Enhancing consecutive interpreting with ASR: Sight-Terp as a computer-assisted interpreting tool[1]

Cihan Ünlü
Aymil Doğan

Cihan Ünlü
PhD Student at Boğaziçi University,
Research Assistant at İstanbul Yeni Yüzyıl University;
unlucihan@yahoo.com;
ORCID:0000-0001-7857-2031

Aymil Doğan
Independent Scholar;

Abstract

This experimental study investigates the potential impact of employing automatic speech recognition (ASR) and speech translation (ST) in consecutive interpreting (CI) through the use of a computer-assisted interpreting (CAI) tool. The tool used is Sight-Terp, an ASR-supported CAI tool developed and designed by the first author of this study. It offers multiple features, such as ASR, real-time ST, named entity highlighting, and automatically enumerated segmentation. The methodology adopted in this research involves a within-subjects design, assessing participants' output in scenarios with and without the use of Sight-Terp on a tablet. Twelve participants were recruited for the experimental setup and asked to interpret four English speeches into Turkish in long CI mode, using Sight-Terp for two of them and a pen and paper for the other two. The data analysis is grounded on parameters of both accuracy and fluency. In distinguishing the variance in accuracy across the two settings, accuracy metrics were rooted in the mean count of correctly rendered semantic units (units of meaning), as defined by Seleskovitch (1989). On the other hand, fluency was quantified by tracking the frequency of disfluency markers, including false starts, frequency of filled pauses, filler words, whole-word repetitions, broken words, and incomplete phrases in each session. The results show that the integration of ASR into two CI tasks led to an enhancement in the accuracy of the participants' rendition. Concurrently, however, it led to an increase in disfluencies and extended task durations compared to the tasks in which Sight-Terp was not used. The study outcomes also suggest potential areas of improvement and modifications that could further enhance the utility of the tool. Future empirical studies using Sight-Terp will tell us more about the feasibility of ASR in the interpreting process and cognitive aspects of human-machine interaction in CI.

Keywords: computer-assisted interpreting, automatic speech recognition, interpreting technology, consecutive interpreting, note-taking, tablet interpreting.

Rebuda: 31 de octubre de 2023 / Acceptació: 4 de novembre de 2024 / Publicació avançada: 20 de desembre de 2024

[1] This article is derived from the MA thesis titled "Automatic Speech Recognition in Consecutive Interpreter Workstation: Computer-Aided Interpreting Tool 'Sight-Terp'", completed at Hacettepe University, Ankara, in 2023.

Citació recomanada: Ünlü, Cihan; Doğan, Aymil (2024). «Enhancing consecutive interpreting with ASR: Sight-Terp as a computer-assisted interpreting tool». Revista Tradumàtica. Tecnologies de la Traducció, 22, 401-425. <https://doi.org/10.5565/rev/tradumatica.382>

http://revistes.uab.cat/tradumatica
ISSN 1578-7559

Resumen

Este estudio experimental investiga el impacto potencial de emplear el reconocimiento automático de voz (ASR) y la traducción de voz (ST) en la interpretación consecutiva (CI) utilizando una herramienta de interpretación asistida por ordenador (CAI). La herramienta CAI empleada en este contexto es “Sight-Terp”, una herramienta con soporte ASR desarrollada y diseñada por el primer autor de este estudio. Sight-Terp ofrece múltiples funciones, como ASR, traducción automática en tiempo real, resaltado de entidades nombradas y segmentación enumerada automáticamente. La metodología de la investigación adopta un diseño intra-sujetos, evaluando el rendimiento de los participantes en escenarios con y sin el uso de Sight-Terp en una tablet. Se reclutaron 12 participantes para el experimento, y se les pidió que interpretaran cuatro discursos en inglés en modo interpretación consecutiva larga al turco: dos utilizando Sight-Terp y los otros dos con papel y bolígrafo. El análisis de datos se basa en parámetros de precisión y fluidez. Para distinguir la variación en la precisión entre los dos escenarios, las métricas de precisión se fundamentaron en el promedio de unidades semánticas correctamente interpretadas (unidades de significado) según Seleskovitch (1989). Por otro lado, la fluidez se cuantificó rastreando la frecuencia de marcadores de disfluencia, incluidos falsos inicios, pausas innecesarias, palabras de relleno, repeticiones de palabras completas, palabras fragmentadas y frases incompletas en cada sesión. Los resultados muestran que la integración de ASR en dos tareas de interpretación consecutiva mejoró la precisión en las interpretaciones de los participantes. Sin embargo, esto también incrementó la frecuencia de disfluencias y prolongó la duración de sus rendimientos en comparación con las tareas realizadas sin Sight-Terp. Los hallazgos del estudio también sugieren áreas potenciales de mejora y modificaciones que podrían optimizar aún más la utilidad de la herramienta. Estudios empíricos futuros con Sight-Terp podrán ofrecer más información sobre la viabilidad del ASR en el proceso de interpretación y sobre los aspectos cognitivos de la interacción humano-máquina en la interpretación consecutiva.

Palabras clave: interpretación asistida por ordenador, reconocimiento automático de voz, tecnología de interpretación, interpretación consecutiva, toma de notas, interpretación con tablet.

Resum

Aquest estudi experimental investiga l’impacte potencial d’utilitzar el reconeixement automàtic de veu (ASR) i la traducció de veu (ST) en la interpretació consecutiva (CI) utilitzant una eina d’interpretació assistida per ordinador (CAI). L’eina CAI utilitzada en aquest context és “Sight-Terp”, una eina amb suport ASR desenvolupada i dissenyada pel primer autor d’aquest estudi. Sight-Terp ofereix múltiples funcions, com ara ASR, traducció automàtica en temps real, ressaltat d’entitats i segmentació enumerada automàticament. La metodologia de la investigació adopta un disseny intra-subjectes, avaluant el rendiment dels participants en escenaris amb i sense l’ús de Sight-Terp en una tablet. S’han reclutat 12 participants per a l’experiment, i se’ls ha demanat que interpretin quatre discursos en anglès en mode interpretació consecutiva llarga al turc: dos utilitzant Sight-Terp i els altres dos amb paper i bolígraf. L’anàlisi de dades es basa en paràmetres de precisió i fluïdesa. A fi de distingir la variació en la precisió entre tots dos escenaris, les mètriques de precisió s’han fonamentat en la mitjana d’unitats semàntiques correctament interpretades (unitats de significat) segons Seleskovitch (1989). D’altra banda, la fluïdesa s’ha quantificat rastrejant la freqüència de marcadors de disfluència, inclosos falsos inicis, pauses innecessàries, paraules per omplir, repeticions de paraules completes, paraules fragmentades i frases incompletes en cada sessió. Els resultats mostren que la integració d’ASR en dues tasques d’interpretació consecutiva ha millorat la precisió en les interpretacions dels participants. De tota manera, això també ha incrementat la freqüència de disfluències i ha prolongat la durada dels seus rendiments en comparació a les tasques realitzades sense Sight-Terp. Les troballes de l’estudi també suggereixen àrees potencials de millora i modificacions que podrien optimitzar encara més la utilitat de l’eina. Estudis empírics futurs amb Sight-Terp podran oferir més informació sobre la viabilitat de l’ASR en el procés d’interpretació i sobre els aspectes cognitius de la interacció humà-màquina en la interpretació consecutiva.

Paraules clau: interpretació assistida per ordinador, reconeixement automàtic de veu, tecnologia d'interpretació, interpretació consecutiva, presa de notes, interpretació amb tauleta.

1. Introduction

The advancement of technology has introduced numerous tools and solutions designed to enhance the accuracy and efficiency of interpreters. Integrating computer-assisted interpreting (CAI) tools and natural language processing (NLP) applications has paved the way for new linguistic and technical possibilities for interpreters and the interpreting process. Moreover, in recent years, the advent of tailored technological solutions for interpreters has made interpreting technology a topic of particular interest in academia, with a surge in the number of empirical studies.

CAI tools emerged as a set of software specifically designed to assist interpreters in various subprocesses of interpreting (Fantinuoli, 2018: 12), encompassing tasks ranging from easing cognitive load (Van Cauwenberghe, 2020; Defrancq and Fantinuoli, 2021) to conference preparation and terminology organisation (Fantinuoli, 2017b). Technological trends in interpreting have evolved with developments in NLP, speech technologies, general artificial intelligence, and the changing role of interpreters with the rise of remote simultaneous interpreting (RSI) and the rapid *platformisation* of the market. This so-called technologisation process, or "technological turn" (Fantinuoli, 2018), has reshaped interpreters' and end users' perception of CAI and interpreting technology in general, and is set to raise many questions about professionalism, automation, and ethics in the age of artificial intelligence.

