well as another one suggested by the team (SR):

8. Correct rendition (CR) = the source task is faithfully and completely reproduced in the target speech (TS);
9. Partial rendition (PR) = the source task is reproduced in the TS omitting some relevant items;
10. Minor error/Missing detail (M) = the source task is reproduced in the TS omitting some irrelevant items;
11. Generalisation (G) = the source task is reproduced in the TS either by a hypernym or a condensed sentence;
12. Omission (O) = the source task is not reproduced in the TS;
13. Semantic error (SE) = the source and TS meaning do not match, or the target sentence is incoherent;
14. Self-repair (SR) = the target language contains false starts or unfinished words subsequently completed.

SR was added considering that students could be expected to anticipate the source content and then check the SmarTerp input and possibly correct themselves.

Students’ performances were submitted to a two-tier analysis: at the level of task (the trigger itself) and of sentence (the effect of trigger rendition on the overall ac-

2 <https://www.fbk.eu/it/>.

curacy and coherence of the sentence in which it occurred). We assessed coherence in terms of target text (TT) internal logical and semantic structure, as well as TT content corresponding to ST content. Each pair of researchers attributed one of the above-mentioned categories to each trigger rendition, both at task and sentence level. By way of example, one Italian rendition of the task Numeral + Complex reference (NR), originally in English, was rated as follows:

ST	*TT*	*Analysis*
Shurtan Gas Chemical Complex	-	Task: O
USD 1.812 billion in funding	-	Task: O
Shurtan Gas Chemical Complex has received USD 1.812 billion in funding to expand its polymer production capacity	hanno ricevuto ehm ingenti somme per aumestr- ehm aumentare le proprie capacità *[have received huge sums to increase its capacity]*	Sentence: PR, SR

Table 2: Example of assessment at task and sentence levels

The data of all performances were aggregated, and the count mean of each category was calculated for statistically significant differences. For the levels ‘Task’ and ‘Sentence’, the mean of the values obtained by the 24 students in each of the seven categories, with and without SmarTerp, were tested with the paired t-test when data were normally distributed, or with the Wilcoxon matched-pairs signed-rank test when data were not normally distributed. The Shapiro-Wilk test was used to test normality in the distribution. Statistical analyses were performed with STATA 18 (2024).

3. Results

3.1 Students’ performances

The analysis was initially carried out by keeping language combination groups separate. However, since the analysis of results for all combinations (EN>IT, IT>EN, ES>IT, IT>ES, with 6 participants each) yielded the same patterns in both directionalities, an aggregate analysis of performances will be offered here to provide a larger dataset.

To better appreciate the effect of the CAI tool use over time, a comparison was made between performances with the CAI tool and without it at sessions 1 and 3 (total n. of performances = 48).

3.1.1 Task

The results for the task rendition (trigger level) are summarised in Figure 1.

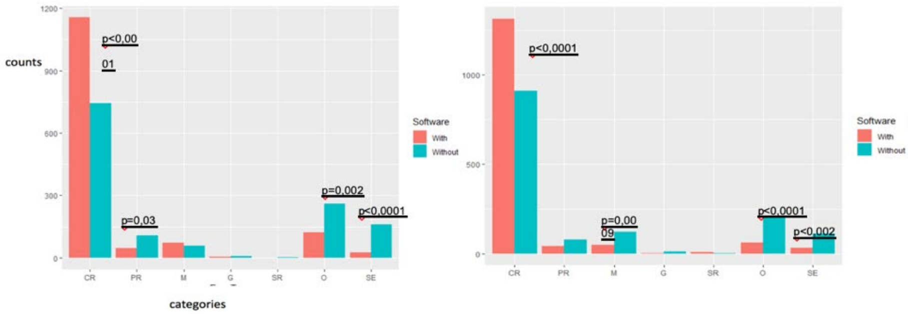

Figure 1: Management of triggers (task level) with *vs* without SmarTerp, and comparison of session 1 (left) *vs* session 3 (right)

In both sessions, comparatively more successful handling of triggers (correct renditions, partial renditions, minor errors) and fewer instances of unsuccessful management (omissions, semantic errors) were found with the CAI tool. The difference between with *vs* without conditions was highly significant for several categories (correct renditions, omissions and semantic errors). The difference is significant also for fewer partial renditions, but only in session 1, and for fewer minor errors only in session 3. The count levels appear to be rather similar for correct renditions between the two sessions, while there is a slightly higher reduction in omissions with SmarTerp in session 3. Generally, a similar pattern was recorded between the two sessions with an improvement in both with *vs* without conditions in session 3 as shown by the decrease in unsuccessful management of triggers (omissions and semantic errors). This result may be explained with general better familiarity with the setting, speech type and CAI tool acquired by the students over time.

