Du, Z., & Muñoz Martín, R. (2025). Interpreters in the loop: Situating CAI tool assessment. *Linguistica Antverpiensia, New Series: Themes in Translation Studies, 24*, 86–113.

Interpreters in the loop: Situating CAI tool assessment

Zhiqiang Du (corresponding author)
Alma Mater Studiorum Università di Bologna
zhiqiang.du2@unibo.it
https://orcid.org/0000-0002-6659-250X

Ricardo Muñoz Martín
Alma Mater Studiorum Università di Bologna
ricardo.munoz@unibo.it
https://orcid.org/0000-0001-6049-9673

Abstract

This study proposes new methodological features to evaluate computer-assisted interpreting (CAI) tools in remote simultaneous interpreting by comparing performance with different tools when compiling glossaries and consulting them during tasks. The traditional information-processing paradigm is superseded by a situated cognition framework that examines multimodal multitasking and interaction in authentic settings. The research project involved a mixed-methods pretest–post-test design across three rounds of data collection (baseline and two post-test rounds). Chinese L1 and English L2 interpreting trainees (*n* = 22) were split into an InterpretBank group (experimental) and an Excel group (control). Term accuracy, speech fluency, cognitive effort, and overall quality were measured through both qualitative indicators and quantitative metrics. CAI tools were found to enhance interpreter performance, but their effectiveness depends on the usage environment, task complexity, and individual adaptation strategies. This study addresses two significant gaps: the need for more robust evaluation methods for CAI tools and the limited research on their use with distant language pairs. Our findings underscore the importance of keeping interpreters involved in tool development and evaluation processes. The methodology and results offer practical insights into CAI tool design, interpreter training, and professional practice in an increasingly technology-dependent field.

Keywords: remote simultaneous interpreting, computer-assisted interpreting tools, situated cognition, mixed methods, task dynamics

1. Background

Managing multilingual terminology poses significant challenges for interpreters, which is why they have long relied on various aids, ranging from traditional tools, such as pen and paper, to modern resources such as electronic dictionaries and instant messaging. An evolving understanding of computer-assisted interpreting (CAI) tools is refining and narrowing the concept to dedicated software packages. Fantinuoli (2023) has identified three generations of such CAI tools: the first generation replaced traditional glossaries, the second introduced features such as automatic glossary extraction, and the third incorporated artificial intelligence (AI) features such as automatic speech recognition (ASR) to support automatic term retrieval.

Even though CAI tools hold promise for improving the compilation and use of glossaries (Fantinuoli, 2018, p. 4), they have yet to see widespread adoption among interpreters. Instead, many still prefer to use spreadsheets such as MS Excel for preparing and managing their glossaries (Jiang, 2013; Woesler, 2021), because Excel is both affordable and easy to use. Professionals and trainees alike easily update their Excel glossaries, share them with colleagues, and integrate them into CAT tools (Matis, 2010). However, Excel has several limitations, such as passive query functionality, slow performance with large files, and limited display options (Amelina & Tarasenko, 2020; Yang, 2021). These shortcomings have in the past decade prompted developers to adopt more specialized CAI tools such as InterpretBank (Fantinuoli, 2016).

We chose InterpretBank (version 8.8) for our study, as it represents third-generation CAI tools in Europe, where it is very popular (Guo et al., 2023) and "the tool students are most often introduced to" (Prandi, 2020, p. 4). Apart from the ASR feature, InterpretBank offers four primary modes for glossary creation and term management:

(1) *Edit* handles terminology management and suggests translation from Wikipedia, MyMemory, and Bing.
(2) *Docs* processes many file formats (.docx, .pdf) for extracting both standard and smart terms.
(3) *Memo* is a flashcard system for bidirectional translation practice.
(4) *Booth* enables rapid term searches with an adjustable three-second reset interval as the default.

InterpretBank (IB, henceforth) offers three methods for creating glossaries: unaided term entry, extraction from web pages, and document import. Document import includes two options: comprehensive extraction of all terms or selective extraction based on relevance algorithms.

CAI tools with such features are geared toward simplifying term management and supporting interpreters by enhancing their accuracy. Reducing the need for complex consultations should also foster smoother and uninterrupted speech delivery. This study was part of a broader PhD project by Du (2024) on CAI-tool use by Chinese-speaking interpreting trainees. Here, we focused on a critical prerequisite question that ultimately contributed to the larger quest: Can we improve how we evaluate the impact and quality of CAI tools on interpreter performance?

In answer to that preliminary question we developed a methodology with innovative approaches. In what follows, we build the argument bottom-up and then we briefly report on some results that Du (2024) obtained with it.

2. Introduction

Both CAT and CAI tools use AI, including natural language processing (NLP) and machine learning techniques. CAI tools add speech recognition and synthesis but otherwise they generally adopt CAT tool strategies. For instance, speech-to-speech translation has traditionally involved three sequential steps: speech-to-text conversion, (written) machine translation (MT), and text-to-speech synthesis. Now, end-to-end speech translation systems are emerging which bypass intermediate text representations by directly translating speech in one language into speech in another, using a single unified model. This can reduce both latency and compounding errors from separate modules and can possibly lead to more natural and efficient translations. Yet, despite this progress, the development of CAI tools continues to build upon the mature field of CAT.

CAT tools use translation memories to store and retrieve text segments through fuzzy matching, which enables content reuse and efficient handling of repetitive content, even in batch mode. Adaptive glossaries standardize terminology whereas quality checks enforce style guides. Integrated AI-powered MT algorithms provide timely suggestions. These features blur the boundaries between translation and post-editing. CAT tools also support remote teamwork and project management for planning, tracking, and resource allocation. Viewed from some distance, the core benefits of CAT support seem to be that the tools enhance text consistency and reduce errors, improve translator efficiency and teamwork, and enable language service providers to meet shorter deadlines, handle higher volumes, and reduce costs. As we shall see, CAI tools are claimed to increase interpreter accuracy for terms, names, and numbers.

However, CAI tools are still taking their first steps and so a key opportunity exists to remedy a longstanding oversight in the development of CAT tools: the insufficient involvement of users in the design process. Now is the moment to prioritize keeping interpreters *in the loop.* This expression refers to a strategy in software development where the target users (in this case, interpreters) are actively involved throughout the design, development, and evaluation process. This participatory approach ensures that the application really meets the needs of its users by incorporating their feedback, expertise, and real-world usage patterns. And yet, the development of CAI tools often lacks sufficient interpreter involvement and this, coupled with the limitations of the prevailing cognitive paradigm, hinders the development of truly effective and user-centered CAI tools.

The testing methodologies for both CAT and CAI tools remain under-explored. On closer examination, they often seem to build implicitly on the information-processing (IP) paradigm, which portrays the mind as a computer and envisages information processing as a sequence of input–output operations in working memory (Atkinson & Shiffrin, 1968; Baddeley & Hitch, 1974; Palmer & Kimchi, 1986; see also Muñoz, 2016a). Therefore, communication and cognitive tasks are broken down into discrete sequential steps, much in the way a computer program is. The IP model has deeply shaped Cognitive Translation & Interpreting Studies (CTIS), which fosters the study of individual cognitive *acts* in encapsulated task-specific input–output operations, as opposed to cognitive *events,* approached from a social perspective at a higher level of granularity (cf. Chesterman, 2013 vs Muñoz, 2016b). Research has often zeroed in on aspects such as working-memory capacity (Boos et al., 2022)—viewed as a stable property of the mind or a computer system—and information consultation (Enríquez & Cai, 2023; Lu et al., 2022) as the flow of information in and out of the biological or silicon processor.

However, the computer metaphor for the IP paradigm fails to capture brain complexity, which surpasses that of current artificial neural networks (Courellis et al., 2024; Muñoz, 2025; Ostrow & Fiete, 2024). Its sequential notion of processing cannot account for the brain's parallel operations and interconnected cognitive functions. This limitation has an impact on multilectal mediated communication research, where simultaneous processing, environment, and experience matter, and problem-solving is adaptive and non-linear (Risku, 2020; Stadler et al., 2024). Early IP models also assumed that information flow was unidirectional (Vasey et al., 2010), leading SI research towards reductive linear transfer models.

Whereas the computational model influenced early SI research, recent studies emphasize the multimodal, dynamic, and interactive nature of SI (e.g., Li & Chmiel, 2024; Tiselius, 2018; Xiao et al., 2023). Research also indicates the cognitive advantages associated with interpreter training (Henrard & Van Daele, 2017). Tzou et al. (2012) note faster L2 reading and better memory spans, whereas Ouyang (2018) identifies unique information-processing patterns. These findings are often framed as improvements in mental machinery rather than as adaptive skills. Alternative CTIS frameworks grounded in situated cognition—for example, cognitive translatology (Mellinger, 2023; Muñoz, 2010, 2023; Muñoz & Tiselius, 2024), socio-cognitive studies (Risku, 2020; Risku & Rogl, 2020), and human–computer interaction (O’Brien, 2023)—deal with SI complexity more effectively.