Automatic speech recognition (ASR) is considered to be one of the game-changing innovations for the new generation of CAI tools, owing to its potential assistance during the task of interpreting. The complexity of interpreting is influenced by several factors involved in the task (Korpal and Stachowiak-Szymczak, 2019). These include the presence of numbers, lists, and proper names in the source text, as well as the speaker's rapid pace of delivery. Gile (2009: 171) refers to such elements as "problem triggers", which are factors that raise the processing capacity required during interpreting. As such, numerous studies have investigated the potential of ASR technology as an automated querying system (Hansen-Schirra, 2012; Fantinuoli, 2017a) and its application in enhancing CAI tools to address problem triggers during the process of interpreting (Van Cauwenberghe, 2020; Defrancq and Fantinuoli, 2021; Rodríguez et al., 2021; Pisani and Fantinuoli, 2021; Rodríguez González et al., 2023; Prandi, 2023). However, the emphasis has predominantly been on simultaneous rather than consecutive interpreting, indicating a need for more comprehensive empirical studies that examine human-computer interaction within CAI across other modes. This study diverges from previous studies primarily examining ASR use in simultaneous interpreting. Instead, it uses an ASR-supported CAI tool in consecutive interpreting (CI). The study aims to fill a gap in the literature by proposing digital software and provides insights into the effectiveness of ASR in CI with a mixed-methods experiment. The CAI tool developed and designed within the scope of this study is called Sight-Terp, a web-based digital application that initiates continuous speech translation (ST), named entity recognition (NER), and automatic speech segmentation.

In terms of methodology, this study employs a within-subjects design to assess 12 participants' output in scenarios both with and without the use of Sight-Terp.

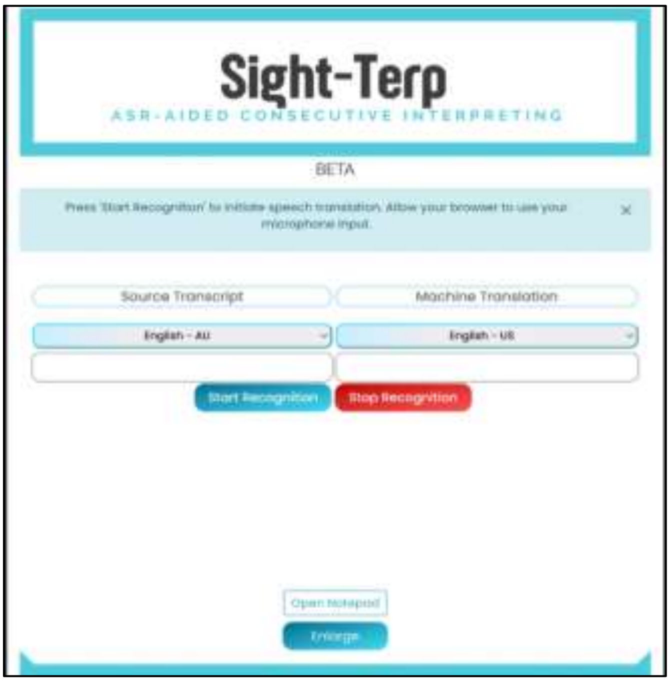
This study aims to answer the following research questions:

1. Does the use in CI of the CAI tool Sight-Terp, which provides both a source transcription and machine translation output, lead to a significant improvement in the interpreting accuracy of interpreters compared to their performance without technological aid?
2. Are there significant differences in the number of disfluencies (pauses, hesitations, repetitions, stuttering, false starts) between pre-test performances without CAI support and post-test performances with Sight-Terp support?
3. What are interpreters' perceptions of the interface and ergonomic design in terms of usability during the interpreting process?

Section 2 describes the functions and architecture of Sight-Terp. The third section explains the methodology of this study in detail. Then, section 4 outlines and discusses the findings. The fifth and last section will serve as a conclusion and make future research recommendations.

2. Sight-Terp: design and functionalities

Sight-Terp is a web-based, ASR-supported CAI tool designed and developed for processes that involve long CI. The tool mainly initiates continuous speech recognition (real-time ST) through a third-party application programming interface (API). Once the speech input is processed, the end-to-end ST model generates translated output in segmented units. Consequently, the system produces two automatically derived textual representations: a transcription of the original speech and its corresponding machine translation (MT). The two outputs are displayed as chunks in an enumerated style, with spaces between each to improve readability and comprehension for the user (interpreter). Concurrent with the transcription process, a NER model highlights identified entities within the text once the text chunk is completed on the source transcription side.



Press 'Start Recognition' to initiate speech translation. Allow your browser to use your microphone input.

Figure 1. The main layout of Sight-Terp (tablet view)

As shown in Figure 1, the user interface, which is mainly designed for tablet use, has buttons bound to certain additional functions. The interface is optimised for tablets due to their lightweight form, portability, touch-screen functionality, and widespread use in interpreting settings. Upon initiating an ST session with the "Start Recognition" button, two juxtaposed text boxes display the output of ASR and the corresponding neural machine translation in real time. Sight-Terp also integrates an optional, third-party digital note-taking application powered by MyScript Nebo. The "Open Notepad" button launches a simple artificial notepad. When used in conjunction with a stylus-equipped tablet, this embedded note-taking area emulates the experience of traditional pen-and-paper note-taking, enabling users to carry out additional note-taking, underlining, and circling. If preferred, this feature can be incorporated into the workflow for digital note-taking for contextual cues as the ST session continues in the background. In short, Sight-Terp incorporates four main functions for its workflow: ST, automatic segmentation, NER, and digital notepad. This section will proceed to delineate the components of Sight-Terp and their purpose in the process. The subsections below briefly explain the background and feasibility of each function component.

2.1. Automatic speech recognition (ASR) and speech translation (ST)

ST, also called speech-to-text translation, is an automated process of converting spoken input, a speech signal, from a source language into a target language in a textual representation. Its applications span a wide array of domains, such as lecture translation (Müller et al., 2016), virtual reality applications (Stefanel Gris et al., 2024), and subtitling (Saboo and Baumann, 2019).

Traditionally, ST systems relied on a cascaded architecture, where ASR and MT systems operated in a sequential way, and the output of the ASR system served as the direct input for the MT system (Stentiford and Steer, 1988). However, this approach has inherent limitations, notably error propagation from the ASR stage to the MT stage, which can degrade translation quality (Sperber and Paulik, 2020). If we look at more recent approaches to ASR and ST, end-to-end (E2E) deep learning models can achieve high-accuracy results by training a single E2E model that can perform both speech recognition and MT tasks simultaneously (without a separate ASR stage) (Bérard et al., 2016; Chiu et al., 2020; Baevski et al., 2020). These direct ST models are based on extensive datasets of paired speech and translation examples. Sight-Terp utilises Microsoft Azure Speech Translation (Xiong et al., 2018), which is one of the state-of-the-art models commercially available for API use at the time of writing. This API employs an E2E deep neural network architecture for robust speech processing. Notably, such E2E trainable encoder-decoder models have demonstrated superior performance compared to the aforementioned conventional cascaded approaches that utilise separate, non-unified ASR and MT systems (Bérard et al., 2016; Weiss et al., 2017). The ST model, implemented in JavaScript, facilitates accessibility through web browsers and uses the system's default microphone as the audio input source. When a user's vocal input is received, the browser transmits a request via the WebSocket protocol to geographically distributed Microsoft servers (situated in Europe). Once a secure, persistent connection is established, the ST session commences. As the process continues, the model tries to recognise individual utterances with minimal latency, where an utterance might comprise three to four sentences or a single phrase. Full pauses or brief silences serve as delimiters, signalling the end of an utterance to the model; this is called a sentence boundary. The model dynamically predicts sentence boundaries while processing the auditory input stream. Furthermore, the model is able to apply text normalisation and automatic punctuation for the text. Automatic punctuation and capitalisation prediction are also crucial for NLP tasks that rely on acoustic cues, such as pauses and pitch variations, to mitigate ASR and segmentation errors (Nozaki et al., 2022: 1). Thus, for a tool to provide a look-up mechanism in an interpreting task, accuracy in punctuation (commas, question marks, periods), based on the intonation and pacing patterns present in the input audio data, is of high importance. Needless to say, the other important factor is the accuracy and precision of the ST model. Admittedly, state-of-the-art ASR systems and ST models are not without flaws and face several issues. The speech type (casual or formal), speaker variability, and homonyms can challenge accurate transcription. Such factors may have an influence on word boundary prediction and can lead to transcription errors. In our case, accurate transcription is crucial for Sight-Terp to minimise segmentation errors and comprehension flaws for the interpreter.

For Fantinuoli (2017: 5), there are several criteria that an ASR system must meet to function effectively with a CAI tool. We adopted those criteria as they offer a comprehensive framework for evaluating ASR systems in not only simultaneous interpreting scenarios but also technology-enabled CI contexts.

