The Interpreters' Newsletter Special Issue 2025 (30), 17-36
DOI: 10.13137/2421-714X/37794 ISSN 1591-4127 (print) ISSN 2421-714X (online)
https://www.openstarts.units.it/dspace/handle/10077/2119

How usable are CAI tools in professional interpreting settings? Insights from a highly technical assignment

Francesca Maria Frittella
University of Surrey
Bianca Prandi
University of Bologna
Susana Rodríguez
SmarTerp

Abstract

The usability of computer-assisted interpreting (CAI) tools has largely been studied in highly controlled conditions. Research in settings that more closely resemble interpreters' working conditions could provide a deeper understanding of how CAI tools can currently support professional interpreting, particularly in challenging assignments and settings characterised by high complexity. This paper reports on a usability study of the CAI tool SmarTerp conducted during a multilingual meeting at the European Patent Office. Six experienced interpreters worked simultaneously from three remote 'dummy' booths (German, English, French). This context involved fast, impromptu speech, a high density of problem triggers, and highly specialised terminology, offering a test of CAI performance under demanding conditions. Data were collected through interpreter observations recorded in digital logs, a post-task usability questionnaire, and focus groups. The findings emphasise the importance of tailoring the automatic speech recognition engine to handle highly technical content, as well as the necessity of providing targeted training to help interpreters integrate CAI tools effectively. The results also point to the value of conducting further research in naturalistic settings to evaluate CAI tools' usability in addition to (quasi-)experimental studies and guide their ongoing development.

Keywords

Computer-assisted interpreting, automatic speech recognition, remote simultaneous interpreting, usability, interpreting technology.

Introduction

Computer-assisted interpreting (CAI) tools are supporting technologies (Braun 2019) that interpreters can use throughout all phases of an assignment. While many practitioners rely on tools developed for general users, this paper focuses on bespoke CAI solutions designed specifically for interpreters, particularly for simultaneous interpreting (SI).

CAI tools can support both preparation and the SI process itself. Their potential has grown in recent years with the integration of automatic speech recognition (ASR), which enables the real-time display of common "problem triggers" (Gile 2009), such as numbers, specialised terms, and named entities (Fantinuoli 2017) of the running transcript of the source speech or live intra- and interlingual captions (see Guo *et al.* 2022).

Most research on CAI has examined the impact of CAI tools on interpreters' performance and cognitive processes in experimental studies, shedding light on human-machine interaction (HCI) in interpreting (see Prandi 2025 for an overview and discussion of the current body of research). More recently, scholars have also examined CAI tool usability, aiming to refine tool functionalities and interface design (e.g. Ghent University 2021; Frittella 2023). Despite this growing interest, research on CAI adopting an explicit usability framework remains scarce, with most studies being conducted under controlled conditions in laboratory settings. As usability testing should closely reflect the context in which a tool is used (Lewis 2012), examining CAI usability in naturalistic tasks is an important next step.

To address this need, the present study investigated interpreters' perceived usability of SmarTerp, a CAI tool that displays isolated problem triggers, when integrated into a remote simultaneous interpreting (RSI) platform during a professional interpreting assignment. The study was conducted in dummy booths during a European Patent Office (EPO) hearing characterised by impromptu speech, a high density of problem triggers and specialised terminology. Data collection combined interpreter and researcher observations, a post-task usability questionnaire, and focus groups.

This paper is structured as follows: section 1 reviews current research on CAI and usability, introducing usability testing as the methodological framework adopted for the study; section 2 presents the methodology; section 3 reports the study findings; and section 4 discusses the results in the broader context of current research on CAI and reflects on their implications. Section 5 offers concluding remarks.

1. Research on CAI and usability

In recent years, research on HCI in interpreting has gained momentum, prompted by the increasing multimodality of interpreters' work environments, for instance in remote settings (Zhang *et al.* 2023), and by staggering advances in supporting technologies and AI. The integration of ASR in CAI tools is a prime example of this, with a growing number of publications exploring their impact on the interpreting process and product.

Research has primarily examined CAI tools' impact on the accuracy of problem triggers, such as numbers (e.g. Desmet *et al.* 2018; Defrancq/Fantinuoli 2021; Pisani/

Fantinuoli 2021), specialised terminology (Van Cauwenberghe 2020; Prandi 2023), and named entities (Yuan/Wang 2023). The impact of these tools on interpreters' cognitive load is also receiving increasing attention (e.g. Prandi 2023; Defrancq *et al.* 2024; Du 2024), also in relation to captions (e.g. Li/Chmiel 2024) and the integration of CAI into augmented reality (Gieshoff *et al.* 2024). Current findings suggest that CAI tools have the potential to effectively improve the accuracy in the rendition of problem triggers, particularly numbers, while results related to the impact on overall quality and on cognitive load paint a less clear picture. Research highlights challenges in interpreter-computer interaction, such as distraction (Wang/Wang 2019) and the difficulty of dividing attention between the tool and other information sources. Against this background, research into CAI usability is gaining momentum (e.g. Frittella 2023; Havelka/Valacchi 2024; Wang *et al.* 2025), aiming to deepen our understanding of interpreter-tool interaction, to identify reasons behind the aforementioned issues, and to find potential solutions through the optimisation of tools' design.

Usability is broadly defined as a property of interactive systems that affects user performance, satisfaction, and acceptance (Bevan *et al.* 1991). The ISO (2018) specifies it as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". It is inherently context-dependent, shaped by interactions among users, tasks, and environments (Nielsen 2010a). Usability can be studied through various methods (Lazar *et al.* 2017), with usability testing – observing representative users completing realistic tasks while reporting perceptions – playing a central role (Nielsen 2010b; Lewis 2012).

It is particularly important to ensure usability for CAI tools used in cognitively demanding tasks like SI, where minor interface issues can disrupt performance (Frittella 2023). Design optimisation is also essential for fostering trust in CAI tools and supporting their adoption in professional practice (Mellinger/Hanson 2018; Fan 2024; Gieshoff *et al.* 2024).