3.1.2 Sentence

The same dataset of students' performances (n.= 48) was analysed also at sentence level to check for overall coherence and faithfulness. These were operationalised using the same seven categories. The results for the sentence rendition are summarised in Figure 2.

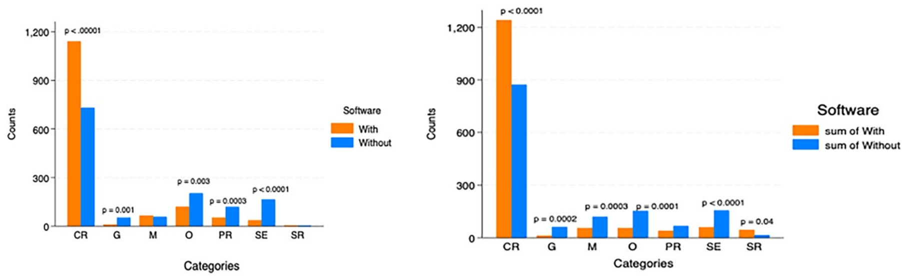

Figure 2: Management of sentences with *vs* without SmarTerp, and comparison of session 1 (left) *vs* session 3 (right).

In both sessions, comparatively more successful processing of the source speech (correct renditions, partial renditions and minor details), and fewer instances of unsuccessful processing (omissions, semantic errors) were obtained with the aid of the CAI tool. The difference between with and without the tool was highly significant for correct renditions, omissions, semantic errors (in both sessions), partial rendition (only in session 1) and minor details (only in session 3). There was also a statistically significant decrease in generalisations with the aid of SmarTerp in both sessions. Finally, slightly more instances of self-repairs were observed with the CAI tool, but the difference was not statistically significant.

3.2 Participants' perceptions

As anticipated, at the end of session 1 and 3, participants were asked to fill in a short online usability questionnaire where they were asked to answer each question or rate each statement using a 7-point Likert scale, with 1 = Strongly dissatisfied/disagree/very unlikely and 7 = Strongly satisfied/agree/very likely. The options on the Likert scale that were not chosen by any participants will not be present in Figures 3-10 illustrating the analysis below. The wording of each of the eight questionnaire items is to be found in the captions.

The vast majority of respondents appeared to be satisfied or very satisfied with their SmarTerp experience (Figure 3). The CAI Tool was found to be user-friendly by the vast majority who strongly agreed with the stated question, with only a slight decline in "strongly agree" responses in session 3 (Figure 4).

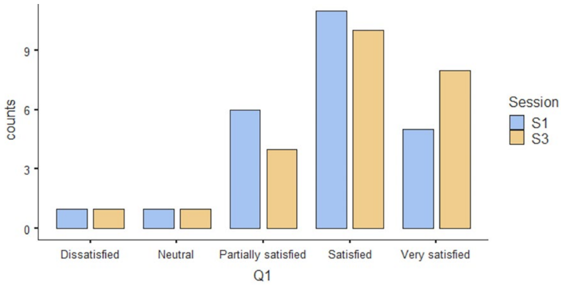

Figure 3: Overall, how satisfied are you with the support of the CAI tool SmarTerp during testing?

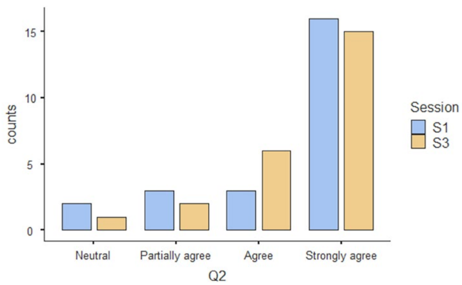

Figure 4: The CAI tool was easy to use

The vast majority of participants also perceived the CAI tool as an effective aid in achieving accuracy of rendition, with a marked increase in the "agree" option in session 3, and only a slight decline in the "strongly agree" option in session 3 (Figure 5).