- Speaker-independence
- Capability to manage continuous speech
- Support for large-vocabulary recognition (bigger models)
- Ability to customise vocabulary for specialised term recognition
- High-performance accuracy, i.e. low word error rate (WER)
- High speed, i.e. a low real-time factor (speed of the ASR system)

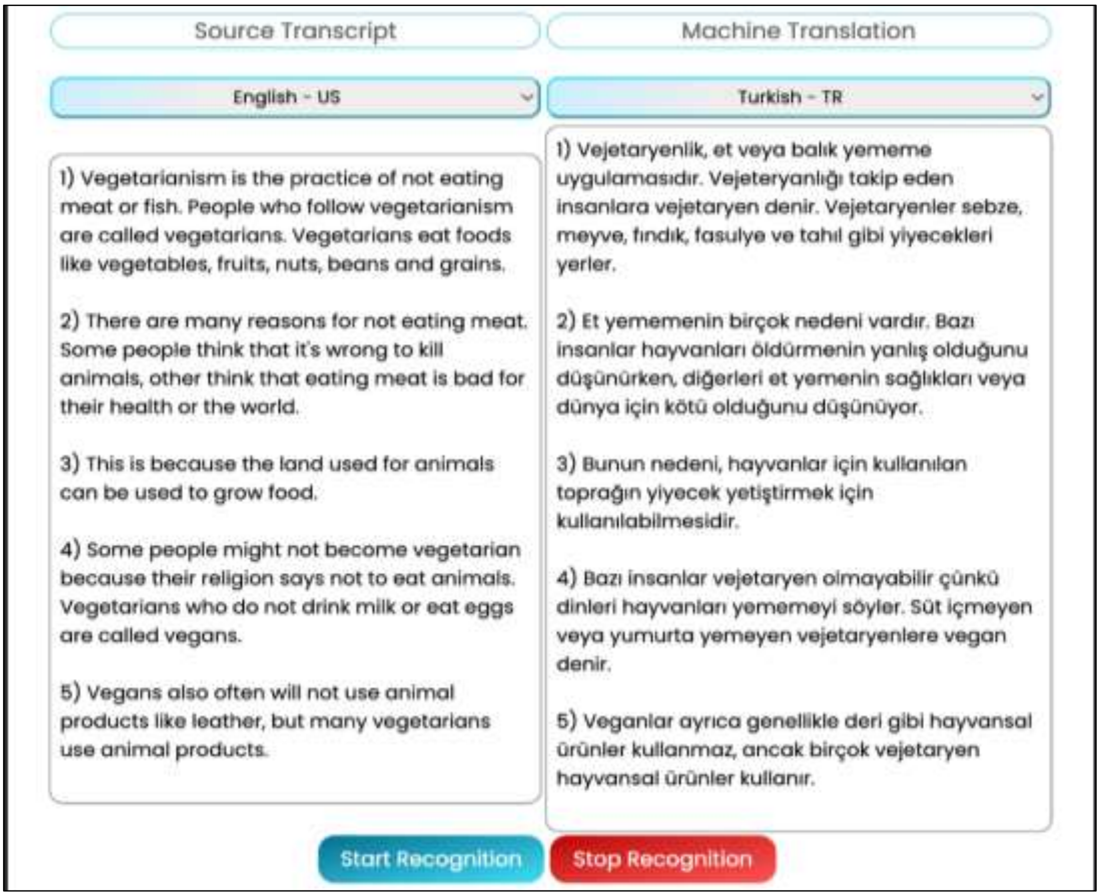
The ST session in Sight-Terp runs on the cloud server, and the source transcript and MT output that is generated through direct ST are displayed on the main interface in real time, thus creating two reference texts for the user (interpreter). As outlined in the above sections, this feature of Sight-Terp is the cornerstone of the system, which is intended to provide additional reference texts for the interpreter in the CI process. Thus, the tool aims to help the user render the source text in a “sight-consecutive” mode by improving the look-up mechanism and providing a memory prompt in both the source and target languages, especially in long CI. We applied NER and enumerated segmentation to the two outputs to improve such look-up and obtain aid from the reference source and target text. The subsections below will elaborate on these functionalities embedded in Sight-Terp.

2.2. Enumerated text segmentation and vertical display

Automatic text segmentation in Sight-Terp allows both the source and machine-translated texts to be displayed concurrently in a vertical format in adjacent text boxes during ST.

More precisely, the final text outputs are displayed in segments with ordinal numbers starting from one. The different segments are created during long pauses (approximately two seconds). The rationale behind this feature is to make sure the reference text is presented in an easily readable manner and to enable the interpreter to follow the source segment alongside its machine-translated output thanks to the aligned format.

This vertical segmentation approach is similar to conventional note-taking methods that Jean Herbert (1952) and Jean-François Rozan (1956) recommend in their seminal works about interpreting and note-taking because of the aid they provide in the CI process. Presenting information vertically (top to bottom) using “shifts” (Rozan, 1956) creates structured notes and aids the interpreter’s memory. Dörte Andres argues that “[t]he segmentation and the arrangement of the notes on the page can facilitate assignation (of the meaning) and have a positive effect on oral reproduction” (Andres, 2002, as cited by Gilles, 2017: 277). Sight-Terp’s auto-segmented outputs are displayed as manageable segments, keeping the user interface organised and aiding the interpreter in comprehending and skimming through the speaker’s arguments. Also, the segmentation on the target side in this format allows users to easily locate the corresponding MT text for specific units of interest. It is important to note that this vertical segmentation approach, while inspired by the traditional note-taking methods recommended by Herbert (1952) and Rozan (1956), differs from them in key ways. In fact, both Herbert and Rozan advocate for condensing information into key words and symbols, often arranged diagonally, which allows interpreters to quickly grasp the essence of the message. In contrast, ASR-generated text presents continuous speech in segmented form, which, while structured, still requires the interpreter to skim through full sentences or phrases, potentially increasing cognitive load by introducing two sources to process: the ASR output and its corresponding MT.



1) Vegetarianism is the practice of not eating meat or fish. People who follow vegetarianism are called vegetarians. Vegetarians eat foods like vegetables, fruits, nuts, beans and grains.

2) There are many reasons for not eating meat. Some people think that it's wrong to kill animals, other think that eating meat is bad for their health or the world.

3) This is because the land used for animals can be used to grow food.

4) Some people might not become vegetarian because their religion says not to eat animals. Vegetarians who do not drink milk or eat eggs are called vegans.

5) Vegans also often will not use animal products like leather, but many vegetarians use animal products.

1) Vejetaryenlik, et veya balık yememe uygulamasıdır. Vejeteryanlığı takip eden insanlara vejetaryen denir. Vejetaryenler sebze, meyve, fındık, fasulye ve tahıl gibi yiyecekleri yerler.

2) Et yememenin birçok nedeni vardır. Bazı insanlar hayvanları öldürmenin yanlış olduğunu düşünürken, diğerleri et yemenin sağlıkları veya dünya için kötü olduğunu düşünüyor.

3) Bunun nedeni, hayvanlar için kullanılan toprağın yiyecek yetiştirmek için kullanılabilmesidir.

4) Bazı insanlar vejetaryen olmayabilir çünkü dinleri hayvanları yememeyi söyler. Süt içmeyen veya yumurta yemeyen vejetaryenlere vegan denir.

5) Veganlar ayrıca genellikle deri gibi hayvansal ürünler kullanmaz, ancak birçok vejetaryen hayvansal ürünler kullanır.

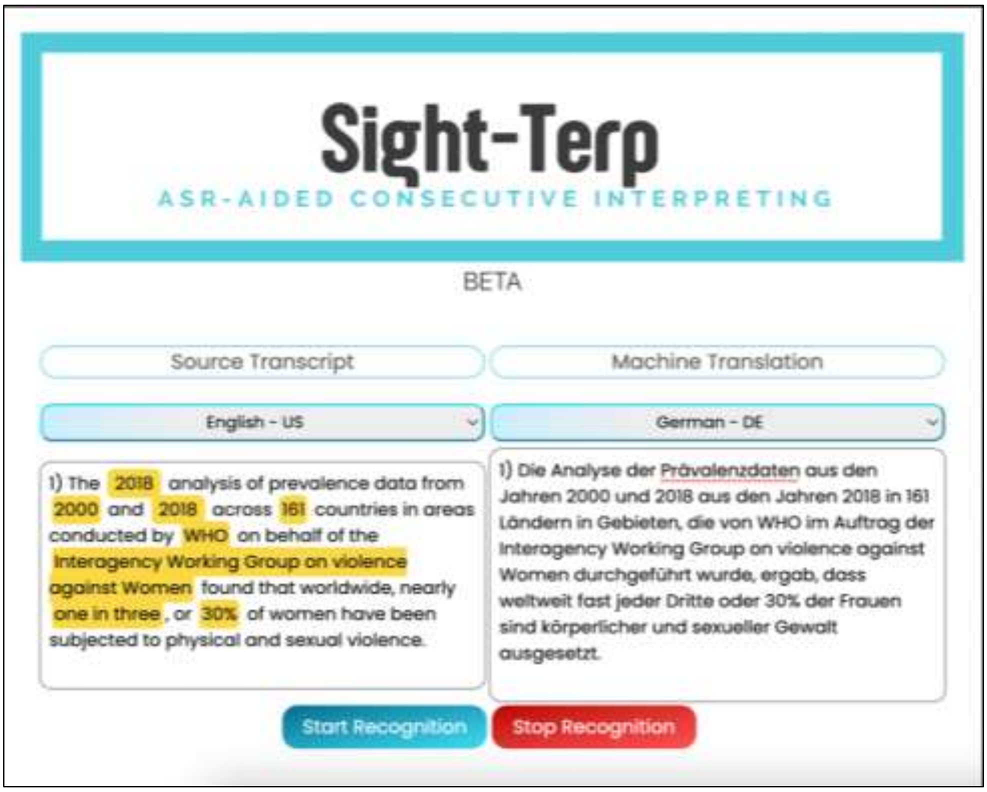
Figure 2. A segmented text on the Sight-Terp interface

Difficulties with locating necessary information and effective skimming were among the conclusions of the study conducted by Wang and Wang (2019), where they aimed to see whether providing an MT reference in CI could improve accuracy. In their experiment, participants were given an MT reference as a full paragraph generated by ASR and MT as they began interpreting in CI mode. However, in the post-experiment questionnaire, nine out of 10 participants reported difficulty in locating necessary information, leading to hesitations, pauses, and lower fluency scores (Wang and Wang, 2019: 135). This finding indicated some challenges of interpreting with long, unsegmented paragraphs, and the authors recommended displaying texts in sentence or utterance chunks to facilitate easier information retrieval and focus on semantic/lexical units in the output. Although it does not replicate the conciseness and flexibility of traditional note-taking, the segmentation in Sight-Terp can assist the interpreter by organising information into manageable units.