To date, most CAI usability research has been conducted with high experimental control, often using scripted test speeches or ASR mock-ups in laboratory settings (e.g. Frittella/Rodríguez 2022; Saeed *et al.* 2022; Frittella 2023). Methodologically, these empirical studies have typically combined interpreters' performance data with their perceptions, gathered through interviews, focus groups, or post-test questionnaires (see Frittella 2023 for a detailed review), reflecting common practice in usability research (Lewis 2012). This research has identified key principles for problem-trigger-only displays (Frittella 2023) and full running transcripts (Saeed *et al.* 2022, 2023; Rodríguez González *et al.* 2023) and has enabled scholars to assess CAI tools' impact under 'ceiling' performance conditions while controlling for potentially confounding variables such as speed, unplanned speech and the variability inherent in live CAI performance. However, this approach limited the ecological validity of these studies: i) their ability to generalise results to participants' natural environments (Mellinger/Hanson 2022) and ii) their ability to explore usability of CAI tools in professional interpreting. The present study aims to contribute to the existing body of research by investigating to assess whether CAI tool functionality and interface design meet the demands of live interpreting settings.

2. Methodology

2.1 Research questions

The study aimed to explore the usability of the SmarTerp CAI tool in a highly demanding, professional setting, with the goal of advancing our understanding of CAI in terms of user interface design, general usability principles and other requirements for effective user interaction. The following research questions guided the study:

4. How did participants perceive the CAI tool's usability?
5. What factors influenced participants' perception of the tool's usability?
6. How did participants use the CAI tool during interpreting?
7. What insights can inform further improvements of the tool?

The study focused specifically on users' perceptions of the CAI tool. Due to practical constraints and stringent confidentiality requirements (see 2.2), data on the tool's or the interpreters' actual performance were not collected.

2.2 Context and participants

The study took place during a meeting at the EPO. Interpreters at the EPO typically work in hearings on patent disputes, such as oral proceedings before the Boards of Appeal, which may occur either on-site or, frequently, via videoconference. Non-co-located teams of two interpreters per booth commonly work remotely on Zoom, making them highly familiar with RSI. During hearings, parties debate patent texts, with speakers often quoting documents verbatim. Interpreters must handle fast-paced discourse filled with legal jargon and specialised terminology, as well as complex alphanumeric references to legislation or case numbers. Given the strictly confidential nature of the hearings, these are not public, and sessions are not recorded. Interpretation is provided into one of the EPO's three official languages (English, French, German), with interpreters working from their B/C languages into A. The present study therefore, recreated the interpreters' natural working environment.

The meeting observed for this study featured a French-speaking patent proprietor, two German-speaking opponents, one English-speaking opponent and three members of the Board of Appeal. Six professional conference interpreters were recruited for the study. They interpreted in dummy booths, as the stakes were too high to test the CAI tool for the first time for a real audience. They were remunerated, which reinforced the ecological validity of the study: their motivation and approach mirrored those of a real assignment (Mellinger *et al.* 2025: 52) and ensured that they could dedicate time to the pre-study onboarding and preparation.

An anonymised pre-experiment questionnaire gathered demographic and professional data. Two interpreters had German (DE) as their A language, two English (EN) and two French (FR). Interpretation was provided into their respective A languages, with relay interpreting as needed. Experience levels varied: one interpreter reported six to ten years of experience; two reported 11–20 years; and three reported over 20

years (mean using category mid-points: 19 years). As for EPO-specific experience: one interpreter had up to two years; four had three to five years; and one had more than 20 years, with a mean of approximately seven years. Only one participant had previously used automatically generated captions, making participants highly experienced interpreters but novice users of ASR-based CAI tools. Concerning their attitudes towards technology, four participants defined themselves as 'enthusiastic' and two described their attitude as 'open' to trying new technologies if found potentially beneficial.

2.3 The SmarTerp CAI tool

The SmarTerp CAI tool (Figure 1) is integrated into the SmarTerp RSI platform. While the tool offers functionalities supporting the preparation phase, this study focused on the in-booth CAI function powered by ASR.

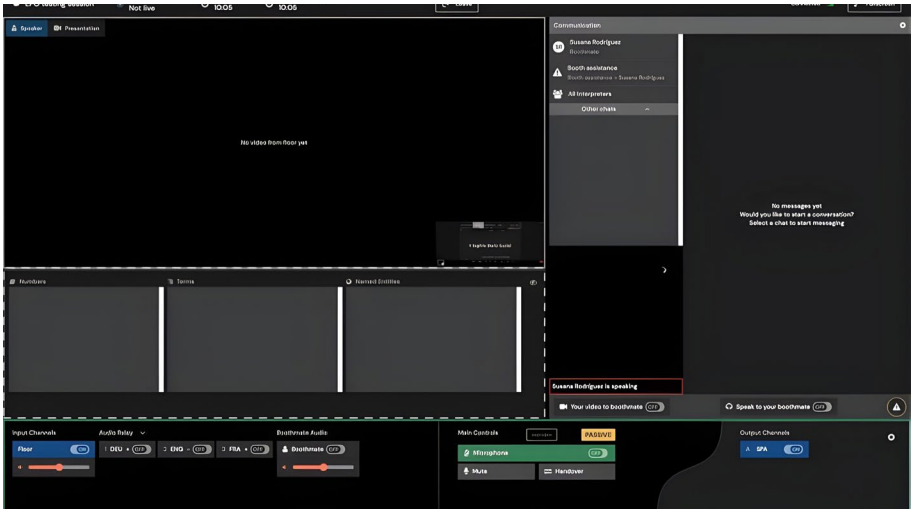

Figure 1: The CAI tool (dotted line) integrated into the SmarTerp RSI platform

The interface is organised into three columns, one for each problem trigger class: numerals, terms and acronyms, and named entities. New items appear at the top of the interface and are highlighted in blue for three seconds. Repeated items already visible in the column are not repeated but only highlighted. A distinctive feature of SmarTerp is that terms recognised via the glossary appear in the 'Terms' column, whereas automatically extracted terms appear in the 'Named Entities' column. For more details on the interface, its functioning, and the user-centred development process, see Fossati (2021) and Frittella (2023).

In highly technical assignments, the ASR engine is usually fine-tuned for each client to maximise accuracy. For this study, customisation was not possible due to resource constraints and strict privacy requirements, therefore, the generic ASR was used instead.

2.4 Procedure

The division of roles among researchers in the study procedure was as follows. The third author of the paper organised the test sessions, prepared the CAI tool, and was present during the test in the role of technician to address any possible technical issues or questions from participants. She was not involved in the study design nor in the collection and analysis of data, which were conducted by the first two authors (see 2.5), to prevent possible bias given her leading role in the development of SmarTerp.