Cognitive translatology suggests that many task difficulties arise from cognitive resource management rather than from capacity limitations. This calls for an examination of interpreters’ entire performance and interaction with CAI tools, including the temporal analysis and multitasking features. Moving beyond the linear sequence and isolated minimal phenomena of the IP paradigm reveals varying responses at different task stages and shows how human beings steer and adapt their cognitive processes based on experience and circumstances. A cognitive-translatological (i.e., situated) approach therefore favors naturalistic environments with realistic input which promote realistic data collection and unobtrusive methods (cf., e.g., Mellinger et al., 2025). This shift from controlled environments makes possible larger samples and better ecological validity.

As for human–computer interaction (HCI), at its core, it examines the ways in which technologies and human actions create cohesive systems within cognitive ecosystems. Grinschgl and Neubauer (2022) emphasize the crucial role that technology plays in enhancing cognitive abilities. This is particularly relevant in SI contexts, where cognitive offloading—transferring cognitive tasks to the environment—reduces cognitive effort and shows how individuals can enhance their cognitive capabilities by using tools and resources in their environments (Grinschgl et al., 2021; Morrison & Richmond, 2020).

Our methodology involves both qualitative and quantitative assessments of the effectiveness of CAI tools, focusing as it does on the ways in which Chinese-speaking interpreting trainees use IB before and during SI tasks. This entails performance metrics such as term retrieval accuracy, speech fluency, and cognitive effort so as to provide a comprehensive understanding of the role of technology in enhancing cognitive processes in interpreting. First, we will examine the literature on interpreters’ cognition and performance while using CAI tools.

3. Literature review

Research on CAI tools builds on CAT research and can be classified into two main areas: productivity—or workflow optimization—and user experience. Studies on user experience examine behavioral changes, skills development, professional engagement, and resistance to adoption.

Research on the effectiveness of CAI tools yields positive results across several kinds of potential problem triggers. Pérez (2018) examined the impact of a corpus management program on vocabulary preparation among 27 final-year translation and interpreting students and showed that those using the tool consistently performed better across topics. Frittella (2023) found the use of a CAI tool effective for simple items such as acronyms, but less so for complex terms, where the level of accuracy dropped to between 45% and 79%. Challenges included pronunciation errors for rare entities and misunderstandings of numerical magnitudes. Defrancq and Fantinuoli (2021) found 96% precision in supported number interpretation; and while the user feedback regarding its ergonomic design was positive, significant benefits were observed in only two out of six participants.

Atabekova et al. (2018) and Prandi (2018, 2020) analyzed the usability and effectiveness of CAI tools in various European language pairs and noted their integration into professional interpreting workflows. Their findings point to improvements in interpreting quality and reductions in cognitive effort (see also Corpas, 2021; Prandi, 2023)—although the adoption patterns vary by region and also by language pair. We return to this point below. IB has shown particularly strong results in the booth. Advanced CAI tool functions, such as automated term extraction and ASR, benefit interpreters in digital environments (Defrancq & Fantinuoli, 2021; Frittella & Rodríguez, 2022; Tammasrisawat & Rangponsumrit, 2023).

Time latency, as measured through ear-voice span (EVS), remains another crucial indicator. Timarová et al. (2011) found a 4.1-second average EVS among 16 professional interpreters. Su's (2020) English-to-Chinese study revealed EVS averages of 0.93–3.25 seconds for novices and 1.17–3.93 seconds for professionals. This suggests that experience enables *longer* latencies. Fantinuoli and Montecchio (2023) report that the system can tolerate latencies of up to three seconds without degrading performance significantly—a crucial finding for SI scenarios. However, EVS may not capture all the relevant temporal aspects, and Zhou et al. (2021) advanced this understanding through a comparison of sentence-initial and sentence-final EVS. This still lacks a complete HCI multimodal perspective, though. For instance, Chmiel and Lijewska (2023) measured eye-voice span (IVS) in sight translation and found averages of more than 8 seconds among conference interpreters.

Term-rendering accuracy remains a primary quality indicator in SI research, particularly for evaluating CAI tools. While it represents just one factor in quality assessment, its measurement has become standard in the field (Gieshoff & Albl-Mikasa, 2022). Quality assessment in SI typically falls into two dichotomies: quantitative vs. qualitative and rubric-based vs. holistic approaches (and their interaction; Crisp, 2016; Phung & Michell, 2022; Sadler, 2009). Analyses usually cluster around *accuracy* and *fluency*—sometimes they also *target speech quality*—and recently other factors seem to be coming center stage, including emotions, intonation, and speaker accents (Ehrensberger-Dow et al., 2020).

Yet, legitimate concerns prevail about the alignment of expert evaluations with end-user perceptions. In evaluating interpreter performance, Chen et al. (2022) found that holistic approaches yield more reliable scores, regardless of the specific interpreting instructions. Han (2021) advocates a multi-method scoring approach used by several US testing agencies. His approach combines the itemized analysis of overall quality using a Likert-type scale, particularly in high-stakes interpreting scenarios.

Despite growing evidence about their potential benefits, other questions remain about the integration of CAI tools into interpreting practice. The conditions of use, for instance, may also be important. Prandi (2023) tested CAI tool-supported SI performance under three conditions: with ASR (sim-ASR), without ASR (no-ASR), and traditional PDF glossary consultation. Terminological accuracy reached 96.3% in the sim-ASR condition, compared to 86.26% with no-ASR, and 78% with the PDF glossary. The differences between sim-ASR and both no-ASR and PDF suggest that automatic term recognition enhanced accuracy more than manual lookups. This suggests that interpreters' glossary consultation methods do indeed affect their term-rendering precision.

Another example is glossary compilation strategies. IB's automated extraction autonomously identifies terms from source texts but deprives glossary compilers of a deeper understanding of terms in their contexts. In contrast, a *read-first* method enhances contextual understanding and connects new information to prior knowledge, promoting deeper comprehension (Kintsch, 1998, 2013). Additionally, the choice and implementation of digital tools can shift researchers' focus from essential knowledge acquisition to the mechanics of tool operation. Most studies measure term accuracy in isolation rather than examine its relationship to overall interpreting quality and the management of cognitive effort. Interpreters develop their preparation routines based on background knowledge, language proficiency, and experience (Chen et al., 2021; Jiang, 2013).

In brief, whereas AI-powered tools may reduce cognitive effort in many areas, they might limit deeper content engagement (Stadler et al., 2024). Still, when interpreters find CAI tools effective, they can better select and integrate terminology more effectively into their workflow. This has the potential to improve their performance and outputs in the case of remote SI. Yuan and Wang (2023) report that the proper tool integration of CAI tools can reduce cognitive effort in SI tasks by up to 30%.

Research on the effectiveness of CAI tools indicates that they share some weak points with other SI research strands. In the first place, CAI tool studies often use small samples and focus on European languages, with limited research being conducted on English–Chinese interpreting, despite the growing demand for this combination. Then, typically, studies employ controlled experimental designs grounded in the IP paradigm, which may not reflect real-world usage patterns. Furthermore, many experiments restrict access to CAI tools to isolated term-finding instances and overlook other possible areas of interest. Moreover, the documentation processes in interpreting involve the complex integration of various information sources while maintaining working memory. This presents particular challenges for inexperienced interpreters.

Finally, Guo et al. (2023) find that IB remains less well known among Chinese interpreters, among whom online dictionary use predominates (Wan & Yuan, 2022). Costa et al. (2017) and Liu (2022) point out the limitations of using software such as IB as being financial costs and learning curves. These limitations mirror the early challenges faced during the implementation of CAT tools. Research in the Chinese context has been expanded through three master's theses, though—those of Ge (2023), Zhang (2021) and Zhou (2019). They examine the effects of IB on Chinese interpreting trainees' performance. Ge's (2023) pretest–post-test study of eight MA interpreting trainees reveals a 16.7% improvement in their term rendering accuracy and a 23.1% improvement in their term retrieval efficiency, as verified by four professional interpreters who checked their glossary-term matches.

This methodological study advocates comparing InterpretBank with Excel by using cognitive translatology, a situated-cognition framework. Such an approach considers the ways in which interpreters actually interact with these tools in authentic settings, and accounts for contextual factors and adaptive behaviors. Instead of viewing tool use as a simple input–output process, the present study examines the way interpreters integrate tools into their broader cognitive strategies and workflow patterns. By examining these aspects, we hope to better understand the ways in which interpreters actually use these tools and the factors that influence their effectiveness in real-world contexts. This understanding could inform both tool development and interpreter training approaches.

4. Methodology

The study used a mixed-methods, pretest–post-test design with three data-collection rounds to study the differential effects of using a CAI tool for glossary compilation and use at remote SI (RSI). Ecological validity was a crucial consideration in our study setup. We chose telephone interpreting because the typical lack of visual contact with the speaker, the frequent absence of a booth mate, and null feedback from the audience mirrored the conditions of our study, which enhanced its verisimilitude. Round I (pretest) was used as the baseline, to establish the levels of the participants' performance. After round I, the participants were split into an Excel or XL group (control) and an InterpretBank or IB group (experimental), depending on the tool they consistently used both for glossary compilation and for consulting the glossary. The two post-test rounds of data collection aimed to capture the impact of novelty in CAI tool use. During the initial uses of a CAI tool, the participants might struggle with one more source of attentional demands, such as remembering how the tool works and integrating it into the action dynamics. Learning and habituation might also lead to different results.

In round II, either IB or XL use was assigned by the researchers. In round III, the participants were allowed to use whichever tool they wanted.