2.3. Named entity recognition (NER)

NER is a subfield of computer science and NLP which aims to identify and classify named items in unstructured text (Kim Sang et al., 2003; Cui et al., 2021). The NER process locates specific words or phrases that refer to real-world entities, such as people, organisations, locations, dates, and quantities. These entities are placed into predefined categories, this being the goal of NER. In fact, NER is a crucial component in various applications, including MT, question-answering systems and information retrieval (Keraghel et al., 2024: 1).

Sight-Terp utilises a NER model from the Microsoft Cognitive Services Text Analytics API, a neural NER model that delivers high confidence scores in generic texts. Such convolutional neural network-based models require large volumes of labelled training data. After each speech segment is displayed in the result boxes, the chunk (source) text is sent to a Node.js application running on a separate server. This server-side JavaScript application identifies the entities in the raw text and returns a heterogeneous array of results. The main application listens to the server through the WebSocket communication protocol. If entities from the predefined categories are found, they are directly highlighted in Sight-Terp's main interface while speech recognition continues for subsequent segments.



1) The 2018 analysis of prevalence data from 2000 and 2018 across 161 countries in areas conducted by WHO on behalf of the Interagency Working Group on violence against Women found that worldwide, nearly one in three , or 30% of women have been subjected to physical and sexual violence.

1) Die Analyse der Prävalenzdaten aus den Jahren 2000 und 2018 aus den Jahren 2018 in 161 Ländern in Gebieten, die von WHO im Auftrag der Interagency Working Group on violence against Women durchgeführt wurde, ergab, dass weltweit fast jeder Dritte oder 30% der Frauen sind körperlicher und sexueller Gewalt ausgesetzt.

Figure 3. Named entities highlighted on the Sight-Terp interface

The model highlights categories such as organisation names, personal names, dates, numerical data (e.g. percentages, ordinal numbers, temperatures), location names, and currency data (e.g. two million dollars). Accordingly, such categories recognised by the model contain critical technical and contextual information and are units of interest. The highlighting feature aims to simplify reformulation while reading from the reference aids in CI. This study does not aim to investigate the impact of automatic highlighting through NER but the impact of the combination of the main features of Sight-Terp. This feature is still in its infancy and its efficiency in interpreting is yet to be tested separately.

2.4. Digital notepad

Positioned below the main layout, the digital notepad is another experimental feature of Sight-Terp. This digital notepad uses the iinkJS JavaScript library provided by MyScript,[2] operating fully on the cloud in a client-server configuration. It offers functionalities such as handwriting recognition, digital ink capture, and rendering, supporting 65 languages. The digital notepad is particularly effective and suitable for use on tablets with a stylus like the Samsung S Pen or Apple Pen. Handwriting recognition refers to pen actions or strokes for editing or marking content, like crossing off or striking out text. For instance, scratching out text erases it, and drawing a frame around or underlining a word highlights it. As a matter of fact, these pen actions are commonly used by interpreters to emphasise meaning or indicate positive/negative content. Users can clear the page by clicking the three-dot icon and selecting the “Clear” button; this can be repeated for each turn taken in CI. It is worth noting that this study’s experimental procedure does not include tests on the digital notepad, as the primary research questions focus on the usability of ASR and NLP applications without involving basic note-taking with a pen. As mentioned before, the default interface displays the digital notepad at the very bottom of the page. However, for optimal usability, when used alongside the ASR function, it would be more effective to position the notepad on one side of the screen (either the right or left), with the ASR results displayed on the opposite side. This horizontal arrangement would allow interpreters to easily view both the ASR output and their notes simultaneously, without having to scroll or switch between sections. Future empirical studies with Sight-Terp could explore the interoperability of the digital notepad with ST.

3. Methodology

This study employed an experimental design to investigate the impact of an ASR-supported CAI tool on the accuracy and fluency of sight-consecutive interpreting performance among novice interpreters (n=12). The first test was a baseline for comparison with the second score. Pre-recorded speeches from English to Turkish were used for interpreting tasks and validated for difficulty through various readability indexes.

To assess the impact of Sight-Terp on accuracy, the study compared accuracy results between two conditions (with and without technological support) for the same participants. The within-subjects factor was the condition (with or without technological support), and the dependent variable was accuracy, measured as the percentage of accurately rendered “units of meaning”, which will be elaborated on in the data analysis section. Fluency, on

[2] https://myscript.github.io/iinkJS/docs/

the other hand, was measured by calculating the total number of occurrences of disfluency markers, such as false starts, frequency of filled pauses, filler words, whole-word repetitions, broken words, and incomplete phrases, for each performance.

Each participant interpreted a speech without technological aid in the initial test. They then received training in using Sight-Terp and subsequently interpreted another speech on the same topic using the tool. In total, the test was conducted twice in a row, with different but similar speeches being given each time. For the second round of the experiment, no training was provided. The rationale for using two consecutive pre-post-test designs with similar stimuli was to improve the reliability of the results by minimising random variability. Additionally, this approach allowed for a comparison between the results of the first and second rounds of tests, helping to assess the consistency of participants' performances across similar tasks. The following sections will provide a detailed explanation of the experimental procedure.

3.1. Materials

The stimuli used in the experiment consisted of four speeches delivered in English, by a native speaker, to be interpreted consecutively into Turkish, the participants' mother tongue. The speeches were categorised into two broad subjects, with two speeches on each topic. The first two speeches addressed violence against women,[3] while the third and fourth focused on earthquakes in Japan.[4]

Ensuring a consistent level of difficulty across the speeches was crucial for fair evaluation. Various readability indexes (Table 1) were applied to all the speeches to validate their difficulty levels. The Automated Readability Index (ARI), developed by Senter and Smith (1967), provides a formula for evaluating readability based on character count per word and words per sentence. The SMOG (Simple Measure of Gobbledygook) index, introduced by McLaughlin (1969), is designed to estimate the years of education required to understand a text. The Flesch–Kincaid Grade Level, derived by Kincaid et al. (1975), is widely used to assess the grade level of a text by considering sentence length and syllable count. The Coleman-Liau Index, proposed by Coleman and Liau (1975), computes readability using characters per word and sentence length, focusing on machine scoring. The Gunning-Fog Index, created by Gunning (1952), gauges the number of years of formal education needed to understand a text based on sentence length and complex words. The Flesch Reading Ease score, developed by Flesch (1948), evaluates readability by analysing word and sentence lengths, with a higher score indicating easier readability.

The results generally indicated closely comparable ratios, ensuring a consistent level of difficulty. This careful selection and consistent difficulty ensured internal validity, reducing the influence of confounding variables on the results. A limitation of this process is that while the speeches used were deemed comparable in terms of readability indexes,

[3] Reformulated from a speech with the same name publicly available at www.speechpool.net.

[4] Reformulated from a speech with the same name publicly available from the Speech Repository of the European Commission's Directorate-General for Interpretation.

other important linguistic factors were not controlled for. Variables such as the frequency and distribution of technical terms, the syntactic complexity of sentences, and the placement of key terms within the speech were not systematically taken into account. As a result, the speeches may not have been fully linguistically parallel.

Table 1 lists the readability index results and lexical density ratios of the speech materials.

Reading index	Subject: Earthquakes in Japan		Subject: Violence against women	
	Speech A1	Speech A2	Speech B1	Speech B2
Automated Readability Index	9.47	10.75	9.06	9.56
SMOG	10.91	11.13	11.15	11.71
Flesch–Kincaid Grade Level	8.88	9.24	8.5	9.66
Coleman-Liau Index	10.61	12.11	11.08	12.46
Gunning-Fog Index	11.12	11.40	11.24	12.14
Average Grade Level	10.2	10.93	10.21	11.67
Median Grade Level	10.61	11.12	11.08	12.06
Flesch Reading Ease	60.207	58.298	56.084	40.906
Lexical density	51.57%	56.09%	50.00%	54.93%

Table 1. Readability index results and lexical density ratios of speech materials

It was also crucial to make sure the speeches had similar durations and contained approximately equal amounts of named entities. This not only added a level of challenge to the task but also exposed the participants to a speech with entities highlighted in the scenario involving the use of Sight-Terp. Table 2 shows the duration, the number of units of meaning, and the length of the speeches in words. The number of units of meaning was defined and calculated by the first author of the study.