All participants completed a self-paced online onboarding programme covering both the RSI platform and the SmarTerp CAI tool approximately one week before the test. Completion of all activities within the agreed timeframe was verified via learner analytics data. The training introduced the tool interface (appearance and layout) and the underlying technology and included practical exercises to familiarise participants with its use. These simulation-based exercises allowed participants to practice typical actions on the RSI console (turn-taking, relay, etc.), and a pre-recorded speech with the CAI tool enabled them to experience interpreting with real-time tool support. Each participant prepared a glossary which was fed to the CAI tool for the provision of terminological suggestions.

The study procedure lasted approximately five hours including breaks. Participants interpreted remotely from home with their usual set-up, using fibre-optic (three), cable (two), or DSL (one) internet connections. During the pre-test sound check, four participants reported excellent audio quality and two rated it as good.

The session began with a briefing, followed by a two-hour test phase. Both researchers and participants accessed the SmarTerp RSI platform. Due to stringent EPO privacy requirements, the sessions were not recorded, and interpreters' performance was not analysed.

Data were collected from three sources: digital logs, a usability questionnaire, and focus group discussions. After each interpreting turn, participants logged observations in the digital logs, which were reviewed by researchers alongside any issues reported in the platform chat. All personal data was anonymised. Following the test, participants completed the usability questionnaire. After a one-hour break, they participated in focus group sessions concluding with a plenary discussion. The following sections describe the three data sources in more detail.

2.4.1 The digital log

Each interpreter used a shared online document accessible to the researchers to record observations about the tool's performance after each interpreting turn. The document contained a two-column table: one column for negative impacts on performance and another for positive impacts. In parallel, the first two authors monitored the CAI tool during the sessions, noting any relevant observations about its performance and its influence on interpreting in a separate log.

2.4.2 The usability questionnaire

A questionnaire was used to explore interpreters' experiences with the SmarTerp CAI tool. The same instrument was employed in previous usability tests of the tool (Frittella/Rodríguez 2022; Frittella 2023). Although the questionnaire has not been formally validated, it was specifically designed to address limitations of applying existing usability instruments, such as UEQ (Laugwitz *et al.* 2008), SUS (Brooke 1996), and NASA-TLX (Hart/Staveland 1988), to a CAI context. In earlier studies, interpreting and usability experts noted that some descriptors in standard instruments could be ambiguous for CAI tasks (Frittella 2023: 86). Consequently, the questionnaire adapted items from the UEQ to better suit CAI, enabling comparison of user responses across studies.

Participants gave Likert-type ratings in a 7-point range. The questionnaire addressed the following areas:

- Satisfaction: overall satisfaction with the SmarTerp CAI tool (1 = very dissatisfied, 7 = very satisfied).
- Perception: agreement with statements on key usability attributes (1 = fully disagree, 7 = fully agree):
  - *Effectiveness*: the CAI tool helped improve the quality of my delivery.
  - *Ease of use*: I could use the CAI tool without unnecessary effort or confusion.
  - *Ease of learning*: the CAI tool can be used successfully without training.
  - *Dependability*: I felt that I could trust the CAI tool at all times.
  - *Timeliness*: the CAI tool's support was delivered when needed.
  - *Stressfulness*: the CAI tool made me feel more confident and less stressed.
- Likelihood of future adoption: likelihood of using an ASR-based CAI tool during SI in the future (1 = very unlikely, 7 = very likely). This question was asked for the first time in the interpreters' demographic questionnaire and repeated in the usability questionnaire, to record any variations in perception.
- Improvement suggestions: open-ended recommendations for improving the tool.

2.4.3 Focus group

A one-hour focus group was conducted in virtual Zoom breakout rooms to further explore participants' perceptions, experiences and suggestions for improvement of the CAI tool. Despite the potential for bias in focus groups, such as groupthink (Morgan 1997), we chose this method for qualitative data collection due to participants' limited time availability and considering that the data collection procedure was already extensive. Moreover, individual perspectives had already been captured through the questionnaires, so the focus groups served to explore issues in greater depth rather than to establish a comprehensive individual-level picture. Participants were divided into two groups of three (Group A and Group B), each including one interpreter from each booth. The discussion was moderated by the first and second authors of this paper and focused on participants' experiences and perceptions of the SmarTerp CAI tool, guided by the following questions:

– Did the SmarTerp CAI tool help you perform better than without support?
– What do you see as the main benefits of the CAI tool?
– Are there any major drawbacks to using the CAI tool?
– What changes or additions would improve the CAI tool for interpreting at the EPO?

Following the breakout sessions, a 30-minute plenary discussion was held. The researchers summarised the main points from each group, presenting them to the other group to stimulate further reflection (e.g. "Group A highlighted the following points: … Does Group B agree?"). Any differences in responses were noted, and additional considerations were discussed by the researchers. The session also allowed participants to ask final questions on the study or the tool.

2.5 Data analysis

Questionnaire responses to the structured items were converted to a range from -3 (most negative) to +3 (most positive) to facilitate comparison across participants and align with previous studies using the same instrument (Frittella/Rodríguez 2022; Frittella 2023). For data analysis, participants' responses regarding effectiveness after testing were compared with their pre-test responses in the demographic questionnaire.

To enhance the reliability of findings, qualitative data from the three sources – digital logs, open-ended questionnaire responses, and focus groups – were initially analysed independently by the first and second authors using a bottom-up coding approach. Subsequently, the researchers compared their codes and agreed on a shared set of categories. Finally, results were triangulated by integrating insights from all data sources.

Open-ended questionnaire responses were coded using thematic analysis. Each response was reviewed, and recurring themes were identified, with frequency counts recorded to indicate the prevalence of specific issues or suggestions. Codes were both concept-driven, i.e. derived from previous literature and from the structured items in the questionnaire (see 2.4.2), and data-driven following an inductive approach (Gibbs 2021).

Focus group discussions were audio-recorded and transcribed verbatim. Responses to each guiding question were coded thematically following the same approach as for open-ended questionnaire responses, with attention paid to the frequency of themes mentioned by participants across groups to facilitate the identification of common patterns (see Saldanha/O'Brien 2014: 189).

Observations recorded in the digital interpreter logs were first aggregated by booth to identify recurring issues and patterns related to the language combination. The researchers cross-checked these logs with their own logs to ensure completeness and to capture any additional observations concerning the CAI tool's functionality. Recurring positive and negative impacts were coded inductively and summarised into categories to highlight the most salient points across booths.