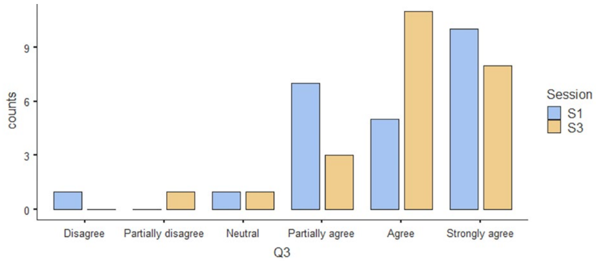

Figure 5: The CAI tool has been helpful in improving accuracy

The range of selected options for the item "The CAI tool has been a source of distraction during my performance" (Figure 6) is more varied, highlighting a marked subjective difference in managing attentive resources. All considered, a slight upward trend seems to appear in session 3, but there is still a full set of replies in the central area of the spectrum, which indicates a certain difficulty in perceiving SmarTerp affordances.

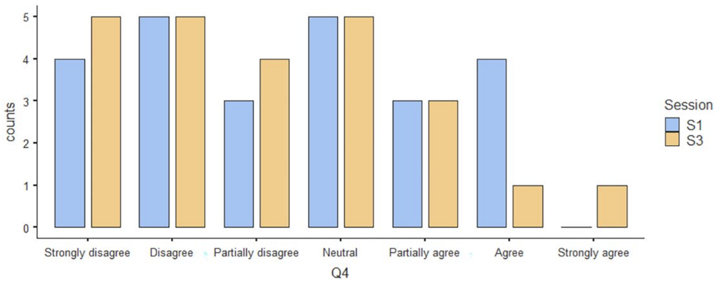

Figure 6: The CAI tool was a source of distraction during my performance

Participants also felt the need for specific training for SmarTerp use to be effective (Figure 7). As to the timeliness of on-screen suggestions, once again, reactions to the stated question were mixed, but the general trend indicates a certain dissatisfaction with the timeliness of SmarTerp (Figure 8).

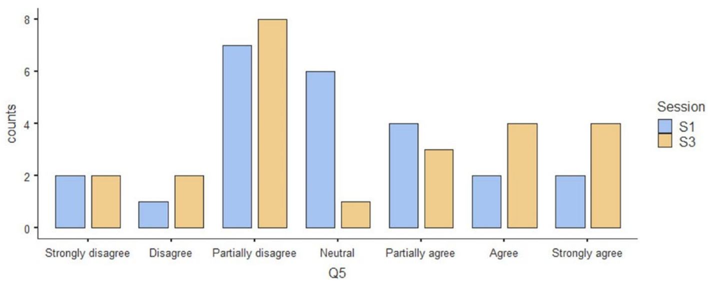

Figure 7: No training is required to use the CAI tool effectively

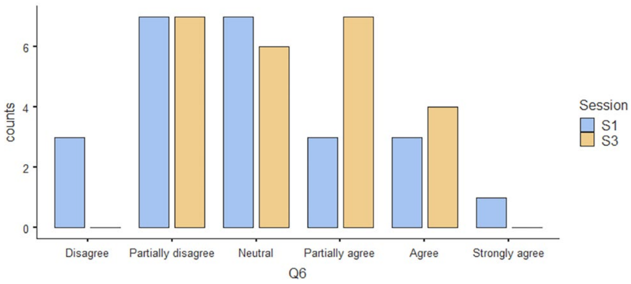

Figure 8: The display suggestions provided by the CAI tool were timely

Overall, however, only few students felt they could not trust the CAI tool (Figure 9). The degree of trust seems to be lower after session 1, but after session 3 too, there was a slight decline in "strongly agree" responses. In replying to the last question (likeliness of using ASR-equipped CAI tools in the future), participants seemed to express a certain degree of confidence that the market will soon provide viable ASR solutions that they will be able to employ in their professional careers (Figure 10).