Material name	Duration	Length (words)	Number of units of meaning
Speech A1 Earthquakes in Japan	04:29	465	109
Speech A2 Earthquakes in Japan	04:35	452	127
Speech B1 Violence against women	04:01	513	159
Speech B2 Violence against women	03:39	404	125

Table 2. Detailed descriptions of speech materials (duration, length, units of meaning)

Before conducting the study, it was important to evaluate the accuracy of the ASR system by calculating the word-error rate (WER) for each speech (based on the source text). The WER measures the percentage of incorrectly recognised words in a speech, providing an objective measure of ASR quality and identifying potential issues. Table 3 shows the WER and the ASR system's NER precision for the speeches used in the second tests with Sight-Terp.

Material name	Word-error-rate (WER) by ASR	Named entity precision by ASR
Speech A1 Earthquakes in Japan	N/A	N/A
Speech A2 Earthquakes in Japan	9.7%	30/30
Speech B1 Violence against women	N/A	N/A
Speech B2 Violence against women	7.4%	30/32

Table 3. WER results and NER precision of ASR system

3.2. Questionnaire

The questionnaire included Likert-scale and open-ended questions to gather comprehensive feedback on the tool's effectiveness, usability, and reliability. It aimed to uncover any potential factors influencing or challenging participants' performance or learning processes. Expert opinions were incorporated into the questionnaire development process to ensure validity and relevance. The questions were as follows:

- How would you evaluate your experience with Sight-Terp?
- (Five-point Likert scale) I think that Sight-Terp is an easy-to-use tool.
- (Five-point Likert scale) Using automatic speech recognition (ASR) during consecutive interpreting tasks negatively impacted my performance.
- (Five-point Likert scale) I believe that the functions available in Sight-Terp contributed to my consecutive interpreting performance.
- Do you think that the automatic speech recognition (ASR) function in Sight-Terp is accurate and reliable?
- Which automatically generated output did you use for support during consecutive interpreting?
- Would you use Sight-Terp in your future professional life?
- Is there any feature/function that you would like to see in Sight-Terp?

3.3. Participants

Twelve participants were recruited for the experiment. All participants were third or fourth-year students in the English Translation and Interpreting (TIS) programme taught at Istanbul Yeni Yüzyıl University, who had achieved B grades or higher in the "Introduction to Consecutive Interpreting" and/or "Note-taking for Interpreting" courses. Participants who did not meet these criteria were excluded from the study. Of the 12 participants, five were female and seven were male. They were aged 20 to 24, with a mean age of 22 years. Table 4 shows the distribution of speeches in the two conditions by participant.

Participants	Earthquakes in Japan		Violence against women	
	Speech A1	Speech A2	Speech B1	Speech B2
Interpreter 1	No support	CAI support	No support	CAI support
Interpreter 2	No support	CAI support	No support	CAI support
Interpreter 3	No support	CAI support	No support	CAI support
Interpreter 4	No support	CAI support	No support	CAI support
Interpreter 5	No support	CAI support	No support	CAI support
Interpreter 6	No support	CAI support	No support	CAI support
Interpreter 7	No support	CAI support	No support	CAI support
Interpreter 8	No support	CAI support	No support	CAI support
Interpreter 9	No support	CAI support	No support	CAI support
Interpreter 10	No support	CAI support	No support	CAI support
Interpreter 11	No support	CAI support	No support	CAI support
Interpreter 12	No support	CAI support	No support	CAI support

Table 4. Distribution of speech materials by participant

3.4. Procedure

For the actual experiment, participants were invited, one by one, to the room where the experiment took place. The room was equipped with a table, chair, notepad, pen, 11-inch Apple iPad Pro (to run Sight-Terp), and a computer with speakers (to play the speeches). Upon arrival, participants were informed of the study's objectives and procedures, and their rights as voluntary participants. Initially, each participant interpreted speech A1 (earthquakes in Japan) into Turkish in consecutive mode using notes made with pen and paper. Participants were then provided with training in using Sight-Terp. Features such as ASR, real-time ST, NER, and automatic segmentation of speech were briefly introduced. Each participant was allowed time to use the tool and practise speaking into it with their own voice.

After 30 minutes of training, participants took the second test, which involved interpreting speech A2 into Turkish in consecutive mode using Sight-Terp. The experiment was repeated with another set of tests using the other two materials: speech B1 (violence against women 1) without technological aid and speech B2 (violence against women 2) with Sight-Terp. Following the tests, a qualitative survey was conducted to gather participants' perceptions and comparative feedback on tool usage, which formed the basis of the qualitative analysis. It is important to note that only participants' perceptions were included in the analysis; no external observations or field research by the experimenter were incorporated into the discussion. Finally, the fluency and accuracy results of each performance were manually analysed, as outlined in the data analysis section above.

In summary, the study followed a structured experimental procedure, starting with the first test, where participants performed CI without technological aid (using only pen and paper). After a brief break, participants received training in Sight-Terp. Following another break, the second test was conducted, where participants used Sight-Terp for the CI task. This sequence was repeated with a second set of tests involving similar speech materials, with no additional training for the second round. After completing the tests, the performances were transcribed and the number of units of meaning was calculated. Propositional analysis was used to measure the accuracy of the rendered units of meaning, followed by an assessment of disfluency markers for each performance. Lastly, a post-experiment questionnaire was administered to gather participants' feedback.

3.5. Pilot study

A pilot study was conducted at İstanbul Yeni Yüzyıl University during October and November 2022 to refine the design and implementation of the main study, to ensure it effectively addressed the research questions, and to validate the stimuli designed for data collection. A small sample of four participants was used for the preliminary study. Of the four participants, two were recent graduates of Translation and Interpreting Studies with a focus on interpreting, and the other two were senior TIS students, similar to the main study participants. All participants underwent the same procedure described and their performances were recorded for data analysis. The findings of the pilot study highlighted the potential benefits of incorporating Sight-Terp into CI tasks, as evidenced by the increase in accurately rendered units of meaning when participants used the tool. In an informal interview about the process, three out of the four participants reported that the ASR output did not fit the screen, requiring constant scrolling. As a result, an "Enlarge" button was added to the interface to expand the screen.

3.6. Data analysis techniques

During all tests, participants' voices (interpreting performances) were recorded. The gathered audio data for all four tests were analysed on the basis of two variables: accuracy and fluency. For performance analysis, accuracy was measured using a propositional analysis, and transcription of the performances was divided into the "units of meaning" proposed by Danica Seleskovitch (1989), with the total number of units calculated for comparison. These units represent the structural meaning of a sentence and can be broken down into smaller elements. This method, which involves propositional analysis techniques, primarily focuses on the semantic aspect of interpreting performance and is widely used by researchers to assess the quality of interpreting (Dillinger, 1994; Tommola and Heleva, 1998; Orlando, 2014).

After the experiment, the total number of accurately rendered units of meaning was calculated for each participant to assess their accuracy rate. This was done by dividing the number of accurately rendered units by the total number of units of meaning in the material and multiplying by 100 to obtain the percentage of correctly rendered units.

The aggregated accuracy rates of the participants were then compared between the first and second test to evaluate the tool's effectiveness in enhancing interpreting accuracy.

It should be noted that, in line with some interpreting strategies, interpreters may, consciously or unconsciously, subtract, add or substitute information in or introduce errors to semantic units. Clauses or phrases with additions and deletions that changed the meaning expressed in the source text were not counted as units of meaning. Additionally, rewordings and additions that did not negatively affect the text's context and the speaker's intention were not excluded from the designation of a meaning unit.

Some scholars view fluency as a measure of speech smoothness and continuity, while others perceive it as the interplay of temporal speech variables, such as pause length and uninterrupted speech runs, along with factors like "voice clarity, enunciation, and speaker confidence" (Freed, 2000: 261). In interpreting studies, there is a consensus that speech rate, pauses, hesitations, lengthened syllables, repetitions, self-corrections, and false starts are prosodic features affecting fluency. In this study, fluency was measured by analysing disfluency markers, which include the overall frequency of disfluencies, false starts, filled pauses, filler words, whole-word repetitions, broken words, and incomplete phrases (Lickley, 2015). The occurrences of these markers in the participants' renditions were counted and aggregated. The total number of disfluencies was calculated for each participant in both the first and second tests. The frequencies in the two conditions (with and without Sight-Terp) were then compared, and the results were outlined in a graph. Since not all the data are normally distributed, using a non-parametric test would be more appropriate. Additionally, the Wilcoxon signed-rank test was chosen as the statistical test due to our paired data design.