3. Results

In the following sections, study findings are presented with reference to each research question, highlighting convergent and divergent insights across data sources for each of the categories identified during data analysis.

3.1 How did participants perceive the CAI tool's usability?

Our first research question was explored primarily through the questionnaire. Therefore, the results below are structured around the key questionnaire themes, which are integrated with relevant insights from the focus groups and digital logs to gain deeper understanding of participants' perceptions.

The first theme that we explored was participants' usability perception. Figure 2 presents participants' perceptions for each usability criterion. Given the small sample size, only median values are reported as a measure of central tendency. Participants generally did not perceive the tool as highly effective in improving their delivery. However, it was considered relatively easy to use, causing minimal effort or confusion, though a certain level of training was deemed necessary to use it effectively. Opinions on the timeliness of support were neutral, while the most negative evaluations concerned the tool's dependability and the stress it induced.

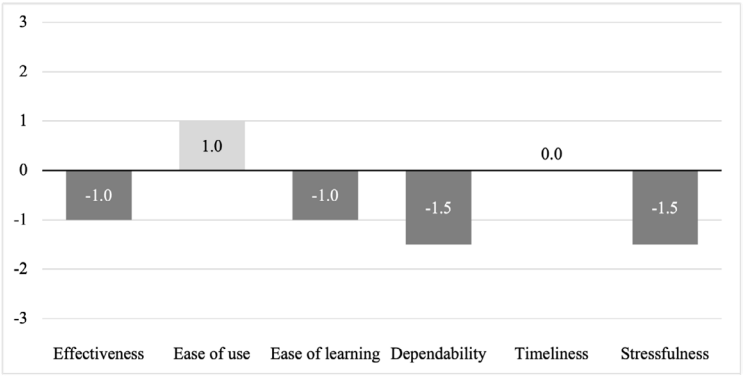

Figure 2: Questionnaire results - Perception of SmarTerp by usability criteria

Satisfaction also varied by problem trigger class. Figure 3 shows participants' median satisfaction levels for each class supported by the tool. Participants were most satisfied with suggestions for specialised terms, less satisfied with acronyms and EPO-specific triggers (references to claims, paragraphs, applications, and patent numbers), and notably disappointed with support for named entities and numbers. Overall, the level of satisfaction with the tool appears to be quite low.

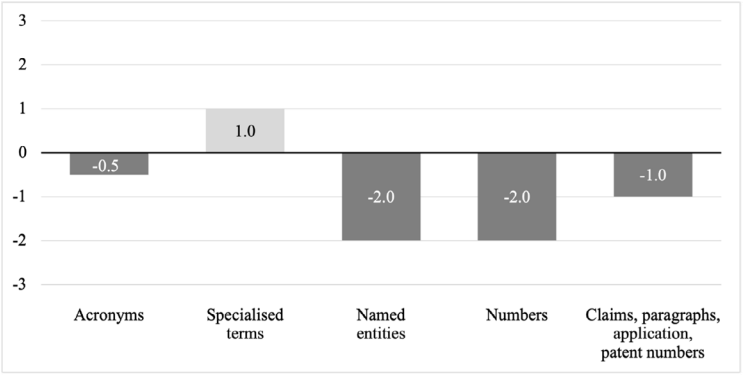

Figure 3: Questionnaire results - Satisfaction by problem-trigger class

The low scores for effectiveness, dependability and stressfulness appear to reflect several issues with the tool's output for the different problem-trigger classes. The thematic analysis of the interpreters' logs and focus group discussions led to the identification of the overarching theme of *problem triggers* for which we defined three subcategories, as not all problem triggers were present in the qualitative data: *specialised terms*, *named entities*, and *numbers* (which also included EPO-specific numbers). Two more codes (*positive impact*, *negative impact*) were defined to identify benefits and issues for each problem trigger category. For the code *negative impact*, we identified three categories of issues which can help understand the relatively low usability ratings: *omissions*, *inconsistencies* and *errors*.

Omissions affected both terms and numerals. For terms, the system sometimes failed to recognise multi-word expressions (e.g. 'Druck-Durchfluss-Paar' [pressure-flow pair]). For numbers, the tool suggested numerals described as 'obvious' by participants (e.g. 'claim 3', 'paragraph 6'). However, it did not aid with those numerals considered as most complex by participants, such as quickly cited phone numbers, patent and application numbers, or complex numerical codes. For example, interpreter DE1 expected the tool to reproduce numbers calculated according to a specific formula and was puzzled that it did not, despite the speaker's moderate pace. In the focus group discussion, participants reported being irritated and disappointed by the insufficient support for numerals.

Participants also reported *inconsistencies* in the visual aids, meaning that system aids occasionally appeared in the wrong interface section. This issue was identified for terms and named entities: certain terms (e.g. 'Gleitshole' [sliding sleeve], 'Verstellkörper' [adjustment element]) were classified as named entities rather than terms.

Occasional *errors* in the visual aids were reported for all problem triggers. For instance, interpreter DE1 remarked: "It is important to have the features and claims displayed correctly. In one instance, it said 'Anspruch 1' [claim 1] and directly after that came 'Anspruch 1000000' [claim 1,000,000] for no obvious reason". Interpreters highlighted how such issues increased cognitive effort: "As I was working from the

relay, I didn't get a German source word, but a French one, which added to the cognitive load", stated EN1.

The code *positive impact* was found only in relation to specialised terms and named entities. For instance, FR2 noted that when the system correctly recognised technical terms, the suggestions were highly beneficial, while EN2 appreciated the tool's suggestions for speakers' names. We could not identify this category for numbers, which is in line with the negative results for these problem triggers from the questionnaire.

Lastly, the questionnaire inquired into participants' self-reported likelihood of using the tool in the future to help shed light on variations in participants' attitude towards CAI tools following the test. Figure 4 presents a comparison of the response to this question before and after the test for each interpreter.