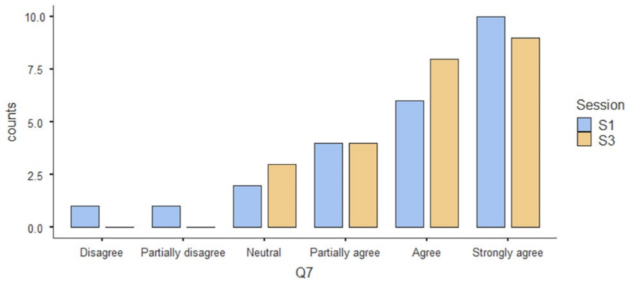

Figure 9: I felt I could trust the support of the CAI tool

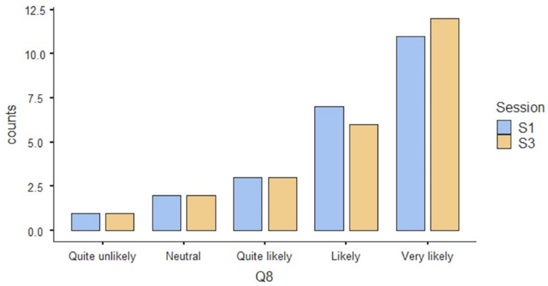

Figure 10: Is it likely that you will use a CAI tool with automatic voice recognition in the future during simultaneous interpretation?

4. Discussion

At task, trigger level, in both sessions 1 and 3, the use of SmarTerp seems to correlate with more positive operations (correct renditions, fewer partial renditions) and fewer negative operations (semantic errors and omissions with minor errors in session 3). All these differences are statistically significant and show that SmarTerp improves student performance in trigger renditions. These results confirm Defrancq/ Fantinuoli's (2021) observations that ASR-based CAI tools increase the accuracy of number renditions. However, usability results seem to indicate that SmarTerp is more ergonomic and efficient than the version of InterpretBank they had employed, which was used only in 55% of cases. Defrancq and Fantinuoli attribute this scant use of InterpretBank in their experiment to participants not being familiar with the tool, which calls into question the need for preliminary training, and being distracted by the running transcript where numbers were highlighted.

Furthermore, we observed statistically significant improvements between sessions 1 and 3, both with and without the tool (correct renditions, partial renditions only without, minor mistakes only with and omissions). Therefore, students appear to require some time to adapt to SmarTerp or to familiarise with speech type and task.

At the sentence level, the differences between successful and unsuccessful source sentence processing with or without CAI tool, in terms of accuracy and coherence, were also highly significant. This result is particularly insightful because studies on the usefulness of CAI tools in SI have so far focused on the micro level of specific problem triggers, thus risking losing sight of other key factors in the source sentence rendition such as fidelity, informativity, coherence, and completeness. This seems to be less so in Chinese studies (Guo *et al*. 2023, 2024).

As for usability, students were generally satisfied or very satisfied with SmarTerp and found it easy to use. They generally perceived an improvement in their performance thanks to the CAI tool but felt the need for preliminary specific training. This same need is reported in various studies involving students (Defrancq/Fantinuoli 2021; Pisani/Fantinuoli 2021; Giudice 2024), but also professionals (Defrancq/Pleșca 2024), which shows that the benefits of a CAI support are jeopardised by untrained attention resource management.

The majority of our students considered SmarTerp trustworthy, as was the case with other students using InterpretBank/ASR (Defrancq/Fantinuoli 2021). However, our study seems to outline an overreliance on the tool, as documented in previous research (Prandi 2015; Defrancq/Fantinuoli 2021; Frittella 2022b; Russello/Carbutto 2023). This shows the importance of stressing that a CAI tool is useful when in need, but does not replace the interpreter's autonomous processing effort.

As to the distraction effect caused by SmarTerp and its latency, our students displayed mixed views. This appears slightly in contrast with other studies, where the intrusiveness and challenges of the extra visual input were reported by the majority (Defrancq/Fantinuoli 2021; Pisani/Fantinuoli 2021; Russello/Carbutto 2023), whereas the tool's excessive latency appears to elicit a general negative perception in all reported studies.

This last feature may explain the tendency we observed in our students to anticipate the SmarTerp input and self-correct after checking. This long latency will have to be redressed by the tool developers.

5. Conclusions

Our study confirms that the use of CAI tools during SI is certainly beneficial for students. Using a "digital boothmate", however, constitutes an extra cognitive effort that does require time and specific training, as pointed out by our trainee participants as well as professionals (Defrancq/Pleșca 2024). Unlike other studies, we gave our students the opportunity to prepare for the use of SmarTerp, both asynchronously and synchronously (the warm-up video), but more time and materials are needed. Attention managing skills do need time before becoming routine procedures.

Students' opinions about latency further point out that there is still room for SmarTerp technical improvement. Our data also confirm the general trend of overreliance on CAI tools, which should be taken into careful consideration in interpreter (self-)training.