4. Findings and discussion

The experiment followed a repeated measures design, with participants interpreting two sets of speeches both with and without the Sight-Terp tool. Figure 4 shows the percentage of accurately rendered units of meaning. The Wilcoxon signed-rank test yielded a W value of 78.00 (p = 0.002), which indicates a statistically significant difference in accuracy when using the Sight-Terp tool. Conditional analysis showed higher interpreting accuracy with Sight-Terp (mean values: 87.05 and 90.10) compared to without (mean values: 55.41 and 54.22). Moreover, the effect size (r = 0.882) indicates that Sight-Terp had a substantial influence on accuracy (correctly rendered units of meaning). In summary, all 12 interpreters showed higher accuracy with Sight-Terp.

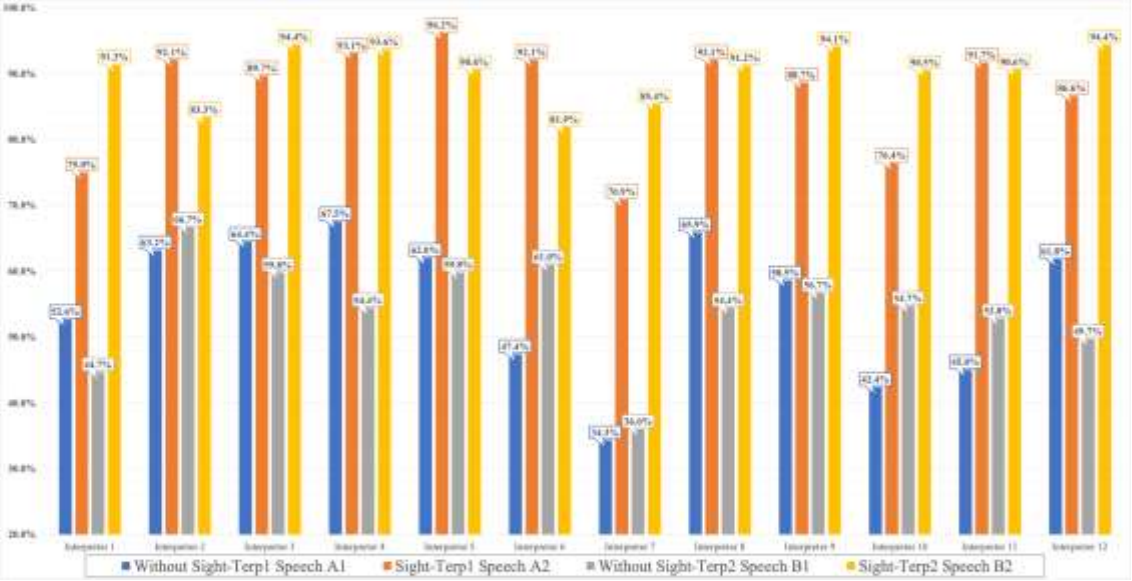


Figure 4. Comparable results of the main test: complete renditions of units of meaning in %

Looking at the fluency results, as shown in Figure 5, participants spent a longer time interpreting with the Sight-Terp tool compared to when using pen and paper. Higher accuracy with Sight-Terp was accompanied by longer duration and, as shown in Table 5, a higher frequency of disfluency markers.

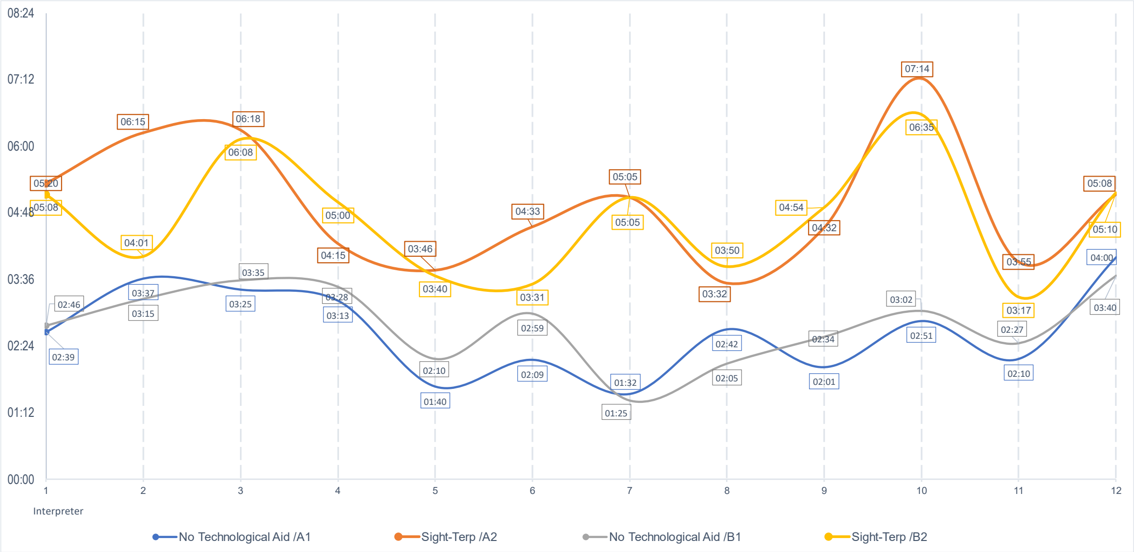


Figure 5. Comparable results of the main test: time taken to complete interpreting tasks (in minutes and seconds)

The higher disfluency rate in the Sight-Terp condition could be due to increased cognitive demand, as interpreters had to manage the key information units from the whole spoken input they had heard and perform sight translation into the target language. The dual references (MT+ASR) might have also interrupted their flow, leading to more disfluencies. The linguistic similarities in the transcriptions of their renditions further support this observation, suggesting that interpreters closely followed the ASR output when Sight-Terp was used. Moreover, the responses from the post-experiment questionnaire provide additional insights, with several participants indicating that the availability of the full source text made them more cautious and deliberate in their interpretation, potentially prioritising completeness over speed. However, it is also possible that the novelty of the technology or inaccuracies in the ASR output contributed to this effect.

Table 5 shows the frequency of disfluencies in numbers. A green upward arrow indicates an improvement in the interpreter's performance compared to the previous task. A red downward arrow indicates a decline in the interpreter's performance compared to the previous task. An orange rightward arrow indicates no significant change in the interpreter's performance compared to the previous task.

Participant	**No Tech. Aid**	**Sight-Terp**	**No Tech. Aid**	**Sight-Terp**
	Speech A1	**Speech A2**	**Speech B1**	**Speech B2**
Interpreter 1	⇩ 10	⇧ **20**	⇩ 11	⇘ **12**
Interpreter 2	⇩ 11	⇧ **29**	⇩ **14**	⇩ 13
Interpreter 3	⇩ 16	⇧ **25**	⇩ 17	⇘ **19**
Interpreter 4	⇩ 16	⇘ **19**	⇨ 22	⇧ **27**
Interpreter 5	⇩ 7	⇧ **8**	⇩ 7	⇧ **8**
Interpreter 6	⇗ 13	⇧ **15**	⇩ 10	⇨ **12**
Interpreter 7	⇩ 13	⇗ **22**	⇘ 16	⇧ **28**
Interpreter 8	⇨ 9	⇧ **10**	⇩ 8	⇨ **9**
Interpreter 9	⇩ 5	⇘ **9**	⇨ 11	⇧ **17**
Interpreter 10	⇩ 12	⇧ **29**	⇘ 19	⇧ **38**
Interpreter 11	⇗ 9	⇧ 10	⇘ 8	⇩ 7
Interpreter 12	⇩ 11	⇧ **18**	⇘ 13	⇨ **14**

Table 5. Instances of disfluency markers by participant

As outlined before, the questionnaire aimed to assess participants' experiences with and opinions about the Sight-Terp tool through a combination of Likert-scale, multiple-choice, and open-ended questions. Participants were generally positive about their experience with Sight-Terp, with most ratings leaning towards the positive end of the scale. Regarding ease of use, the majority of participants (11 out of 12) found Sight-Terp easy to use. When asked about the impact of ASR on their performance, most participants reported no negative effect, though three out of 12 were uncertain in that regard. Additionally, most participants (11 out of 12) felt that the features of Sight-Terp contributed positively to their performance. One participant reported struggling with the urge to use all the on-screen information, which made them stutter

Responses regarding the accuracy and reliability of the ASR function were generally positive, though participants were cautious and noted minor errors. Participants primarily used both the MT and source text outputs for support during CI. Preferences varied according to sentence type and specific needs while interpreting. Some participants felt more fluent using the MT output once they were reasonably sure of its accuracy, while others preferred the source transcription to avoid the discomfort of reading the automatic translation out loud.

When asked about potentially using Sight-Terp in their professional lives, most participants indicated a positive inclination. Some expressed a desire to see the tool used in various contexts before fully committing. A few participants mentioned that they would like to use Sight-Terp while still taking notes on paper, despite the tool's digital note-taking capabilities.

In addition, several suggestions regarding improvements emerged from the feedback. Participants noted the need for improved segmentation consistency: flawed segmentation sometimes led to incoherent MT output. Adding features for manually merging or splitting segments was also suggested.