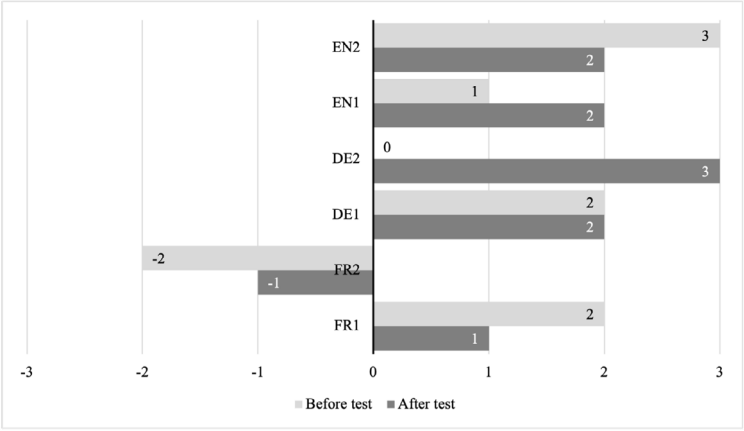

Figure 4: Questionnaire results - Self-reported likelihood to use the CAI tool in the future before and after the test

Improvements in the likelihood of future use were observed for three of the six participants (EN1, DE2, FR2). One interpreter (DE1) maintained their opinion, which was already quite positive prior to the test, while two interpreters (EN2, FR1) reported a one-point decrease. These results align with what emerged from focus group analysis, as several participants' comments were related to the theme *tool potential*. Participants in Group B stated that, although they would not use the tool in its current state, they recognised its potential, noting that CAI tools are very useful, and they will be even more useful when they improve. A similar view emerged from Group A, who indicated that with the necessary customisation of the ASR engine – even if only for terminology – they would consider using the tool in the future.

3.2 What factors influenced participants' perception of the tool's usability?

In addition to the factors strictly related to the tool's perceived performance on problem triggers, additional factors emerged from the focus group analysis which seem

to have affected perceived usability. Three themes were identified: *latency*, *perceived redundancy*, *unmet expectations*.

System *latency* was mentioned by all participants. Opinions were generally positive to neutral, which is in line with the results for the timeliness dimension reported in 3.1. Two participants, however, expressed a more negative opinion: DE1 felt that the tool was too slow at the start, while EN2 judged the latency as excessive, noting that the tool was "much too slow – so slow that any suggestions come too late to be used". At the EPO, interpreters must maintain a very short decalage: the same interpreter, who has extensive experience using a CAI tool with manual look-up, reported being much faster when consulting their own glossary.

Another common theme was the *perceived redundancy* of the tool's suggestions. As interpreter FR1 put it, "legal and procedure terms [e.g. 'patent'] are unnecessary. They're distracting more than helping". FR2 added: "It makes it difficult to focus on the technical terms which are the ones that are really needed". Hence, participants identified the provision of redundant suggestions as a distracting factor which could undermine the effectiveness of the interaction.

The *expectations* theme emerged from the analysis as a further factor determining participants' perceived usability of the tool. For example, EN2 reported that the French word 'recours' was automatically translated as 'appeal'. While correct in certain contexts, it appeared in the phrase 'avoir recours' [to rely on], so the suggestion was confusing rather than helpful. Hence, the participant expected that the tool would provide context-aware translation suggestions, albeit this is currently unfeasible from a technical standpoint. Similar expectations concerned EPO-specific numbers, which were not easily recognisable by the ASR.

3.3 How did participants use the CAI tool during interpreting?

Related to this research question is the theme of *strategic use* identified analysing focus group discussions and open-ended questionnaire responses. The focus groups revealed five main strategies for interacting with the tool: *disengagement*, *selective engagement*, *stalling*, *use of additional support* and *strategic preparation*.

As mentioned in the focus groups, participants deliberately *disengaged* from the tool when its support was considered unhelpful, such as when translation suggestions were unsuitable for the context or when expectations were not met. Previous experience helped participants use the tool strategically: interpreter DE1 noted, "it was my second test of one of these solutions. I have learnt that you have to look away from the tool".

A related strategy was *selective engagement*: some participants observed that not all columns consistently provided useful support, and after a while, they began selectively consulting only the column they deemed relevant.

Stalling was mentioned as a way of coping with perceived excessive latency. Although this was not a prevalent theme that emerged from the analysis, we report it as it reveals a proactive way of engaging with the tool even when it does not entirely meet the user's expectations.

As emerged from the questionnaire, five of the six interpreters reported *using additional support* tools and resources, such as notepads, digital glossaries, dictionaries,

or printed documents. Three participants combined a digital glossary with the ASR input, one used a different CAI tool (Interpreters' Help), and another relied on their own terminology programme. Four participants also consulted patent and/or claims documents, either in print or PDF format, to manage read-out passages that required sight translation.

In the focus group discussions, participants reflected on the theme of *preparation,* which was recognised as an important strategic component in the interaction with the tool. Some participants indicated that they would continue creating glossaries manually rather than relying solely on SmarTerp's automatic approach, as the process helped them memorise terms. Other participants stated that they would generate glossaries automatically with the CAI tool but emphasised that knowledge-based preparation remains essential for comprehension during interpreting.

3.4 What insights can inform further improvements to the tool?

To address our final research question, we looked at focus group remarks which were summed up in the categories of *ASR customisation*, *interface design*, and *strategies for redundancy reduction*.

The theme of *ASR customisation* was the most prevalent and it highlighted the specific needs of interpreters working in this setting. Participants suggested that ASR be customised to recognise specialised jargon, particularly alphanumerical combinations such as document numbers (e.g. 'ES-94, ES-95, D-14') and other EPO-specific numerical references, including decision, case law and Board of Appeal case numbers. These items are often read rapidly and in lists, significantly contributing to interpreters' cognitive load, as noted by participant DE2. In the questionnaire, one participant also suggested that it would be very helpful if the tool could automatically pull up the relevant document texts.

Several comments concerned the *interface design.* Participants suggested that all terms should appear in a single, central column, independent of the source (glossary *vs*. NER). Some terms were displayed by the tool in the NER column. This caused confusion, and sorting through named entities and terms within the same NER column was described as unfeasible during interpreting. Other requests related to the interface design for terminology included a function for real-time editing of suggested terms and displaying multiple translation options where available.

Also related to interface design, in the questionnaire participants mentioned potential approaches to *reducing redundancy* in the items displayed automatically. One participant suggested adding a button to manually stop displaying repeated items, another wished that general legal or EPO-related terms could be excluded automatically, and a participant suggested finding a way to automatically prioritise low-frequency terms over high-frequency, essential ones.