4.1 Participants

This study had two kinds of participants: interpreters and raters. Twenty-two Chinese L1 speakers enrolled in master's programs in interpreting at three leading Chinese universities were recruited as interpreters. We used a convenience sampling method and had no prior acquaintance with any of the participants. The sample was balanced for gender and had an age range of 22–34 (avg. age 24.7 ± 2.9). All of the participants had English as their L2 and at least two semesters of SI training behind them, so they could be assumed to have a basically similar understanding and command of interpreting techniques, which would help to reduce differences in skill level that could affect the results.

The interpreting participants were matched in pairs according to their RSI performance in data-collection round I and each pair was then split into the IB group (*n* = 12) and the XL group (*n* = 10). The assignment was quasi-randomized in that previous experience with the CAI tool was taken into account: the six participants with previous experience with IB were placed in the XL group. In round III, when the participants could choose which tool to use, two participants switched tools, one in each direction. They were both excluded from the analysis to ensure that the results accurately reflected consistent tool use throughout the study.

As for the raters, we selected five PhD candidates through convenience sampling. We did not know them before and we used a form to collect their socio-demographic information. All of them were Chinese L1 speakers with English as their L2. They had obtained competitive fellowships in the United Kingdom, Beijing and Shanghai, were researching the cognitive aspects of interpreting, and had some professional experience in conference interpreting. These features were requirements, to ensure that they had developed their own SI quality criteria, based on their training and experience. Chen et al. (2022) and Han (2021) provide evidence on the reliability of holistic assessments over traditional rubric-based evaluations. Hence, instead of using rubrics, we adopted a holistic approach to assessment and asked the raters to evaluate the RSI quality intuitively while also allowing for the analysis of their rating patterns and performance metrics later. This approach supports ecological validity because conventional interpreting quality assessment activities such as note-taking and re-listening are somewhat artificial, which may be useful for interpreting training purposes. However, this means that we depart from a listeners' perspective, since listeners' reception can differ substantially between controlled evaluation settings and real-world interpreting scenarios (Guo et al., 2024; Murr, 2018).

The research protocol received approval from the University of Bologna's research ethics committee. We made sure that all the participants understood their role in the study and its objectives, and we obtained their signed informed consent. The interpreting participants were compensated for their time with an amount of 600 RMB (approx. 77 euros), following ethical guidelines for compensation. The raters received no financial compensation. To ensure their anonymity, all of the participants were identified with unique self-generated codes (Direnga et al., 2016).

4.2 Materials

The source speeches for the RSI tasks came from Dr. Huberman Lab Podcast episodes, so that the content was both engaging and relevant.[1] This podcast website provides high-quality, single-speaker recordings with consistent audio, uniform scientific content difficulty, and a structured format that renders it ideally suited for studies needing controlled variables in popular science materials. Thematically matched episode pairs with consistent speakers covered (1) time perception and dopamine, (2) the immune system, and (3) emotions. These pairs were randomly assigned to different study rounds, but their order was kept invariable for all the participants. This could cast some doubts on the results, but randomizing the source speeches would have compromised the remote data collection. Out of each pair, one was assigned to the participants' glossary compilation, and the other to the RSI task.

The episodes were downloaded and transcribed automatically using MS Stream, part of MS 365. We then corrected the transcripts and edited the contents to remove irrelevant elements, such as advertisements, in order to create structured source texts that are typical in SI. All of the edited transcripts underwent linguistic analysis to ensure consistency across rounds (Du, 2024, p. 31), as shown in Table 1. From these transcripts, for the RSI tasks, we selected specialized terms (for a non-specialized audience) as *potential problem triggers.* This was done with a careful strategy to avoid excessive frequency and personal biases. Our principled selection and validation also aimed to reflect regular and specialized language usage within the domain.

We first extracted keywords from the podcast transcripts assigned to RSI tasks using AntConc (Anthony, 2022), a freeware corpus analysis toolkit for text analysis.[2] Second, we used eight specialized terms—four unigrams and four plurilexical expressions—as seeds to build full-text corpora from online resources using BootCaT (Baroni & Bernardini, 2004), a free web crawler that extracts information from online sources.[3] We configured tuples with two seeds for combination queries which enabled BootCaT to gather the relevant documents and create a customized, domain-specific English corpus for each transcript.

Table 1

Linguistic features of texts in glossary tasks and RSI tasks

task	**glossary texts**			**booth texts**		
round	**I**	**II**	**III**	**I**	**II**	**III**
word count	8432	8247	8228	1686	1673	1752
complex word count*	1303	1044	1068	228	239	181
complex word share	15.45	12.66	12.98	13.31	14.28	10.33
avg. word frequency for content words	2.37	2.29	2.4	2.28	2.29	2.38
avg. word frequency for all words	3.10	3.06	3.13	3.07	2.99	3.13
full words % (lexical density)	50.66	49.76	47.31	51.43	52.6	49.29
nouns %	26.22	24.58	22.44	29.36	27.62	28.01
adjectives %	7.15	7.98	6.71	7.12	8.49	6.90
verbs %	10.40	10.53	11.06	10.68	11.48	9.30
sentence count	481	398	405	86	84	84
passive sentences count	59	66	85	25	19	17
passive sentences %	12.27	16.58	20.99	29.06	22.62	20.48
sentence length, number of words, mean	11.89	13.83	16.1	8.9	8.78	9.58
number of long sentences**	110	128	108	19	18	20
syllable count				2558	2383	2470
duration (s)				776.35	793.75	777.53
speech rate (syllables per second)				3.29	3	3.18

* Complex words are those with three or more syllables.

** As a rule of thumb, sentences with more than 25 words are considered long.

Next, we used AntConc again to refine our keyword selection, using the ukWaC corpus (Baroni et al., 2009) as a benchmark for medium-frequency word usage, which confirmed that the domain terms were contextually appropriate. Finally, we generated prioritized terms with high keyness values related to the corresponding speech topic and assessed the overlaps between the keywords from AntConc and the terms in the raw speech transcript. To ensure an even distribution, we selected 33 terms as potential problem triggers: these comprised the most frequent 11 unigrams, 11 bigrams, and 11 trigrams.

Potential problem triggers were strategically embedded in the transcripts, following Frittella (2022) and Prandi (2017). We manipulated the sentences in two ways: 23 contained a single problem trigger each, and five contained two problem triggers each, totaling 28 target sentences distributed throughout the source speech. The transcripts were longer, so problem triggers were embedded at a rate of one in every two sentences, that is, the target sentences were not always consecutive and could be interspersed with sentences lacking problem triggers. This ensured a balanced distribution that averaged approximately three triggers per minute.

To enrich our analysis, we included two repetitions each of one unigram, one bigram, and one trigram, which totaled six repetitions per transcript. The repetitions served as stimuli with which to observe the participants' ability to remember and accurately interpret non-new terms; they focused on the participants' memory rehearsal during RSI tasks. This secondary goal aimed to explore the possibility of studying cognitive dynamics during task performance, that is, how subtasks interacted with each other as the task unfolded. Breedveld's (2002) insights into translation's dynamic processes and the temporal indicators of translation as vital proxies for understanding cognitive translation processes was another source of inspiration for our approach to capturing individual variations and cognitive processes during task engagement.

Nevertheless, we included a minimal number of items simply to assess whether further exploration is warranted. If it shows promise, future studies should incorporate this approach with more stimuli under different conditions. The transcripts included other repetitions, but they were not considered in this study. Consequently, each RSI transcript contained 39 strategically placed problem triggers, 33 first-time terms (or *first-timers*), plus six repetitions. The latter will be analyzed separately for recall, but no distinction between them will be made when analyzing potential problem triggers. Because of the way the data was collected, we would have been unable to determine the actual memory recall, very much like the old methods cannot determine whether CAI tools were really used at a certain point.

An English L1 professional interpreter and interpreting trainer reviewed and modified the enriched RSI transcripts into scripts suitable for SI tasks of less than 15 minutes. In Chinese simultaneous interpreting programs, SI practice sessions are typically capped at 15 minutes, limiting the input to about 1,800 words at a speaking rate of 120 words per minute. We considered that setting RSI tasks at approximately 13 minutes would ensure that the participants were slightly challenged but not overly stressed, which we thought would foster an optimal level of arousal. The set of potential problem triggers serves both as a benchmark for assessing the participants' RSI term accuracy and their recall capacity, and as an indicator of the way IB participants respond to these challenges.

Three American English L1 speakers recorded the scripts in MP3 format for telephone interpreting. We then analyzed the acoustic properties of the recordings to ensure their comparability, including the syllable count, duration, and speech rate (Table 1), using Praat software. A 30-minute initial blank segment was added to the recordings so that the participants could use that time to review pronunciation and annotate the master glossary. Two alert signals marked the end of the preparation, followed by recorded task instructions. The source speech began after the third signal and ended with an additional ten seconds for completing the interpretation.

4.3 Data-collection tools

The data collection was constrained, due to the international travel restrictions imposed by the COVID-19 pandemic. In any event, we opted for remote data collection because it also aligns with contemporary research practices aimed at expanding the scope and diversity of the participants (Dolmaya, 2023; Rodd, 2024). The data was collected using MS Stream, Pynput Keylogger, and TechSmith Capture.[4] These tools minimized any interference with the participants' tasks and enabled the key metrics to be tracked in ~~a~~ detail. The participants accessed the audio files for playback through MS Stream via shared links that were restricted to their email addresses, to ensure confidentiality. To maintain consistent listening conditions, the participants could not alter the audio reproduction by pausing, stopping, re-running, or adjusting the speed.