ASR and AI-integrated CAI tools have great potential to support interpreting students' workflow both in preparation for an assignment (documentation, parallel terminology management, etc.) and during SI performance (transcription and translation of the aforementioned problem triggers). Even after the assignment, these can be useful to check students' proficiency not only to reproduce the problem triggers but also the whole source speech. The ASR feature could be used to transcribe the student's performance on a different device and have the transcription ready to check semantic faithfulness, terminological accuracy, and the general quality of the delivery. Furthermore, AI systems could be applied to produce alternative versions of the same source speech or play with register and style to improve the student's version, in order to expand their lexical flexibility. Ideally, these further self-training activities are best carried out with peers, so as not to deprive interpreters of invaluable human exchange during their training.

Funding information: The study presented in this paper was conducted as part of the EIT Digital grant for the Innovation Activity 'SmarTerp'.

Division of work: Although the article was jointly planned, laid out and revised, Russo is mainly responsible for Introduction and sections 1.1, 3.1 (and subsections), 4 and 5; Torresi for sections 2 (and sub-sections) and 3.2; Amato for sections 1 and 1.2.

References

Arriaga C. / Villamayor I. / Serrano A. / Cascallana A. / Rodríguez S. / Pozo A. / Alonso A. (2023) "SmarTerp educational: a virtual laboratory to train simultaneous interpreting", in L. Gómez / C. González / J. Lees (eds) *EDULEARN23 Proceedings, 15th International Conference on Education and New Learning Technologies*, Valencia, IATED, 3257-3264 <10.21125/edulearn.2023.0900>.

Baigorri-Jalón J. (2014) *From Paris to Nuremberg: The birth of conference interpreting*, Amsterdam/Philadelphia, John Benjamins.

Braun S. / Clarici A. (1996) “Inaccuracy for numerals in simultaneous interpretation: neurolinguistic and neuropsychological perspectives”, *The Interpreters’ Newsletter* 7, 85-102.

Chmiel A. (2024) “Interpreting and language production”, in C.D. Mellinger (ed.) *The Routledge Handbook of Interpreting and Cognition*, London, Routledge.

Defrancq B. / Fantinuoli C. (2021) “Automatic speech recognition in the booth: assessment of system performance, interpreters’ performances and interactions in the context of numbers”, *Target* 33/1, 73-102.

Defrancq B. / Pleşca A.M. (2024) *Interpreter Interaction with the Artificial BoothMate: An Overview of Participant Experience*, report, Ghent University, <http://hdl.handle.net/1854/LU-01J1D03MXBW78DN4HZQ6BT784M>.

Defrancq B. / Snoeck H. / Fantinuoli C. (2024) “Interpreters’ performances and cognitive load in the context of a CAI tool”, in M. Winters / S. Deane-Cox / U. Böser (eds) *Translation, Interpreting and Technological Change: Innovations in Research, Practice and Training*, London, Bloomsbury, 37-58.

Delisle J. / Woodsworth J. (eds) (1995) *Translators through History*, Amsterdam/Philadelphia, John Benjamins.

Desmet B. / Vandierendonck M. / Defrancq B. (2018) “Simultaneous interpretation of numbers and the impact of technological support”, in C. Fantinuoli (ed.) *Interpreting and Technology*, Berlin, Language Science Press, 13-27.

Fagot C. / Pashler H. (1992) “Making two responses to a single object: implications for the central attentional bottleneck”, *Journal of Experimental Psychology: Human Perception and Performance* 18, 1058-1079.

Fantinuoli C. (2017) “Speech recognition in the interpreter workstation”, in J. Esteves-Ferreira / J. Macan / R. Mitkov / O.M. Stefanov (eds) *Translating and the Computer 39: Proceedings*, Geneva, Tradulex, 25-34.

Fantinuoli C. (2018) “Interpreting and technology: the upcoming technological turn”, in C. Fantinuoli (ed.) *Interpreting and Technology*, Berlin, Language Science Press, 1-12.

Fantinuoli C. (2023) “Towards AI-enhanced computer-assisted interpreting”, in G. Corpas Pastor / B. Defrancq (eds) *Interpreting Technologies – Current and Future Trends*, Amsterdam/Philadelphia, John Benjamins, 46-71.