4. Discussion

In the present study, overall satisfaction with the CAI tool was markedly lower than in previous evaluations of SmarTerp conducted in controlled conditions (Frittella/

Rodríguez 2022; Frittella 2023). This may be due to differences in the study design and setting. The engine had not been fine-tuned to EPO-specific problem triggers in this study, meaning that the tool's 'ceiling' performance during a live assignment could not be tested, as was done in previous research on CAI where mock-ups were used (e.g. Desmet *et al.* 2018; Prandi 2023). However, the outlook is not entirely negative. Although satisfaction ratings were low, the self-reported likelihood of future use increased slightly for several participants and focus group discussions revealed that interpreters recognised the potential of CAI technology. These results suggest that first-hand experience may foster openness towards future experimentation with CAI tools, despite usability challenges, and are in line with the cautious openness towards technology identified in recent surveys (Fan 2024).

The major sources of dissatisfaction were the tool's suggestions for numbers. This aligns with previous studies on SmarTerp (Frittella 2023), but contrasts sharply with current findings on CAI tool's impact on number accuracy, which have generally been positive (Desmet *et al.* 2018; Defrancq/Fantinuoli 2021; Pisani/Fantinuoli 2021). This discrepancy can be attributed to the high complexity of EPO-specific numerical references, underscoring the setting's impact on the usability of CAI tools. Similar observations can be made regarding latency, which was technically within interpreters' average ear-voice span (Fantinuoli/Montecchio 2023), and had been identified as satisfactory in previous tests on SmarTerp (Frittella/Rodríguez 2022). However, it was sometimes deemed excessive, reflecting the high pace and density of problem triggers in the source speech.

The higher level of satisfaction reported for terms is consistent with previous product-oriented research indicating a positive impact on terminological accuracy (Prandi 2018, 2023; Van Cauwenberghe 2020). This may be explained by the use of interpreters' glossaries as support for term recognition. Interpreters have been found to prefer their own glossaries to machine-generated suggestions (see Fantinuoli *et al.* 2022). However, we were unable to collect data on interpreters' performance, which might have revealed additional issues, as occurred with the reported interference with sentence planning in Gieshoff *et al.* (2024).

Participants scored the tool low on dependability and stressfulness, expressing feelings of irritation and disappointment. This contrasts with the positive psychological effect of the tool's presence, as reported for instance in Defrancq/Fantinuoli (2021). Redundant suggestions were seen as distracting and frustrating. Similar findings have been reported in previous studies, which found that locating relevant terms in lists was cognitively demanding (Van Cauwenberghe 2020; Prandi 2023), and that visual overwhelm was challenging (Wang/Wang 2019; Defrancq/Fantinuoli 2021). However, it should be noted that what participants perceived as redundant was highly subjective and often linked to their expertise, familiarity with the assignment and glossary preparation practices.

With regard to the impact of interface design on perceived usability, the study emphasised the detrimental effect of interface inconsistencies, such as problem triggers appearing in unexpected locations. This corroborates the findings of previous usability studies, which highlighted the importance of item placement in the interface for users. This can be achieved both through customisation (Gieshoff *et al.* 2024) or by prioritising consistency, as was found in the present study, and also by Saeed *et*

al. (2022, 2023) for complete captions. Interface designs that prioritise consistency may help interpreters develop automatic behaviours, freeing up cognitive capacity for other interpreting processes (Frittella 2024).

This is important because participants mentioned a negative impact on their cognitive processes and referred to cognitive load. Thus far, research on CAI tools and cognitive load has centred on the impact of tool performance, identifying common issues such as overreliance (Prandi 2015; Defrancq/Fantinuoli 2021), while also suggesting a potentially positive influence (e.g. Li/Chmiel 2024). It may therefore be worthwhile to explore the link between perceived usability and cognitive load, especially as our findings emphasise that interpreters' expectations and behaviour can profoundly shape their interaction with CAI tools, sometimes as much as, or more than, tool performance itself.

To mitigate usability challenges, interpreters employed various adaptive strategies. These included selective disengagement from certain interface elements, stalling, and reliance on additional support tools, such as glossaries and reference documents (as also observed in Prandi 2015). This supports the framing of CAI as a complex skill requiring technological literacy and strategic adaptation (Frittella 2024). Effective CAI tool use depends on having realistic expectations, being well prepared, and displaying strategic behaviour. Participants' frustration and disengagement illustrate how unmet expectations can undermine both usability and uptake, as discussed by Mellinger/Hanson (2018) and Gieshoff *et al.* (2024) among others. These results also highlight that interpreters' preferences and approaches to the interaction can vary significantly, as the analysis conducted by Frittella (2023) also revealed.

The study has some limitations. The first relates to its ecological validity. ASR customisation, which is usually carried out by SmarTerp developers prior to assignments, could not be performed due to time and budget constraints. Interpreters worked in dummy booths rather than during a live assignment, which potentially affected their focus and behaviour. Additionally, participants were aware that the study aimed to evaluate usability, potentially leading them to scrutinise the tool more critically. Our study also highlighted some of the challenges that naturalistic studies may face, such as confidentiality constraints and technical limitations, both of which can be more effectively addressed in laboratory settings. For instance, the strict confidentiality requirements at the EPO prevented us from collecting objective performance data and measuring the tool's precision and recall (see Fantinuoli 2017). Controlled studies therefore remain crucial, particularly when investigating design principles that require carefully crafted source material to elicit meaningful interactions (see Frittella 2023).

The sample size in our study was small, which is consistent with qualitative research norms and usability testing standards (Nielsen 2010a; Lewis 2012). We explored a specific setting and a specific CAI tool. Additionally, we did not analyse the preparation phase, which can influence interaction (Du 2024) and performance (Xu 2018). The test reveals participants' perception of usability after a short training phase and their first experience of using the tool during an assignment (for all but one participant). All this limits our ability to generalise the findings and only allows us to partially respond to research questions related to factors influencing perceived usability. Finally, although the usability questionnaire was based on validated scales and had previously been used to evaluate SmarTerp, it has not itself been validated as a

standalone instrument, which limits its construct validity. An important next step will be to validate the questionnaire and compare it with other validated instruments, thus ensuring higher reliability of findings related to perceived usability.

5. Conclusion

In response to the need to evaluate CAI tools in conditions representative of interpreters' complex work realities, this study examined the perceived usability of SmarTerp during a highly technical, live assignment at the EPO.