We keylogged the participants' keyboard actions during the glossary compilation and the RSI tasks with Pynput, an open-source, Python-based keylogger compatible with both macOS and Windows. It records each keystroke with millisecond precision and runs in the background unobtrusively, enabling accurate measurement of the response times. The software is also compatible with the QWERTY keyboard layout, which all of the participants used, and supports Chinese character input via pinyin transcription, capturing all variations accurately.

TechSmith Capture was used for synchronous screen and audio recording. This free screen-capture software allows simultaneous recording of screen interactions and audio via a headset microphone. Integrating these three tools provided us with complementary data streams: MS Stream, for controlled audio delivery and transcription; Pynput, for precise keystroke analysis; and TechSmith Capture, for comprehensive behavioral observation. This methodological framework enabled detailed analyses of interpreter behavior, cognitive processes, and performance patterns in RSI scenarios.

Moreover, the participants followed a standardized setup procedure to install (first time) and activate the Pynput keylogger and TechSmith Capture before each session. For the RSI tasks, the participants enabled TechSmith Capture's audio-recording feature. They were also advised to make sure that the tools were operational prior to beginning each task.

4.4 Constructs and indicators

We analyzed the interpreters' delivery using traditional quality markers (e.g., fillers, repetitions, false starts, self-corrections) and the holistic quality assessments by external raters. We also investigated the interpreters' behavior and the relationship between actions and tasks using novel and established metrics and post-task surveys (see § 3.5). Indicators for both glossary compilation and RSI tasks (see Appendix 1) enable intra-subject analysis, mitigating Simpson's Paradox (Carlson, 2024)—when trends that appear in different groups of data disappear or reverse when such groups are combined. Term recall was measured by proxy, that is, by contrasting the assessed quality of renditions (correct, adequate, wrong, skipped) of first-timers versus their repetitions. For the IB group, we also measured the frequency and accuracy of those terms accessed and correctly rendered through the IB in rounds II and III. The XL participants would not necessarily use the keyboard.

Previous research on information management while at task (e.g., Enríquez & Cai, 2023; Gough, 2023; Lu et al., 2022) tends to concentrate on the use of online resources by participants. In contrast, our study expands on this by incorporating time-related metrics (total task time, time per term), language production measures (term count, diversity rate), and a layered task model (Muñoz, 2014) for a more dynamic understanding of glossary-compilation behavior. Observed behaviors (both keylogged and screen-recorded) were tagged with constructs representing translation behaviors (e.g., *term spotting, search query, search result review*) and grouped into subtasks (e.g., TRANSLATION SEARCH, TERM EXTRACTION). This nested approach captures the complex interplay of actions that occur during glossary compilation. The temporal dimension, derived from timestamps, allows for an analysis of cognitive activities, and the flexible order and potential recurrence of subtasks reveal dynamic behavior.

Our study also examined the interpreters' performance during RSI tasks, analyzing as it did the fluency indicators *(fillers, repetitions, false starts, self-corrections),* accuracy and content measures *(correct, adequate, wrong, skipped terms),* and time spans (see Appendix 1). Adapting Zhou et al.'s (2021) concept of sentence-initial and sentence-final eye–voice spans (EVS1 and EVS2) to apply to oral chunks rather than to sentences, we analyzed the time from the start/end of source-speech bursts to the start/end of their corresponding target outputs.

Inspired by research on eye–voice span (Chmiel & Lijewska, 2023; Su, 2020; Zhou et al., 2021), we also introduced ear–key span (E2K) and eye–voice span (I2V). Specifically, E2K reflects the initiation of a response (from the end of source utterance to the first keyboard action), whereas I2V tracks information integration (from the display of a translation on screen to target utterance). Combined, they capture the ways in which interpreters manage the flow of information and coordinate actions while using CAI tools.

We hypothesize a relationship between E2K and I2V in computer-supported SI, assuming that visual attention shifts to the screen after keystrokes. While typically sequential (E2K preceding I2V), these spans may not always co-occur and can be reversed. Positive I2V values indicate that the translation appeared before vocalization; negative values flag that vocalization began before the translation appeared on screen, possibly indicating predictive behavior. Analyzing these time spans allowed us to study these dynamics and understand the ways in which interpreters manage their subtask interactions, multitasking, and cognitive effort.

Drawing on Muñoz and Apfelthaler's (2022) similar distinction for keylogged interkeystroke intervals, we subcategorized silent or empty spans in RSI as *lags* (< 200 ms, excluded from analysis), *bumps* (200–600 ms), and *respites* (> 600 ms). Whereas, for translating, respites are calculated based on intra-word median time spans, in sight interpreting/translation and SI, the threshold is fixed and based on listeners' noticing the gap in the flow of delivery. Longer spans, particularly respites, reflect the pressure to maintain fluency (Ho, 2021) and they offer insights into cognitive effort, possibly reflecting the struggles experienced with pacing or multitasking. These measures are categorized separately for clarity, but they are interconnected: fluency and time pertain to rendering the target speech, while accuracy and content link source speech to output.

4.5 Procedures

Each data-collection round included two main tasks: (1) a glossary-compilation task lasting 2.5 hours, and (2) a 13-minute RSI task. After round I, the IB group then received online training on IB use during a 2.5-hour session that covered the features and procedures for term selection, glossary compilation, and the use of booth mode. The features of automatic term retrieval from the glossary through ASR and parallel document extraction were intentionally excluded. This was done to avoid potential confounders in the analysis of the participants' behavior regarding glossary use, and to minimize any additional complexity. The participants had a week in which to practice using IB before round II of the data collection started and they were given video tutorials to watch, although this was not strictly controlled to avoid intrusive monitoring. In parallel, the XL group attended a workshop on search techniques and digital resources. Both training sessions were recorded and shared afterward with both groups.

During the glossary task, the IB participants could choose either manual, automatic, or mixed extraction methods. As with any activity—though, perhaps more for those that entail expectations of future information reuse—glossary compilation could leave memory traces. Such recollections might vary between those who choose manual and those who choose automatic methods—for example, participants might remember terms that were particularly difficult to reformulate in the glossary while individual differences might affect their performance anyway. Some participants might not excel at glossary compilation, which could influence their simultaneous interpreting performance, even if they were generally skilled at the booth. We wanted to separate these two effects, retaining only the first of them. In addition, the participants compiled their glossaries using a speech similar in several respects to the one they would interpret, but not really the one they would listen to during the RSI task. As a result, not all of the terms targeted as potential problem triggers might end up being collected in their glossaries.

To minimize the potential impact of glossary quality and compilation skills on RSI performance, the participants submitted their own glossaries to us, and we returned a revised *master glossary* to them. In it we combined all the entries found in at least two individual glossaries plus the 33 problem triggers we had selected for the RSI source speech. Each master glossary contained 95–97 English–Chinese term pairs. The participants received the master glossary 30 minutes before each RSI task and were prompted to review and change the term pairs at will. The source-speech soundtracks regulated the progression of the RSI tasks, with the participants following the prescribed timing and instructions in each soundtrack—from the initial 30-minute period of silence to review the master glossary provided until a designated ending signal.

Following the RSI task in round III, all participants completed a survey. The IB group also completed a survey after round II to capture initial impressions of the new tool. The surveys included questions on general opinions (including performance self-assessment and attitudes towards CAI tools), glossary tasks (preferences for term retrieval tools, pronunciation checking, translation verification, and automatic term extraction), and RSI tasks (reliance on memory, glossary use, CAI tool necessity, and training needs). The IB informants also filled a follow-up survey one year after concluding their data collection, to assess their long-term CAI tool adoption and usage patterns.

Five raters were involved in the evaluation process, performing holistic assessments of the recordings. The 66 recordings (from 22 participants across three rounds) were sampled and randomized. Five recordings per round were selected to represent a range of quality levels and they were also used to analyze inter-rater reliability. Of these 15 audio files, five were randomly selected and placed in identical order at the beginning of each rater's first session, so as to control for initial rating tendencies, such as potential biases related to first impressions. The remaining 51 recordings were randomly assigned to the raters, while ensuring that each recording received three independent evaluations and an even distribution of rater combinations. This resulted in each rater assessing 45–46 files.

The evaluations were completed during eight sittings, with a maximum of six recordings per sitting (approximately one hour and a half of listening). Using PsyToolkit (Stoet, 2010, 2017), the raters categorized the recordings into six quality levels: bad, poor, fair, good, very good, and excellent. They were instructed to base their judgments on their intuitive perception of interpreting quality and to explicitly avoid using any rubrics. Methodological requirements included listening to the full audio recording before making an assessment, no option to re-listen, and the completion of all the recordings of a set within one sitting. These requirements were set to mimic some audience conditions and to minimize fatigue and its effects.

The 39 potential problem triggers were timestamped and synchronized with the source speech timeline in Adobe Audition 2024. The data from various sources—screen recordings with RSI audio, keystroke logs, transcripts, assessment results, and survey responses—were integrated into a synchronized, multimodal timeline or *ethogram*. A universal timeline was created to cross-compare the participants' ethograms, to ensure the reliable analysis of behaviors with millisecond-level precision (for details of the alignment and synchronization, see Du, in press). This allowed for detailed analysis of the participants' behaviors during the glossary and RSI tasks, which facilitated both quantitative and qualitative analyses of the interpreter performance and CAI tool usage.