Currently, the tool stops scrolling once the transcription reaches a certain point, requiring the user to manually scroll down, which participants found inconvenient and difficult to manage. Therefore, automatic scrolling was another feature recommended with a view to enhancing the user experience. Participants felt that a post-editing function to make real-time corrections of minor errors in MT output would be desirable. They also suggested enabling the ability to click on highlighted words to see their equivalents in the target language and highlighting named entities in both the source and target texts. Finally, thin lines between segments were suggested to help distinguish them from one another and avoid confusion while navigating the reference texts.

5. Conclusion

This study aimed to investigate the effectiveness and potential of the ASR-supported CAI tool Sight-Terp for enhancing CI performance on the basis of two variables: accuracy and fluency. It also aimed to introduce the functions and design of Sight-Terp as a publicly available computer-assisted tool developed for CI scenarios. Through quantitative and small-scale qualitative analysis, it has been observed that the use of Sight-Terp leads to a noteworthy improvement in content accuracy when interpreting. The findings of the analysis also showed that participants achieved increased precision in their renditions and were more attentive to and engaged more with the text when using Sight-Terp compared to when interpreting without technological aid (i.e. using only pen and paper). However, despite these improvements in accuracy, participants found it more challenging to deliver a fluent, unfragmented rendition.

Accordingly, data analysis reveals that disfluencies, including pauses, hesitations, repetitions, stuttering, and false starts, occurred with a noticeably higher frequency when participants used the Sight-Terp tool. This increase in disfluency markers suggests that the use of Sight-Terp may have influenced the flow of rendition, potentially due to the cognitive load associated with processing the additional information provided by the tool and/or unfamiliarity with ASR/ST. Participants consistently took longer to complete their rendition when using the Sight-Terp tool than when working with no aid. There are many possible reasons for this, including the additional time required to read and process the text(s) provided by Sight-Terp, participants more meticulously transferring everything into the target language, additional/redundant information provided by Sight-Terp, or unfamiliarity with the tool. However, it is important to clarify that the frequency of disfluency markers observed in this study should not be directly interpreted as an indicator of cognitive load. While disfluencies may correlate with increased cognitive demand, they do not provide a comprehensive measure of cognitive load itself. Further research utilising specific cognitive load measurement frameworks, such as dual-task methods or physiological indicators, would be necessary to draw more definitive conclusions regarding the cognitive strain associated with Sight-Terp usage.

The questionnaire revealed that users found the functions available in Sight-Terp beneficial for their interpreting performance, highlighting the tool's usefulness for supporting interpreters. However, the reliability and accuracy of ASR and MT results were viewed with some scepticism. Interestingly, the study found that users employed different strategies when utilising Sight-Terp's automatically generated outputs for support during CI. For instance, some participants relied on the MT output for more complex sentences, whereas for simpler or critical units they anticipated that such output might fall short and, thus, opted to use the speech transcription instead. Future experimental studies could unveil how interpreters interact with each type of output in a "sight-consecutive" mode.

Such conclusions could certainly lead to a substantial step forward in promoting and ensuring quality in the integration of ASR into interpreting practice. However, this study has several limitations worth mentioning. One limitation is that the study was conducted with students/novice interpreters. Additionally, the language pair used in this study was Turkish and English, with interpreting tasks from English into Turkish. The directionality may introduce other factors that interfere with the accuracy and completeness of renditions, as well as the accuracy of the ASR/ST model, especially in technology-mediated interpreting scenarios. Furthermore, variables such as specific domains, speech characteristics, and accents are highly relevant and may significantly affect the tool's performance and usability. Another limitation is reliance on the Microsoft Azure Speech Recognition API; although it is considered one of the best ASR/ST systems at the time of writing, this reliance should still be taken into account when evaluating the proposed software's overall performance and effectiveness.

There are various options for future research using ASR/ST in CI. First and foremost, the digital notepad feature of Sight-Terp has not been used within the scope of this study. Future studies might explore the interoperability of the digital notepad and ASR when using a tablet. From a cognitive perspective, triangulating eye-tracking data and transcription analysis could provide insight into how users interact with the tool and identify areas for improvement. Experimenting with different language pairs and directions could also indicate how the tool performs across various linguistic contexts.

Additionally, future research with Sight-Terp could examine the balance between efficiency and accuracy when interpreters consult multiple written sources — such as the segmented source text (ASR), the MT output, and their own notes — during CI. While quickly deciding to trust the ASR or MT output without analysing it carefully may lead to more errors, carefully reading and mentally verifying the correctness of the text may increase cognitive load. In this study, cognitive load is partially operationalised and reflected in the increased frequency rates in renditions. Further experimental research could provide deeper insights into how technological aids like Sight-Terp impact this trade-off and identify specific instances where participants deviate from the source transcription or MT output in favour of other reference aids.

6. References

Andres, Dörte (2002). *Konsekutivdolmetschen und Notation*. Frankfurt am Main: Peter Lang. <https://openscience.ub.uni-mainz.de/handle/20.500.12030/1317.>. [Accessed: 20241219].

Bérard, Alexandre; Pietquin, Olivier; Servan, Christophe; Besacier, Laurent (2016). Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint*, arXiv:1612.01744. <https://arxiv.org/abs/1612.01744.>. [Accessed: 20241219].

Chiu, Chung-Cheng; Sainath, Tara N.; Wu, Yonghui; Prabhavalkar, Rohit; Nguyen, Patrick; Chen, Zhifeng; Kannan, Anjuli; Weiss, Ron J.; Rao, Kanishka; Gonina, Ekaterina; Jaitly, Navdeep; Li, Bo; Chorowski, Jan; Bacchiani, Michiel (2018). State-of-the-art speech recognition with sequence-to-sequence models. In: Yvon, François; Hansen, Viggo (eds.). *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB, Canada: IEEE, pp. 4774-4778. <https://doi.org/10.1109/ICASSP.2018.8462105.>. [Accessed: 20241219].

Coleman, Meri; Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, v. 60, n. 2, pp. 283–284. <https://doi.org/10.1037/h0076540.>. [Accessed: 20241219].

Cui, Leyang; Wu, Yu; Liu, Jian; Yang, Sen; Zhang, Yue (2021). Template-based named entity recognition using BART. *arXiv preprint*, arXiv:2106.01760. <https://arxiv.org/abs/2106.01760.>. [Accessed: 20241219].

Defrancq, Bart; Fantinuoli, Claudio (2021). Automatic speech recognition in the booth: Assessment of system performance, interpreters' performances and interactions in the context of numbers. *Target. International Journal of Translation Studies*, v. 33, n. 1, pp. 73–102. <https://benjamins.com/catalog/target.19166.def.>. [Accessed: 20241219].

Dillinger, Mike (1994). Comprehension during interpreting: What do interpreters know that bilinguals don't? In: Lambert, Sylvie; Moser-Mercer, Barbara (eds.). *Bridging the Gap: Empirical Research in Simultaneous Interpretation*. Amsterdam: John Benjamins, pp. 155–190. <https://doi.org/10.1075/btl.3.14dil.>. [Accessed: 20241219].

Fantinuoli, Claudio (2017a). Speech recognition in the interpreter workstation. In: Proceedings of the 39th Conference Translating and the Computer. London, UK: Editions Tradulex, pp. 25–34. <https://www.staff.uni-mainz.de/fantinuo/download/publications/Speech%20Recognition%20in%20the%20Interpreter%20Workstation.pdf.>. [Accessed: 20241219].

Fantinuoli, Claudio (2017b). Computer-assisted preparation in conference interpreting. *Translation & Interpreting: The International Journal of Translation and Interpreting Research*, v. 9, n. 2, pp. 24–37. <https://doi.org/10.12807/ti.109202.2017.a02.>. [Accessed: 20241219].

Fantinuoli, Claudio (2018). Interpreting and technology: The upcoming technological turn. In: Fantinuoli, Claudio (ed.). *Interpreting and Technology*. Berlin: Language Science Press, pp. 1–12. <https://doi.org/10.5281/zenodo.1493289.>. [Accessed: 20241219].

Flesch, Rudolf (1948). A new readability yardstick. *Journal of Applied Psychology*, v. 32, n. 3, pp. 221–233. <https://doi.org/10.1037/h0057532.>. [Accessed: 20241219].

Gile, Daniel (2009). *Basic Concepts and Models for Interpreter and Translator Training*. Amsterdam: John Benjamins Publishing Company. <https://doi.org/10.1075/btl.8.>. [Accessed: 20241219].

Gillies, Andrew (2017). *Note-taking for Consecutive Interpreting: A Short Course*. 2nd ed. London: Routledge. <https://doi.org/10.4324/9781315648996.>. [Accessed: 20241219].

Gunning, Robert (1952). *The Technique of Clear Writing*. New York: McGraw-Hill.

Hansen-Schirra, Silvia (2012). Nutzbarkeit von Sprachtechnologien für die Translation. *trans-kom*, v. 5, n. 2, pp. 211–226. <https://www.trans-kom.eu/ihv_05_02_2012.html.>. [Accessed: 20241219].

Herbert, Jean (1952). *Manuel de l'interprète: Comment on devient interprète de conférences*. Genève: Librairie de l'Université Genève.