Our findings contribute to existing research by suggesting that the use of CAI tools for complex, naturalistic interpreting tasks can produce different usability outcomes to those reported in laboratory studies. This highlights the importance of ecologically valid research, particularly to understand CAI tool users' expectations, satisfaction, and requirements, while also emphasising the challenges and constraints of such studies, and positioning controlled usability studies as a complementary approach within these limitations. Complementing self-reported usability data with objective performance metrics for both tools and interpreters, could provide a more nuanced understanding of CAI tools in professional interpreting practice, promoting interpreters' acceptance (Mellinger/Hanson 2018), and facilitating their long-term integration into professional workflows.

Furthermore, our results suggest that insufficient user preparation can result in dissatisfaction, disengagement from the tool and suboptimal use. This emphasises the critical role of training in facilitating effective interaction with CAI systems and encouraging user adoption.

As is being done in workplace studies in translation, adopting socio-cognitive approaches (see Sannholm/Risku 2024) investigating the influence of booth collaboration, relay interpreting, and RSI on CAI usability and design, could provide further insight into these complex interactions. This may help to address calls to study interpreter-computer interaction from situated cognition perspectives (Mellinger 2023), providing insight into the relationship between usability challenges and their impact on interpreters' cognitive processes.

References

Bevan N. / Kirakowski J. / Maissel J. (1991) "What is usability?", in *Proceedings of HCI International 91*, Stuttgart, Elsevier.

Braun S. (2019) "Technology and interpreting", in M. O'Hagan (ed.) *The Routledge Handbook of Translation and Technology*, London, Routledge, 271-288.

Brooke J. (1996) "SUS: a 'quick' and 'dirty' usability scale", in P. W. Jordan / B. Thomas / I. L. McClelland / B. A. Weerdmeester (eds) *Usability evaluation in industry*, London, CRC press, 189-194.

Defrancq B. / Fantinuoli C. (2021) "Automatic speech recognition in the booth: assessment of system performance, interpreters' performances and in-

teractions in the context of numbers", *Target* 33/1, 73-102, <https://doi.org/10.1075/target.19166.def>.
Defrancq B. / Snoeck H. / Fantinuoli C. (2024) "Interpreters' performances and cognitive load in the context of a CAI tool", in S. Deane-Cox / U. Böser / M. Winters (eds) *Translation, Interpreting and Technological Change: Innovations in Research, Practice and Training*, London, Bloomsbury Academic, 38-58.
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.
Du Z. (2024) *Bridging the Gap. Exploring the Cognitive Impact of InterpretBank on Chinese Interpreting Trainees*, unpublished PhD Thesis, Università di Bologna at Forlì.
Fan D. C. (2024) "Conference interpreters' technology readiness and perception of digital technologies", *Interpreting* 26/2, 178-200, <https://doi.org/10.1075/intp.00110.fan>.
Fantinuoli C. (2017) "Speech recognition in the interpreter workstation", in J. Esteves-Ferreira / J. Macan / R. Mitkov / O-M. Stefanov (eds) *Proceedings of the 39th Conference Translating and the Computer*, London, Editions Tradulex, 25-34.
Fantinuoli C. / Marchesini G. / Landan D. / Horak L. (2022) "KUDO Interpreter Assist: automated real-time support for remote interpretation", in R. Mitkov / J. Macan / J. Esteves-Ferreira / O-M. Stefanov / M. Recort Ruiz / D. Chambers / V. Sosoni (eds) *Proceedings of the 43rd Conference Translating and the Computer*, Geneva, Editions Tradulex, 68-77.
Fantinuoli C. / Montecchio M. (2023) "Defining maximum acceptable latency of AI-enhanced CAI tools", in Ó. Ferreiro 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, 213-225.
Fossati G. (2021) *SmarTerp: Applying the User-Centered Design Process in a Computer-Assisted Interpreting (CAI) Tool*, unpublished MA Thesis, Universidad Politécnica de Madrid at Madrid.
Frittella F. M. (2023) *Usability Research for Interpreter-centred Technology: The Case Study of SmarTerp*, Berlin, Language Science Press.
Frittella F. M. (2024) *Computer-assisted Interpreting: Cognitive Task Analysis and Evidence-Informed Instructional Design Recommendations*, unpublished PhD Thesis, University of Surrey at Guildford.
Frittella F. M. / Rodríguez S. (2022) "Putting SmartTerp to test: a tool for the challenges of remote interpreting", *INContext: Studies in Translation and Interculturalism* 2/2, 137-166, <https://doi.org/10.54754/incontext.v2i2.21>.
Ghent University (2021) *Ergonomics for the Artificial Boothmate (EABM) Survey*, <https://www.eabm.ugent.be/survey>.
Gibbs G. R. (2021) *Analyzing Qualitative Data*, London, SAGE, <https://doi.org/10.4135/9781526441867>.