5. Results

We analyzed how accurately the participants interpreted 39 potential problem triggers using either InterpretBank (IB) or Excel (XL). The XL participants increased their correct interpretations from 22.2% to 37.7%, whereas the IB participants improved from 19.8% to 50.4%. Therefore, the rendition of specialized terms was more accurate using IB (for detailed accuracy results; see Du, 2024). Term-skipping was also reduced: the XL participants, from 68.7% to 50.2% and the IB participants, from 68.6% to 39.9%.

On the one hand, both the accuracy and the omission results align with studies by Defrancq and Fantinuoli (2021), Frittella and Rodríguez (2022), Prandi (2023) and Tammasrisawat and

Rangponsumrit (2023); this supports the advantage of IB’s assistance to simultaneous interpreting. Furthermore, this study extended its results to a distant language pair. On the other hand, the inter-rater reliability regarding RSI quality assessment was low, with Krippendorff’s alpha coefficients being 0.016 across all five raters and 0.166 for subsets of three raters. Further details of the inter-rater reliability and the links to quantitative parameters will be provided as part of an ongoing investigation that adds reassessments (retesting) of some of the recordings one year later.

Our results, based primarily on frequency data and relative frequency comparisons, show clear patterns. Generalization based on this type of analysis is somewhat limited, though, and statistical testing would be a more appropriate way to generalize the results to other groups. Yet, the frequencies are relatively revealing (e.g., a frequency of 70% is much larger than that of the next group), so extensive statistical testing might be excessive in a methodological study.

5.1 Individual glossary and master glossary feature comparisons

We compared the term counts from individual glossaries across three rounds in each participant and also contrasted the aggregated group results. Table 2 displays only three key numbers for each group (IB and XL)—the smallest glossary (min.), the largest glossary (max.), and the average size glossary (AVG.) across participants—to compare them with the master glossary, which contained 95 terms in round I, 96 in round II, and 97 in round III. The *diversity rate* shows what percentage of terms in each individual glossary did not appear in the master glossary. For example, in round I, the XL participant with the shortest glossary (31 terms) had a *diversity rate* of 15.60%. This means that this percentage of terms in this individual glossary was not in the master glossary. The IB participants consistently compiled larger glossaries than the XL participants. Across rounds II and III, the IB participants averaged 99 terms while the XL participants averaged 62. Eight of the ten largest individual glossaries came from the IB participants.

Table 2

Terms in individual glossaries

GR	measure	I	*%*	II	*%*	III	*%*
XL	**min.**	**31**	*15.60*	**60**	*21.83*	**46**	*24.56*
	max.	**94**	*56.11*	**84**	*39.08*	**47**	*31.14*
	AVG.	**59**	*32.41*	**79**	*30.00*	**45**	27.90
IB	**min.**	**56**	*27.81*	**42**	*17.77*	**39**	*21.62*
	max.	**109**	*45.70*	**96**	*29.97*	**82**	*45.56*
	AVG.	**91**	*40.84*	**125**	*36.52*	**73**	*33.37*

Table 3

Glossary task indicators and interpreting quality for the XL and IB groups

Round		**I**		**II**		**III**	
Indicators	**AVG.**	**XL**	**IB**	**XL**	**IB**	**XL**	**IB**
total time (s)	**5313.3**	5749.5	4949.8	4730.1	4554.9	4527.9	4107
term (#)	**76.68**	59.1	91.3	78.8	124.8	44.8	73.3
time per term (s)	**75.81**	97.3	54.2	68.9	41.5	115.1	63.0
diversity rate (%)	**37**	42.8	29.0	45.3	27.7	46.4	27.8

Table 3 summarizes the way the participants in each group performed on the glossary task through averages for overall compilation time, number of terms, time per term, and diversity rates. IB's automatic term-extraction feature led to different patterns occurring between the groups. For instance, the IB participants consistently showed lower diversity rates (round I, 29.0%; II, 27.7%; III, 27.8%) compared to the XL participants. The combination of lower diversity rates and larger term counts in the IB glossaries suggests that IB's automatic extraction works as a one-size-fits-all algorithm. But while this automation makes glossary compilation more convenient and faster per term, it may reduce the variety of terms participants choose and also make individual glossaries more similar to each other.

The next aspect we analyzed was how successfully the participants used IB for term searches and renditions.

5.2 Searches, hits, and correct renditions

We analyzed how effectively the participants used IB's search function during interpretation. Table 4 presents these valid search results—successfully found terms (hits) but excluding unsuccessful searches from this discussion. In round II, the participants performed 242 searches in total. Of these, there were 153 hits, and 134 of these hits led to correct interpretations. Among the hits with correct renditions, 68 involved first-timers (28.1% of all searches and 44.4% of hits) while only five involved repetitions (2.1% of all searches and 3.2% of hits). Their performance improved in round III: hits increased to 72.2% (127 out of 176 searches), and those leading to correct renditions rose to 65.3% (115 out of 176 searches).

Table 4

InterpretBank search counts and percentages in rounds II and III

round	II	%	III	%
total searches	242	*100.0*	176	*100.0*
hits	153	*63.2*	127	*72.2*
correct renditions	134	*55.4*	115	*65.3*
first-timers	68	*28.1*	88	*50.0*
repetitions	5	*2.1*	8	*4.5*

5.3 Terms and searches

We compared the term counts across three sources (Table 5): individual glossaries, IB's automatic extractions (IB glossaries), and the master glossaries (terms selected by at least two participants). To obtain these numbers, we first compiled all individual glossaries from each group into a lemmatized list, removed any duplicates, and then cross-checked this list against both the IB glossary and the master glossary. We categorized the terms based on where they appeared:

- both in the master glossary and the IB glossary;
- in the IB glossary but not in the master glossary;
- in the master glossary but not in the IB glossary;
- neither in the master glossary nor in the IB glossary.

For example, in round II, out of a total of 391 terms—those in at least one individual glossary—27 appeared in both the IB and master glossaries; 157 were only in the IB glossary; 2 were only in the master glossary and 205 appeared in neither.

Table 5

Terms and searches in master and IB glossaries

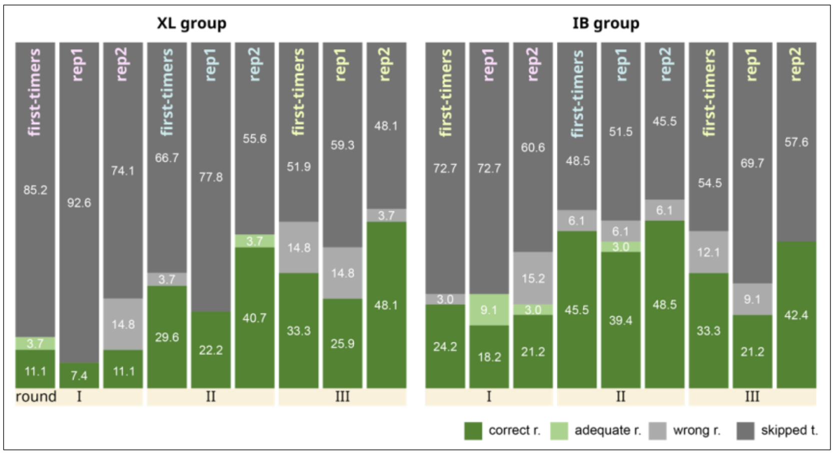
		found in glossary			
item	**round**	**both**	**IB only**	**master only**	**neither one**
terms	II	27	157	2	205
	III	35	106	2	121
searches	II	4	3	22	33
	III	4	0	23	20

5.4 Term accuracy and search patterns across repetitions

We analyzed the ways in which the participants handled first-timers versus term repetitions *(rep1* and *rep2).* Both the accuracy rates and the search patterns varied across rounds. Figure 1 shows that the XL group's accuracy improved progressively across all three rounds.

Figure 1

Repetitions by rounds between XL and IB groups



For first-timers, their correct renditions increased from 11.11% in round I to 29.63% in round II, and 33.33% in round III. A similar upward trend appeared for the rep1 and rep2 terms. However, within each round, the group's accuracy decreased for rep1 terms compared to first-timers, before rising again in rep2 (U-shaped pattern).

The IB group showed a different pattern. Whereas their correct renditions increased from round I to round II, the accuracy declined in round III in both first-timers and repetitions. This suggests that the IB group may have reached a plateau or experienced a slight regression. Table 6 illustrates this by showing search frequency patterns. The number of searches varied both within rounds (across term conditions) and between rounds. For instance, round II first-timers generated six searches, while round III rep2 terms led to nine searches.