Fantinuoli C. / Montecchio M. (2023) “Defining maximum acceptable latency of AI-enhanced CAI tools”, in O. Vázquez / A. Correia / S. Araújo (eds) *Technological Innovation put to the Service of Language Learning, Translation and Interpreting: Insights from Academic and Professional Contexts*, Berlin, Peter Lang, 213-222.

Fantinuoli C. / Prandi B. (2021) “Towards the evaluation of automatic simultaneous speech translation from a communicative perspective”, in M. Federico / A. Waibel / M.R. Costa-jussà / J. Niehues / S. Stuker / E. Salesky (eds.) *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, Bangkok, Association for Computational Linguistics, 245-254.

Frittella F.M. (2017) *Numeri in interpretazione simultanea: difficoltà oggettive e soggettive*, Roma, Europa Edizioni.

Frittella F.M. (2021a) *Early Testing Report. Usability Test of the ASR- and AI-powered CAI Tool '*SmarTerp', unpublished report of the SmarTerp project.
Frittella F.M. (2021b) *SmarTerp CAI Tool training*. Moodle course for registered users. <https://smarter-interpreting.eu/training/> (last accessed: 11 February 2023).
Frittella F.M. (2022a) "The ASR-CAI tool supported SI of numbers: sit back, relax and enjoy interpreting?", in J. Esteves-Ferreira / R. Mitkov / M. Recort Ruiz / O.M. Stefanov (eds) *Translating & The Computer 43*, Geneva, Tradulex, 88-102, <https://www.tradulex.com/varia/TC43-OnTheWeb2021.pdf>.
Frittella F.M. (2022b) "CAI Tool-Supported SI of numbers: a Theoretical and Methodological Contribution", *International Journal of Interpreter Education* 14/1, 32-56.
Frittella F.M. (2023) *Usability Research for Interpreter-centred Technology: The Case Study of SmarTerp*, Berlin, Language Science Press.
Gieshoff A.K. / Schuler M. / Yahani Z. (2024) "The augmented interpreter: an exploratory study of the usability of augmented reality technology in interpreting", *Interpreting* 26/2, 282-315.
Gile D. (2009) *Basic Concepts and Models for Interpreter and Translator Training. Revised edition*, Amsterdam/Philadelphia, John Benjamins.
Giudice G. (2024) *Interpreti in formazione e intelligenza artificiale: trascrizione automatica di Google Meet e CAI tool di SmarTerp a confronto*, unpublished MA Thesis, University of Bologna.
Guo M. / Han L. / Anacleto M.T. (2023) "Computer-assisted interpreting tools: status quo and future trends", *Theory and Practice in Language Studies* 13/1, 89-99, <10.17507/tpls.1301.11>.
Guo M. / Han L. / Li D. (2024) "Computer-assisted interpreting in China", in R. Moratto / C. Zhang (eds) *The Routledge Handbook of Chinese Interpreting,* London, Routledge, 439-452, <10.4324/9781032687766-36>.
Korpal P. / Stachowiak K. (2013) "Numerical data processing in simultaneous interpreting: does visual input facilitate the process?", poster presented at the 11th Symposium of psycholinguistics, Tenerife, Spain.
Korpal P. / Stachowiak-Szymczak K. (2018) "The whole picture: processing of numbers and their context in simultaneous interpreting", *Poznan Studies in Contemporary Linguistics* 54/3, 335-354.
Korpal P. / Stachowiak-Szymczak K. (2020) "Combined problem triggers in simultaneous interpreting: exploring the effect of delivery rate on processing and rendering numbers", *Perspectives* 28/1, 126-143.
Lamberger-Felber H. (2001) "Text-oriented research into interpreting: examples from a case-study", *HERMES - Journal of Language and Communication in Business* 26, 39-63.
Lambert S. (2004) "Shared attention during sight translation, sight interpretation and simultaneous interpretation", *Meta* 49/2, 294-306.
Li T. / Chmiel A. (2024) "Automatic subtitles increase accuracy and decrease cognitive load in simultaneous interpreting", *Interpreting* 26/2, 253-281.
Mankauskienė D. (2016) "Problem trigger classification and its applications for empirical research", *Procedia - Social and Behavioral Scien*ces 231, 143-148.

Mazza C. (2001) “Numbers in simultaneous interpretation”, *The Interpreters’ Newsletter* 11, 87-104.