Keraghel, Imed; Morbieu, Stanislas; Nadif, Mohamed (2024). A survey on recent advances in named entity recognition. *arXiv preprint*, arXiv:2401.10825. <https://arxiv.org/abs/2401.10825.>. [Accessed: 20241219].

Kincaid, J. Peter; Fishburne, Robert P. Jr.; Rogers, Richard L.; Chissom, Brad S. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy Enlisted Personnel. *Chief of Naval Technical Training*. <https://doi.org/10.21236/ADA006655.>. [Accessed: 20241219].

Korpal, Paweł; Stachowiak-Szymczak, Katarzyna (2019). Combined problem triggers in simultaneous interpreting: Exploring the effect of delivery rate on processing and rendering numbers. *Perspectives*, v. 28, n. 1, pp. 126–143. <https://doi.org/10.1080/0907676X.2019.1628285.>. [Accessed: 20241219].

Lickley, Robin J. (2015). Fluency and Disfluency. In: Redford, Melissa A. (ed.). *The Handbook of Speech Production*. Hoboken, NJ: John Wiley & Sons, pp. 445–474. <https://doi.org/10.1002/9781118584156.ch20.>. [Accessed: 20241219].

Lucas Rafael Stefanel Gris, Diogo Fernandes, Frederico Santos de Oliveira, Anderson da Silva Soares, Telma Woerle de Lima Soares, and Arlindo Rodrigues Galvão (2024). "Automatic Speech-to-Speech Translation of Educational Videos Using SeamlessM4T and Its Use for Future VR Applications." In *Proceedings of the 2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, Orlando, FL, USA, pp. 163–166. <https://doi.org/10.1109/VRW62533.2024.00033.>. [Accessed: 20241219].

McLaughlin, G. Harry (1969). SMOG grading: A new readability formula. *Journal of Reading*, v. 12, n. 8, pp. 639–646. <http://www.jstor.org/stable/40011226.>. [Accessed: 20241219]..

Müller, Markus; Nguyen, Thai Son; Niehues, Jan; Cho, Eunah; Krüger, Bastian; Ha, Thanh-Le; Kilgour, Kevin; Sperber, Matthias; Mediani, Mohammed; Stüker, Sebastian; Waibel, Alex (2016). Lecture Translator – Speech translation framework for simultaneous lecture translation. In: DeNero, John; Finlayson, Mark; Reddy, Sravana (eds.). *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. San Diego, California: Association for Computational Linguistics, pp. 82–86. <https://doi.org/10.18653/v1/N16-3017.>. [Accessed: 20241219]. .

Nozaki, Jumon; Kawahara, Tatsuya; Ishizuka, Kenkichi; Hashimoto, Taiichi (2022). End-to-End Speech-to-Punctuated-Text Recognition. *arXiv preprint*. <https://arxiv.org/abs/2207.03169.>. [Accessed: 20241219]..

Orlando, Marc (2014). A study on the amenability of digital pen technology in a hybrid mode of interpreting: Consec-simul with notes. *Translation and Interpreting*, v. 6, n. 2, pp. 39–54. https://doi.org/10.12807/ti.106202.2014.a03.

Pisani, Elisabetta; Fantinuoli, Claudio (2021). Measuring the Impact of Automatic Speech Recognition on Number Rendition in Simultaneous Interpreting. In: Wang, Caiwen; Zheng, Binghan (eds.). *Empirical Studies of Translation and Interpreting: The Post-Structuralist Approach*. 1st ed. London: Routledge, pp. 181–197. <https://doi.org/10.4324/9781003017400-14.>. [Accessed: 20241219].

Prandi, Bianca (2023). *Computer-Assisted Simultaneous Interpreting: A Cognitive-Experimental Study on Terminology*. Berlin: Language Science Press. <https://langsci-press.org/catalog/book/348.>. [Accessed: 20241219].

Rodríguez González, Elena; Saeed, Muhammad Asad; Korybski, Tomasz; Davitti, Elena; Braun, Sabine (2023). Assessing the Impact of Automatic Speech Recognition on Remote Simultaneous Interpreting Performance Using the NTR Model. In: Corpas Pastor, Gloria; Hidalgo-Ternero, Carlos Manuel. (eds.). *Proceedings of the International Workshop on Interpreting Technologies - SAY IT AGAIN 2023*, Málaga, Spain, 2–3 November 2023, pp. 177–186.

Rodriguez, Susana; Gretter, Roberto; Matassoni, Marco; Alonso, Alvaro; Corcho, Oscar; Rico, Mariano; Falavigna, Daniele (2021). SmarTerp: A CAI System to Support Simultaneous Interpreters in Real-Time. In: Mitkov, Ruslan; Sosoni, Vilelmini; Giguère, Julie Christine; Murgolo, Elena; Deysel, Elizabeth (eds.). *Proceedings of the Translation and Interpreting Technology Online Conference (TRITON 2021)*. Held Online: INCOMA Ltd., pp. 102–109. <https://aclanthology.org/2021.triton-1.12.>. [Accessed: 20241219].

Rozan, Jean-François (1956). *La prise de notes en interprétation consécutive*. Genève: Librairie de l'Université.

Saboo, Ashutosh; Baumann, Timo (2019). Integration of Dubbing Constraints into Machine Translation. In: Bojar, Ondřej; Chatterjee, Rajen; Federmann, Christian; Fishel, Mark; Graham, Yvette; Haddow, Barry; Huck, Matthias; Jimeno Yepes, Antonio; Koehn, Philipp; Martins, André; Monz, Christof; Negri, Matteo; Névéol, Aurélie; Neves, Mariana; Post, Matt; Turchi, Marco; Verspoor, Karin (eds.). *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Florence, Italy: Association for Computational Linguistics, pp. 94–101. <https://aclanthology.org/W19-5210.>. [Accessed: 20241219].

Seleskovitch, Danica; Lederer, Marianne (1989). *Pédagogie raisonnée de l'interprétation*. Paris: Didier Érudition/OPOCE.

Smith, E. A.; Senter, R. J. (1967). *Automated Readability Index*. AMRL-TR-66-220. Wright-Patterson Air Force Base, Ohio: Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command.

Tjong Kim Sang, Erik F.; De Meulder, Fien (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Daelemans, Walter; Osborne, Miles (eds.). *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. Edmonton, Canada: Association for Computational Linguistics, pp. 142–147. <https://doi.org/10.3115/1119176.1119195.>. [Accessed: 20241219].

Tommola, Jorma; Helevä, Marketta (1998). Language Direction and Source Text Complexity Effects on Trainee Performance in Simultaneous Interpreting. In: Bowker, Lynne; Cronin, Michael; Kenny, Dorothy; Pearson, Jennifer (eds.). *Unity in Diversity: Current Trends in Translation Studies*. Manchester: St. Jerome Publishing, pp. 177–186.

Van Cauwenberghe, Goran (2020). *La reconnaissance automatique de la parole en interprétation simultanée: étude expérimentale de l'impact d'un soutien visuel automatisé sur la restitution de terminologie spécialisée*. [Master's thesis], Ghent University. Ghent. <https://lib.ugent.be/catalog/rug01:002862551>. [Accessed: 20241219].

Wang, Xinyu; Wang, Caiwen (2019). Can Computer-Assisted Interpreting Tools Assist Interpreting? *Transletters: International Journal of Translation and Interpreting*, 3, 109–139. <https://journals.uco.es/tl/article/view/11575>. [Accessed: 20241219].

Weiss, Ron J.; Chorowski, Jan; Jaitly, Navdeep; Wu, Yonghui; Chen, Zhifeng (2017). Sequence-to-Sequence Models Can Directly Translate Foreign Speech. *arXiv preprint*, arXiv:1703.08581. <https://arxiv.org/abs/1703.08581.

Xiong, Wayne; Wu, Lingfeng; Alleva, Frank; Droppo, Jeffrey; Huang, Xuedong; Stolcke, Andreas (2018). The Microsoft 2017 Conversational Speech Recognition System. In: *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB, Canada: IEEE, pp. 5934–5938. https://doi.org/10.1109/ICASSP.2018.8461870>. [Accessed: 20241219].

Zhang, Yu; Park, Daniel S.; Han, Wei; Qin, James; Gulati, Anmol; Shor, Joel; Jansen, Aren; Xu, Yuanzhong; Huang, Yanping; Wang, Shibo; Zhou, Zongwei; Li, Bo; Ma, Min; Chan, William; Yu, Jiahui; Wang, Yongqiang; Cao, Liangliang; Sim, Khe Chai; Ramabhadran, Bhuvana; Sainath, Tara N.; Beaufays, Françoise; Chen, Zhifeng; Le, Quoc V.; Chiu, Chung-Cheng; Pang, Ruoming; Wu, Yonghui (2022). BigSSL: Exploring the Frontier of Large-Scale Semi-Supervised Learning for Automatic Speech Recognition. *IEEE Journal of Selected Topics in Signal Processing*, 16(6), 1519–1532. <https://doi.org/10.1109/JSTSP.2022.3182537. .>. [Accessed: 20241219].