Table 6

Search counts for first-timers and their repetitions

round	terms	first-timers	rep1	rep2
II	*neuroplasticity*	2	2	2
	blood vessels	1	1	0
	sympathetic chain ganglia	3	0	1
III	*reward prediction error*	0	1	1
	L-tyrosine	0	2	4
	gut microbiome	5	2	4

6. Discussion

The InterpretBank (IB) participants relied heavily on automatic term extraction, which contributed to their glossaries looking similar. Their individual glossaries contained less diverse terms (round I, 29.0%; round II, 27.7%; round III, 27.8%)—more than 70% of the terms appeared in at least two participants' lists. The automated method saved time per term. The IB participants spent 54.2 s in round I; 41.5 s in round II; and 63.0 s in round III. In contrast, the

XL participants' averages were: round I, 97.3 s; round II, 68.9 s; round III, 115.1 s. However, the IB participants' shorter times at term extraction did not reduce the total compilation time they needed. The IB participants spent less time on each term but needed extra time to review the longer, de-contextualized machine-generated lists. This reliance on automation can affect how well students select and prepare their terms.

We found that half the terms (205/391) in round II were not in standard glossaries but were individual choices, showing diverse participants' needs. The next largest group came from IB's automatic suggestions, which the participants often added without reviewing them. This explains why the IB participants had such long glossaries; but their search patterns indicate that they looked up terms from the master glossary more often when they were not only IB-suggested terms. In fact, most of the terms appearing in the IB glossaries only were nearly ignored, so they were hardly useful. The figures are probably different for other language combinations, but the conclusion is that IB's automatic term extraction needs to improve, and so do the renditions suggested for Chinese.

To its credit, though, IB did help the participants to interpret more accurately, although this varied based on how familiar they were with the terms. In round II, the participants searched for terms 242 times and found what they wanted only 153 times. Of these hits, 134 (55.4% of total searches) led to correct renditions. When the participants found terms and interpreted them correctly, 68 cases involved terms they were seeing for the first time (28.1% of all searches), while only five involved repeated terms (2.1%). The results improved in round III: 72.2% of their searches were hits (127 out of 176 searches); and when they found terms, they interpreted them correctly 65.3% of the time (115 out of 176).

We looked closely at the 33 first-timers and found different patterns between the groups. The IB participants achieved more correct term renditions than the XL participants did. The gap in the correct term renditions between the two groups grew from 6.4% in round II to 12.6% in round III, with IB participants doing almost twice as well. Both groups nevertheless continued to skip many terms in round III, but at different rates: the IB participants skipped about 40% of terms, while the XL participants skipped about 50%. This could be explained by the fact that our participants were students and they might skip terms so often because they know that missing information has no real consequences in training, although having machine support seems to have made them skip fewer terms.

Six terms among our 33 potential problem triggers appeared two more times (that is, three times total), and we analyzed them separately in search of particularities. We also tracked the way the participants handled them across rounds, and the results varied: the participants usually did worse when they saw terms for the second time, and they used IB less often to search for these terms. However, these patterns were not consistent. In round III, the IB participants performed worse with repeated terms compared to round II; however, they searched for these terms more often than before. Strong individual variation and very likely the low number of repetitions precluded the detection of clear patterns in the way the IB informants faced them. In any case, while IB helped, the participants nevertheless relied heavily on managing terms from memory.

7. Conclusions

This study explored a cognitive-situated methodological framework to investigate the documentation behaviors and performance of 22 Chinese L1 and English L2 interpreting trainees using either InterpretBank (IB) or Excel (XL) to compile glossaries and retrieve terms during remote simultaneous interpreting (RSI) tasks. Whereas the sample size and the study's focus on a single CAI tool (IB), language pair and direction (English→Chinese) represent important limitations, the mixed-methods pretest–post-test design developed here, which combines new or modified indicators, constructs, and tools for quantitative metrics and qualitative assessments, offers a robust framework for evaluating CAI tools in different scenarios and lays the groundwork for further research.

Our results support previous findings that CAI tools enhance term accuracy but they also underscore the impact of task complexity and user adaptability in determining their effectiveness. In this sense, they extend but also complement Prandi's (2023) examination of the impact of CAI tools on cognitive load in SI. The sketched repeated measures design implemented through two post-test data-collection rounds enabled us to track the changes in performance after the first use, offering insights into the ways interpreters adapt to technological tools and refine their strategies over time. All participants improved their accuracy over time, although the IB users had a clear, if moderately shrinking, advantage as the XL users caught up with them. Telephone interpreting was helpful in reducing confounders and contributing to a naturalistic environment; however, video interpreting, as with face-to-face interpreting, adds layers of complexity to the task—a consideration for future research.

IB's automatic extraction feature found nearly all of the terms considered relevant by at least two participants, which demonstrated a high level of *recall* in information-retrieval terms. However, IB's one-size-fits-all algorithms contributed to considerably longer term lists compared to those compiled manually; but it did not fare as well in respect of *precision*. Nearly all of the terms offered by IB that did not match those chosen by at least two participants through other means were rarely consulted. Furthermore, the time saved during term extraction was then invested in term verification, which resulted in similar overall times being invested in the glossary compilation. This suggests several lines of further research for CAI tool developers: determining the impact of a concordance feature to provide a context for glossary entries; exploring machine learning techniques and customization strategies to tailor automatic extraction to meet individual user needs; and investigating whether additional unused terms may become distractors, potentially increasing the cognitive demands and slowing down delivery during interpreting.

Interpreters are not necessarily good typists, and the attentionally demanding RSI situation likely explains the large number of unsuccessful searches (typos, misspellings, etc.). The novelty of using a CAI tool may also have contributed to the participants' multitasking and control needs, which implies that CAI tool developers should implement fuzzy matching in searches to support error tolerance. Search difficulties experienced with searches may also explain the occurrence of a few abandoned searches.

A key finding was the strong link between glossary-building and the efficiency of term retrieval during RSI tasks. Although the evidence we gathered is scarce, we observed confirmatory searches (interpreters searching for terms after rendering them), which may indicate the users' long-term planning. We also registered more searches during repetitions in the third data-collection round. This raises questions: does access to terminology via CAI tools influence interpreters' reliance on memory, possibly affecting the development of memory skills traditionally linked to interpreter training? CAI tool developers should explore ways in which the users would be able to suppress repetitions from automatic retrieval functions.

This study emphasizes the importance of keeping interpreters in the loop. CTIS researchers should also be more involved in assessing the functionalities of CAI tools and they should support quality under realistic conditions. Our methods suggest ways of studying interpreters' renditions and performance so as to determine whether these technologies meet the real-world needs of professional practice. By focusing on the situated use of these tools, we have been able to provide a clearer picture of the way technology can enhance interpreter performance.

Acknowledgment

This work was supported by PRIN 2022 (CUP J53D23011340006, Project Code 2022EYX28N_001), under the PNRR plan, funded by the European Union – NextGenerationEU. We thank all participants for their valuable contribution.

References

Amelina, S., & Tarasenko, R. (2020). Using modern simultaneous interpretation tools in the training of interpreters at universities. In A. Bollin, H. C. Mayr, A. Spivakovsky, M. Tkachuk, V. Yakovyna, A. Yerokhin, & G. Zholtkevych (Eds.), *Proceedings of the 16th International Conference on ICT in Education, Research and Industrial Applications: Integration, harmonization and knowledge transfer* (Vol. 1, pp. 188–201). CEUR-WS.org. http://ceur-ws.org/Vol-2740/20200188.pdf

Anthony, L. (2022). *AntConc* (Version 4.1.4) [Computer software]. Waseda University. https://www.laurenceanthony.net/software/AntConc

Atabekova, A. A., Gorbatenko, R. G., Shoustikova, T. V., & Valero Garcés, C. (2018). Cross-cultural mediation with refugees in emergency settings: ICT use by language service providers. *Journal of Social Studies Education Research*, *9*(3), Article 3.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *Psychology of learning and motivation* (Vol. 2, pp. 89–195). Academic Press. https://doi.org/10.1016/S0079-7421(08)60422-3

Baddeley, A., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 8, pp. 47–89). Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 1313–1316. http://www.lrec-conf.org/proceedings/lrec2004/pdf/509.pdf

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–226. https://doi.org/10.1007/s10579-009-9081-4

Boos, M., Kobi, M., Elmer, S., & Jäncke, L. (2022). The influence of experience on cognitive load during simultaneous interpretation. *Brain and Language*, *234*, Article 105185. https://doi.org/10.1016/j.bandl.2022.105185

Breedveld, H. (2002). Translation processes in time. *Target*, *14*(2), 221–240. https://doi.org/10.1075/target.14.2.03bre

Carlson, B. W. (2024). *Simpson's paradox*. Encyclopedia Britannica. https://www.britannica.com/topic/Simpsons-paradox

Chen, H., Wang, Y., & Brown, T. P. (2021). The effects of topic familiarity on information completeness, fluency, and target language quality of student interpreters in Chinese–English consecutive interpreting. *Across Languages and Cultures*, *22*(2), 176–191. https://doi.org/10.1556/084.2021.00013

Chen, J., Yang, H., & Han, C. (2022). Holistic versus analytic scoring of spoken-language interpreting: A multi-perspectival comparative analysis. *The Interpreter and Translator Trainer*, *16*(4), 558–576. https://doi.org/10.1080/1750399X.2022.2084667

Chesterman, A. (2013). Models of what processes? *Translation and Interpreting Studies, 8*(2), 155–168. https://doi.org/10.1075/tis.8.2.02che

Chmiel, A., & Lijewska, A. (2023). Reading patterns, reformulation and eye-voice span (IVS) in sight translation. *Translation and Interpreting Studies*, *18*(2), 213–234. https://doi.org/10.1075/tis.21021.chm