Gieshoff A. C. / Schuler M. / Jahany Z. (2024) "The augmented interpreter: an exploratory study of the usability of augmented reality technology in interpreting", *Interpreting* 26/2, 282-315, <https://doi.org/10.1075/intp.00108.gie>.
Gile D. (2009) *Basic Concepts and Models for Interpreter and Translator Training*, Amsterdam/Philadelphia, John Benjamins.
Guo M. / Han L. / Anacleto M. T. (2022) "Computer-Assisted interpreting tools: status quo and future trends", *Theory and Practice in Language Studies* 13/1, 89-99, <https://doi.org/10.17507/tpls.1301.11>.
Hart S. G. / Staveland L. E. (1988) "Development of NASA-TLX (Task Load Index): results of empirical and theoretical research", in P.A. Hancock / N. Meshkati (eds) *Advances in Psychology*, Amsterdam, North-Holland, 139-183, <https://doi.org/10.1016/S0166-4115(08)62386-9>.
Havelka I. / Valacchi M. (2024) "Interpreter-computer interaction in RSI: the role of graphical user interfaces in interpreters' perception", *Revista Tradumàtica. Tecnologies de la Traducció* 22, 377-400, <https://doi.org/10.5565/rev/tradumatica.406>.
ISO (2018) *Ergonomics of Human-system Interaction – Part 11: Usability: Definitions and Concepts*, <https://www.iso.org/standard/63500.html>.
Laugwitz B. / Held T. / Schrepp M. (2008) "Construction and evaluation of a user experience questionnaire", in A. Holzinger (ed.) *HCI and Usability for Education and Work*, Berlin, Heidelberg, Springer, 63-76, <https://doi.org/10.1007/978-3-540-89350-9_6>.
Lazar J. / Feng J. H. / Hochheiser H. (2017) *Research Methods in Human-Computer Interaction*, Cambridge, Morgan Kaufmann.
Lewis J. R. (2012) "Usability testing", in G. Salvendy (ed.) *Handbook of Human Factors and Ergonomics*, 4th edition. Hoboken, NJ, John Wiley & Sons, 1267-1312, <https://doi.org/10.1002/9781118131350.ch46>.
Li T. / Chmiel A. (2024) "Automatic subtitles increase accuracy and decrease cognitive load in simultaneous interpreting", *Interpreting* 26/2, 253-281, <https://doi.org/10.1075/intp.00111.li>.
Mellinger C. D. (2023) "Embedding, extending, and distributing interpreter cognition with technology", in G. Corpas Pastor / B. Defrancq (eds) *Interpreting Technologies – Current and Future Trends*, Amsterdam, John Benjamins, 195-216.
Mellinger C. D. / Hanson T. A. (2018) "Interpreter traits and the relationship with technology and visibility", *Translation and Interpreting Studies* 13/3, 366-392, <https://doi.org/10.1075/tis.00021.mel>.
Mellinger C. D. / Hanson T. A. (2022) "Considerations of ecological validity in cognitive translation and interpreting studies", *Translation, Cognition & Behavior* 5/1, 1-26, <https://doi.org/10.1075/tcb.00061.mel>.
Mellinger C. D. / Spinolo N. / Ehrensberger-Dow M. / O'Brien S. (2025) "Designing studies with naturalistic tasks", in A.M. Rojo López / R. Muñoz Martín (eds) *Research Methods in Cognitive Translation and Interpreting Studies*, Amsterdam, John Benjamins, 49-68, <https://doi.org/10.1075/rmal.10.02mel>.

Morgan D. (1997) *Focus Groups as Qualitative Research*, Thousand Oaks, SAGE, <https://doi.org/10.4135/9781412984287>.
Nielsen J. (2010a) "What is usability?", in C. Wilson (ed.) *User Experience Re-mastered*, Boston, Morgan Kaufmann, 3-22, <https://doi.org/10.1016/B978-0-12-375114-0.00004-9>.
Nielsen J. (2010b) *Usability Engineering*, San Francisco, Morgan Kaufmann.
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: The Post-Structuralist Approach*, New York, Routledge, 181-197, <https://doi.org/10.4324/9781003017400>.
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. (2018) "An exploratory study on CAI tools in simultaneous interpreting: theoretical framework and stimulus validation", in C. Fantinuoli (ed.) *Interpreting and Technology*, Berlin, Language Science Press, 29-59.
Prandi B. (2023) *Computer-Assisted Simultaneous Interpreting: A Cognitive-Experimental Study on Terminology*, Berlin, Language Science Press, <https://doi.org/10.5281/zenodo.7143056>.
Prandi B. (2025) "Computer-Assisted Interpreting (CAI) tools and CAI tool training", in E. Davitti / S. Braun / T. Korybski (eds) *Routledge Handbook of Interpreting, Technology and AI*, London, Routledge, 123-144.
Rodríguez González E. / Saeed M. A. / Korybski T. / Davitti E. / Braun S. (2023) "Reimagining the remote simultaneous interpreting interface to improve support for interpreters", in O. Ferreiro Vázquez / A.T. Vara / S. Lima Gonçalves Araújo (eds) *Technological Innovation for Language Learning, Translation and Interpreting*, Berlin, Peter Lang, 227-246.
Saeed M. A. / Rodríguez González E. / Korybski T. / Davitti E. / Braun S. (2022) "Connected yet distant: an experimental study into the visual needs of the interpreter in remote simultaneous interpreting", in M. Kurosu (ed.) *Human-Computer Interaction. User Experience and Behavior*, Cham, Springer, 214-232, <https://doi.org/10.1007/978-3-031-05412-9_16>.
Saeed M. A. / Rodríguez González E. / Korybski T. / Davitti E. / Braun S. (2023) "Comparing interface designs to improve RSI platforms: insights from an experimental study", in *Proceedings of the International Conference on Human-informed Translation and Interpreting Technology 2023*, INCOMA Ltd., Shoumen, Bulgaria, 147-156, <https://doi.org/10.26615/issn.2683-0078.2023_013>.
Saldanha G. / O'Brien S. (2014) *Research Methodologies in Translation Studies*, London, Routledge, <https://doi.org/10.4324/9781315760100>.
Sannholm R. / Risku H. (2024) "Situated minds and distributed systems in translation: exploring the conceptual and empirical implications", *Target, International Journal of Translation Studies* 36/2, 159-183, <https://doi.org/10.1075/target.22172.san>.

Van Cauwenberghe G. (2020) *Étude expérimentale de l'impact d'un soutien visuel automatisé sur la restitution de terminologie spécialisée*, unpublished MA Thesis, Universiteit Ghent at Ghent.

Wang X. / Wang C. (2019) "Can computer-assisted interpreting tools assist interpreting?", *Transletters. International Journal of Translation and Interpreting* 3, 109-139, <https://www.uco.es/ucopress/ojs/index.php/tl/article/view/11575>.

Wang X. / Wang B. / Yuan L. (2025) "The function of ASR-generated live transcription in simultaneous interpreting: trainee interpreters' perceptions from post-task interviews", *Humanities and Social Sciences Communications* 12/1, 1-12, <https://doi.org/10.1057/s41599-025-04492-w>.

Xu R. (2018) "Corpus-based terminological preparation for simultaneous interpreting", *Interpreting* 20/1, 29-58, <https://doi.org/10.1075/intp.00002.xu>.

Yuan L. / Wang B. (2023) "Cognitive processing of the extra visual layer of live captioning in simultaneous interpreting. Triangulation of eye-tracked process and performance data", *Ampersand* 11, 1-9, <https://doi.org/10/gsc8x8>.

Zhang X. / Corpas Pastor G. / Zhang J. (2023) "Videoconference interpreting goes multimodal: some insights and a tentative proposal", in G. Corpas Pastor / B. Defrancq (eds) *IVITRA Research in Linguistics and Literature*, Amsterdam, John Benjamins, 169-194, <https://doi.org/10.1075/ivitra.37.07zha>.