Miller G.A. (1956) “The magical number seven, plus or minus two: some limits on our capacity for processing information”, *Psychological Review* 63, 81-97.

Moser-Mercer B. / Lambert S. / Darò V. / Williams S. (1997) “Skill components in simultaneous interpreting”, in Y. Gambier / D. Gile / C. Taylor (eds) *Conference Interpreting: Current Trends in Research*, Amsterdam/ Philadelphia, John Benjamins, 133-148.

Olalla-Soler C. / Spinolo N. / Muñoz Martín R. (2023) “Under pressure? A study of heart rate and heart-rate variability using the SmarTerp CAI tool”, *HERMES - Journal of Language and Communication in Business* 63, 119-142.

Pashler H. / Johnston J.C. (1998) “Attentional limitations in dual-task performance”, in H. Pashler (ed.) *Attention*, Hove (UK), Taylor & Francis, 155-189.

Pinochi D. (2009) “Simultaneous interpretation of numbers: comparing German and English to Italian. An experimental study”, *The Interpreters’ Newsletter* 14, 33-57.

Pisani E. / Fantinuoli C. (2021) “Measuring the impact of automatic speech recognition on number rendition in simultaneous interpreting”, in C. Wang / B. Zheng (eds) *Empirical Studies of Translation and Interpreting*, London, Routledge, 181-197.

Prandi B. (2015) “The use of CAI tools in interpreters’ training: a pilot study”, in J. Esteves-Ferreira / J. Macan / R. Mitkov / O.M. Stefanov (eds) *Proceedings of the 37th Conference Translating and the Computer*, London, AsLing, 48-57.

Prandi B. (2023) *Computer-assisted Simultaneous Interpreting: A Cognitive-experimental Study on Terminology,* Berlin, Language Science Press.

Rodríguez S. / Frittella F.M. / Okoniewska A. (2022a) “In-booth CAI tool support in conference interpreter training and education”, in J. Esteves-Ferreira / R. Mitkov / M. Recort Ruiz / O.M. Stefanov (eds) *Translating & The Computer 43*, Geneva, Tradulex, 78-87, <https://www.tradulex.com/varia/TC43-OnTheWeb2021.pdf>.

Rodríguez S. / Gaido M. / Hof M.R. / Fernández Pérez M.M. / Toledano Buendía C. / Hernández Cabrera I. (2022b) “University of La Laguna students’ use of SmarTerp-CAI: a study of users’ perception”, unpublished paper presented at the 3rd HKBU international conference on Interpreting and Technology, Hong Kong, 7-9 December 2022.

Rodríguez González E. (2024) *The Use of Automatic Speech Recognition in Cloud-based Remote Simultaneous Interpreting*, PhD Thesis, University of Surrey.

Russello C. / Carbutto M. (2023) “Enhancing numerical accuracy in simultaneous interpreting students: an exploratory study comparing human and AI-based support”, *The Interpreters’ Newsletter* 28, 69-90.

Seeber K.G. (2011) “Cognitive load in simultaneous interpreting: existing theories — new models”, *Interpreting* 13/2, 176-204.

Seeber K.G. (2015) “Simultaneous interpreting”, in H. Mikkelson / R. Jourdenais (eds) *The Routledge Handbook of Interpreting*, London, Routledge, 79-95.

Seeber K.G. (2017) “Multimodal processing in simultaneous interpreting”, in J.W. Schwieter / A. Ferreira (eds) *Handbook of Translation and Cognition*, Hoboken (NY), Wiley & Sons, 461-475.

Timarová Š. (2012) *Working Memory in Conference Simultaneous Interpreting*, PhD thesis, Prague, Charles University / Antwerp, Lessius University College, <https://dspace.cuni.cz/bitstream/handle/20.500.11956/44172/140015273.pdf?sequence=1&isAllowed=y>.

Xiao Luanyi 肖鸾仪 / Wang Yanyan 王艳艳 (2020) “Moni jiqi fuzhu gongneng dui yinghan tongsheng chuanyi biaoxian de ganyu yanjiu” 模拟机器辅助功能对英汉同声传译表现的干预研究 (In English, “An Intervention Study of Simulated Artificial Intelligence Auxiliary Functions in E-C Simultaneous Interpretation”), *Guangdong di”er shifan xueyuan xuebao* 广东第二师范学院学报 *(Journal of Guangdong University of Education)* 40/6, 52-57.