Corpas Pastor, G. (2021). Interpreting and technology: Is the sky really the limit? *Proceedings of the translation and interpreting technology online conference TRITON 2021*, 15–24. https://doi.org/10.26615/978-954-452-071-7_003

Costa, H., Pastor, G. C., & Durán Muñoz, I. (2017). Assessing terminology management systems for interpreters. In G. Corpas Pastor & I. Durán Muñoz (Eds.), *Trends in e-tools and resources for translators and interpreters* (pp. 57–84). Brill. https://doi.org/10.1163/9789004351790_005

Courellis, H. S., Minxha, J., Cardenas, A. R., Kimmel, D. L., Reed, C. M., Valiante, T. A., Salzman, C. D., Mamelak, A. N., Fusi, S., & Rutishauser, U. (2024). Abstract representations emerge in human hippocampal neurons during inference. *Nature*, 632, 841–849. https://doi.org/10.1038/s41586-024-07799-x

Crisp, V. (2016). The judgement processes involved in the moderation of teacher-assessed projects. *Oxford Review of Education*, *43*(1), 19–37. https://doi.org/10.1080/03054985.2016.1232245

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. https://doi.org/10.1075/target.19166.def

Direnga, J., Timmermann, D., Lund, J., & Kautz, C. (2016). *Design and application of self-generated identification codes (SGICs) for matching longitudinal data*. 44th Annual Conference of the European Society for Engineering Education: Engineering Education on Top of the World: Industry-University Cooperation, Tampere, Finland. https://www.sefi.be/proceeding-author/j-direnga/

Dolmaya, J. M. (2023). *Digital research methods for translation studies*. Routledge. https://doi.org/10.4324/9781003083986

Du, Z. (2024). *Bridging the gap: Exploring the cognitive impact of InterpretBank on Chinese interpreting trainees* [Doctoral dissertation]. Università di Bologna. https://doi.org/10.48676/unibo/amsdottorato/11584

Ehrensberger-Dow, M., Albl-Mikasa, M., Andermatt, K., Hunziker Heeb, A. & Lehr, C. (2020). Cognitive load in processing ELF: Translators, interpreters, and other multilinguals. *Journal of English as a Lingua Franca*, *9*(2), 217–238. https://doi.org/10.1515/jelf-2020-2039

Enríquez Raído, V., & Cai, Y. (2023). Changes in web search query behavior of English-to-Chinese translation trainees. *Ampersand*, *11*, Article 100137. https://doi.org/10.1016/j.amper.2023.100137

Fantinuoli, C. (2016). InterpretBank: Redefining computer-assisted interpreting tools. *Proceedings of the Translating and the Computer 38 Conference*, 42–52. ACL Anthology.

Fantinuoli, C. (2018). Interpreting and technology: The upcoming technological turn. In C. Fantinuoli (Ed.), *Interpreting and technology* (pp. 1–12). Language Science Press. https://doi.org/10.5281/ZENODO.1493289

Fantinuoli, C. (2023). Towards AI-enhanced computer-assisted interpreting. In G. Corpas Pastor & B. Defrancq (Eds.), *IVITRA research in linguistics and literature* (Vol. 37, pp. 46–71). John Benjamins. https://doi.org/10.1075/ivitra.37.03fan

Fantinuoli, C., & Montecchio, M. (2023). Defining maximum acceptable latency of AI-enhanced CAI tools. In Ó. Ferreiro Vázquez, A. T. Varajão Moutinho Pereira, & S. L. Gonçalves Araújo (Eds.), *Technological innovation put to the service of language learning, translation and interpreting: Insights from academic and professional contexts* (pp. 213–225). Peter Lang.

Frittella, F. M. (2022). CAI tool-supported SI of numbers: A theoretical and methodological contribution. *International Journal of Interpreter Education*, *14*(1), 32–56. https://doi.org/10.34068/ijie.14.01.05

Frittella, F. M. (2023). *Usability research for interpreter-centred technology: The case study of SmarTerp.* Language Science Press. https://doi.org/10.5281/zenodo.7376351

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), Article 2. https://doi.org/10.54754/incontext.v2i2.21

Ge, T. (2023). *Usability of terminology—assistance in Chinese to English simultaneous interpretation—taking InterpretBank as an example* [Master's Thesis]. Beijing Foreign Studies University. https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFDTEMP&filename=1023063456.nh&v=

Gieshoff, A. C., & Albl-Mikasa, M. (2022). Interpreting accuracy revisited: A refined approach to interpreting performance analysis. *Perspectives*, *32*(2), 210–228. https://doi.org/10.1080/0907676X.2022.2088296

Gough, J. (2023). Individual variations in information behaviour of professional translators: Towards a classification of translation-oriented research styles. *Translation Studies*, *17*(2), 394–415. https://doi.org/10.1080/14781700.2023.2231933

Grinschgl, S., & Neubauer, A. C. (2022). Supporting cognition with modern technology: Distributed cognition today and in an AI-enhanced future. *Frontiers in Artificial Intelligence*, *5*, Article 908261. https://doi.org/10.3389/frai.2022.908261

Grinschgl, S., Papenmeier, F., & Meyerhoff, H. S. (2021). Consequences of cognitive offloading: Boosting performance but diminishing memory. *Quarterly Journal of Experimental Psychology*, *74*(9), 1477–1496. https://doi.org/10.1177/17470218211008060

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. https://doi.org/10.17507/tpls.1301.11

Guo, W., Guo, X., Huang, J., & Tian, S. (2024). Modeling listeners' perceptions of quality in consecutive interpreting: A case study of a technology interpreting event. *Humanities and Social Sciences Communications*, *11*(1), Article 985. https://doi.org/10.1057/s41599-024-03511-6

Han, C. (2021). Interpreting testing and assessment: A state-of-the-art review. *Language Testing*, *39*(1), 30–55. https://doi.org/10.1177/02655322211036100

Henrard, S., & Van Daele, A. (2017). Different bilingual experiences might modulate executive tasks advantages: Comparative analysis between monolinguals, translators, and interpreters. *Frontiers in Psychology*, *8*, Article 1870. https://doi.org/10.3389/fpsyg.2017.01870

Ho, C.-E. (2021). What does professional experience have to offer?: An eyetracking study of sight interpreting/translation behaviour. *Translation, Cognition & Behavior*, *4*(1), 47–73. https://doi.org/10.1075/tcb.00047.ho

Jiang, H. (2013). The interpreter's glossary in simultaneous interpreting: A survey. *Interpreting, 15*(1), 74–93. https://doi.org/10.1075/intp.15.1.04jia

Kintsch, W. (1998). *Learning from text*. Cambridge University Press.

Kintsch, W. (2013). Revisiting the construction–integration model of text comprehension and its implications for instruction. In D. E. Alvermann, N. Unrau, & R. B. Ruddell (Eds.), *Theoretical models and processes of reading* (6th ed., pp. 807–839). International Reading Association. https://doi.org/10.1598/0710.32

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

Liu, J. (2022). The impact of technologies on interpreting: An interpreter and trainer's perspective. *International Journal of Chinese and English Translation & Interpreting*, *1*, 1–8. https://doi.org/10.56395/ijceti.v1i1.14

Lu, S., Xiangling, W., & Shuya, M. (2022). Investigating the relationship between online information seeking and translation performance among translation students: The mediating role of translation self-efficacy. *Frontiers in Psychology*, *13*, Article 944265. https://doi.org/10.3389/fpsyg.2022.944265

Matis, N. (2010). Terminology management during translation projects: Professional testimony. *Linguaculture*, *1*(1), 107–115. https://doi.org/10.47743/lincu-2010-1-0226

Mellinger, C. D. (2023). Embedding, extending, and distributing interpreter cognition with technology. In G. Corpas Pastor & B. Defrancq (Eds.), *IVITRA research in linguistics and literature* (Vol. 37, pp. 195–216). John Benjamins. https://doi.org/10.1075/ivitra.37.08mel

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 10* (pp. 49–68). John Benjamins. https://doi.org/10.1075/rmal.10.02mel

Morrison, A. B., & Richmond, L. L. (2020). Offloading items from memory: Individual differences in cognitive offloading in a short-term memory task. *Cognitive Research: Principles and Implications*, *5*(1), Article 1. https://doi.org/10.1186/s41235-019-0201-4

Muñoz Martín, R. (2010). On paradigms and cognitive translatology. In G. M. Shreve & E. Angelone (Eds.), *Translation and cognition* (pp. 169–187). John Benjamins. https://doi.org/10.1075/ata.xv.10mun

Muñoz Martín, R. (2014). Situating translation expertise: A review with a sketch of a construct. In J. W. Schwieter & A. Ferreira (Eds.), *The development of translation competence: Theories and methodologies from psycholinguistics and cognitive science* (1st ed., pp. 2–56). Cambridge Scholars.

Muñoz Martín, R. (2016a). Of minds and men: Computers and translators. *Poznań Studies in Contemporary Linguistics*, *52*(2), 351–381. https://doi.org/10.1515/psicl-2016-0013

Muñoz Martín, R. (2016b). Processes of what models?: On the cognitive indivisibility of translation acts and events. *Translation Spaces*, *5*(1), 145–161. https://doi.org/10.1075/ts.5.1.08mun

Muñoz Martín, R. (2023). *Traductología cognitiva: Tratado general* (1st ed.). ULPGC Ediciones. https://doi.org/10.20420/1747.2023.743

Muñoz Martín, R. (2025). Do translators dream of electric brains? *Fachsprache, 47*(1–2), 88–108. https://doi.org/10.24989/fs.v47i1-2.4001

Muñoz Martín, R., & Apfelthaler, M. (2022). A task segment framework to study keylogged translation processes. *The International Journal of Translation and Interpreting Research*, *14*(2), 8–31. https://doi.org/10.12807/ti.114202.2022.a02

Muñoz Martín, R., & Tiselius, E. (2024). Written words speak as loud: On the cognitive differences between translation and interpreting. In C. D. Mellinger (Ed.), *The Routledge handbook of interpreting and cognition* (pp. 15–31). Routledge. https://doi.org/10.4324/9780429297533-3

Murr, K. E. (2018). *Quality in interpreting: Interpreters' vs listeners' perception* [Master's thesis]. University of Geneva. https://archive-ouverte.unige.ch/unige:131157

O'Brien, S. (2023). Human-centered augmented translation: Against antagonistic dualisms. *Perspectives*, *32*(3), 391–406. https://doi.org/10.1080/0907676X.2023.2247423

Ostrow, M., & Fiete, I. (2024). How the human brain creates cognitive maps of related concepts. *Nature*, *632*, 744–745. https://doi.org/10.1038/d41586-024-02433-2

Ouyang, Q. (2018). Assessing meaning-dimension quality in consecutive interpreting training. *Perspectives*, *26*(2), 196–213. https://doi.org/10.1080/0907676X.2017.1369552

Palmer, S. E., & Kimchi, R. (1986). *The information processing approach to cognition*. Lawrence Erlbaum.

Pérez Pérez, P. S. (2018). The use of a corpus management tool for the preparation of interpreting assignments: A case study. *The International Journal of Translation and Interpreting Research*, *10*(1), 137–151. https://doi.org/10.12807/ti.110201.2018.a08

Phung, D. V., & Michell, M. (2022). Inside teacher assessment decision-making: From judgement gestalts to assessment pathways. *Frontiers in Education*, *7*, Article 830311. https://doi.org/10.3389/feduc.2022.830311

Prandi, B. (2017). Designing a multimethod study on the use of CAI tools during simultaneous interpreting. *Proceedings of the 39th Conference Translating and the Computer*, 76–88. www.asling.org.

Prandi, B. (2018). An exploratory study on CAI tools in simultaneous interpreting: Theoretical framework and stimulus validation. In C. Fantinuoli (Ed.), *Interpreting and technology* (pp. 28–59). Language Science Press.

Prandi, B. (2020). The use of CAI tools in interpreter training: Where are we now and where do we go from here? *inTRAlinea*, *Special Issue: Technology in Interpreter Education and Practice.* http://www.intralinea.org/specials/article/2512

Prandi, B. (2023). *Computer-assisted simultaneous interpreting: A cognitive-experimental study on terminology*. Language Science Press.

Risku, H. (2020). *Cognitive approaches to translation*. John Wiley & Sons. https://doi.org/10.1002/9781405198431.wbeal0145.pub2

Risku, H., & Rogl, R. (2020). Translation and situated, embodied, distributed, embedded and extended cognition. In F. Alves & A. L. Jakobsen (Eds.), *The Routledge handbook of translation and cognition* (1st ed., pp. 478–499). Routledge. https://doi.org/10.4324/9781315178127-32

Rodd, J. M. (2024). Moving experimental psychology online: How to obtain high quality data when we can't see our participants. *Journal of Memory and Language*, *134*, Article 104472. https://doi.org/10.1016/j.jml.2023.104472

Sadler, D. R. (2009). Transforming holistic assessment and grading into a vehicle for complex learning. In G. Joughin (Ed.), *Assessment, learning and judgement in higher education* (pp. 1–19). Springer. https://doi.org/10.1007/978-1-4020-8905-3_4

Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior*, *160*, Article 108386. https://doi.org/10.1016/j.chb.2024.108386

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*(4), 1096–1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, *44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Su, W. (2020). *Eye-tracking processes and styles in sight translation*. Springer Singapore. https://doi.org/10.1007/978-981-15-5675-3

Tammasrisawat, P., & Rangponsumrit, N. (2023). The use of ASR-CAI tools and their impact on interpreters' performance during simultaneous interpretation. *New Voices in Translation Studies*, *28*(2), 25–51. https://doi.org/10.14456/nvts.2023.19

Timarová, S., Dragsted, B., & Gorm Hansen, I. (2011). Time lag in translation and interpreting: A methodological exploration. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Methods and strategies of process research* (Vol. 94, pp. 121–146). John Benjamins. https://doi.org/10.1075/btl.94.10tim

Tiselius, E. (2018). Exploring cognitive aspects of competence in sign language interpreting of dialogues: First impressions. *HERMES*, *57*, 49–61. https://doi.org/10.7146/hjlcb.v0i57.106193

Tzou, Y.-Z., Eslami, Z. R., Chen, H.-C., & Vaid, J. (2012). Effect of language proficiency and degree of formal training in simultaneous interpreting on working memory and interpreting performance: Evidence from Mandarin–English speakers. *International Journal of Bilingualism*, *16*(2), 213–227. https://doi.org/10.1177/1367006911403197

Vasey, M. W., Dalgleish, T., & Silverman, W. K. (2010). Research on information-processing factors in child and adolescent psychopathology: A critical commentary. *Journal of Clinical Child and Adolescent Psychology, 32*(1), 81–93. https://doi.org/10.1207/S15374424JCCP3201_08

Wan, H., & Yuan, X. (2022). Perceptions of computer-assisted interpreting tools in interpreter education in Chinese mainland: Preliminary findings of a survey. *International Journal of Chinese and English Translation & Interpreting, 1,* 1–28. https://doi.org/10.56395/ijceti.v1i1.8

Woesler, M. (2021). Modern interpreting with digital and technical aids: Challenges for interpreting in the twenty-first century. In R. Moratto & M. Woesler (Eds.), *Diverse voices in Chinese translation and interpreting* (pp. 191–217). Springer. https://doi.org/10.1007/978-981-33-4283-5_8

Xiao, Y., Hvelplund, K. T., & Ho, C.-E. (2023). Wearable eye trackers: Methodological challenges, opportunities and perspectives for sight interpreting/translation. *Translation, Cognition & Behavior*, *6*(2), 164–186. https://doi.org/10.1075/tcb.00084.xia

Yang, Z. (2021). Effective computer-assisted terminology management through SDL MultiTerm. *Journal of Physics: Conference Series*, *1861*, Article 012106. https://doi.org/10.1088/1742-6596/1861/1/012106

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*, Article 100131. https://doi.org/10.1016/j.amper.2023.100131

Zhang, J. (2021). *An experiment report on the impact of computer-aided interpreting tools on simultaneous interpreting* [Master's thesis]. China Foreign Affairs University. https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202201&filename=1021596437.nh&v=

Zhou, H., Weng, Y., & Zheng, B. (2021). Temporal eye-voice span as a dynamic indicator for cognitive effort during speech processing: A comparative study of reading aloud and sight translation. In R. Muñoz Martín, S. Sun, & D. Li (Eds.), *Advances in cognitive translation studies* (pp. 161–179). Springer. https://doi.org/10.1007/978-981-16-2070-6_8

Zhou, L. (2019). *The impact of computer-aided interpreting tools on simultaneous interpreting performance: Taking InterpretBank as an example* [Master's thesis]. Xiamen University. https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202002&filename=1019069326.nh&v=

Appendix 1: Summary of indicators and constructs

GLOSSARY COMPILATION	
BEHAVIOUR	
term- spotting	searching for terms in source texts
search queries	pasting terms into search boxes
search result review	reviewing search results
translation search	using resources to find translations
glossary review	reviewing glossary contents without modifying entries
entry editing	modifying glossary entries
RESULTS	
time total	total time spent on glossary tasks
term count	number of terms retained in the glossary
time per term	total time divided by the number of terms retained
diversity rate	percentage overlap between glossaries of different groups
SIMULTANEOUS INTERPRETING	
BEHAVIOUOR	
fillers	sounds like *uh* and *mmm* ... used to fill pauses in speech
repetitions	immediate repetition of a sequence of words
self-corrections	adjustments made to rectify errors during speech
false starts	abruptly interrupted beginnings of phrases
RESULTS	
correct	terms matching entries in the master glossary
adequate	terms rendering meaning but not in the master glossary
wrong	unacceptable renditions
skipped terms	terms not rendered
TIME SPANS	
ear–voice (evs)	from start/end of a source speech unit to start/end of its corresponding target output
ear–key (e2k)	from the end of a source speech utterance to the first related keyboard action
eye–voice (i2v)	from when a term’s translation appears on screen to when the participant vocalizes it
silent	• lags (< 200 ms) • bumps (200–600 ms) • respites (> 600 ms)

[1] https://www.hubermanlab.com/podcast

[2] https://www.laurenceanthony.net/software/antconc/

[3] https://bootcat.dipintra.it/

[4] MS Stream: https://www.microsoft.com/en/microsoft-365
Pynpt Keylogger: https://pypi.org/project/pynput/
TechSmith Capture: https://www.techsmith.com/