**The Interpreters' Newsletter Special Issue 2024 (29 bis), 59-78**
DOI: 10.13137/2421-714X/36765 ISSN 1591-4127 (print) ISSN 2421-714X (online)
https://www.openstarts.units.it/dspace/handle/10077/2119

# Automatic Speech Recognition in conference interpreting: an exploratory study on consecutive interpreting assisted by Sight-Terp

Michele Restuccia
University of Trieste

Abstract

*This article aims to observe potential interpreter-machine approaches in Consecutive Interpreting (CI) assisted by Automatic Speech Recognition (ASR). Sight-Terp was used in the study, a CAI tool for CI equipped with ASR and Machine Translation (MT). Six student-interpreters at the end of their second year of the master's degree in Conference Interpreting at SSLMIT, Trieste were asked to interpret one speech in the traditional consecutive manner and one using Sight-Terp in a randomised order. The results show that a) the transcript produced by Sight-Terp does not always enable the interpreters to concentrate more on speech listening and comprehension, b) the duration of the interpreters' renditions tends to increase, owing to an increased dependency on the written text and to additional cognitive efforts involved in the rendition phase, and c) ASR can be a useful resource for the retrieval of names and figures (although there may be errors in the transcript), but it presents limitations in the transcription of spontaneous spoken language. In conclusion, despite the significant advances in ASR technologies, more research is needed to further explore the advantages of ASR-assisted consecutive interpreting and strategic approaches.*

Keywords

Assisted consecutive interpreting, ASR, CAI tool, interpreting technologies, Sight-Terp.

## Introduction

Over the past century, since the emergence of modern interpreting, technology has consistently played a crucial role in conference interpreting (Kalina/Ziegler 2015). This is particularly prevalent in simultaneous interpreting, though possibly less so in consecutive interpreting. However, it has often taken time for any advance in the field to gain acceptance among interpreters (both practitioners and researchers). For example, there was even initial reluctance towards simultaneous interpreting (Gaiba 1998: 163; Fantinuoli 2016), now considered one of the most significant inventions of the 20th century (Stoll 2001). This pattern persists in the present era of the "technological turn" (Fantinuoli 2018), in which interpreters are confronted with Computer-Assisted Interpreting (CAI) tools, the relentless rise of Remote Simultaneous Interpreting (RSI), and, recently, Machine Interpreting (MI).

Unlike other pivotal moments in the brief history of conference interpreting, the current technological turn presents unprecedented characteristics. Firstly, new practices made possible by cutting-edge technologies have rarely been brought about by interpreters, but rather external circumstances, such as the Covid-19 pandemic which contributed to the extraordinary growth of RSI (Fantinuoli 2021: 509). Secondly, unlike previous technological solutions – particularly those developed for terminology management – the most sophisticated tools for interpreters (such as AI-based CAI tools and MI) have a direct impact on the interpreting task itself and the cognitive processes that underpin it. Such tools require interpreters to redistribute their cognitive resources and develop new strategies to obtain genuine quality gains.

There are several factors that explain the scepticism of professional interpreters, especially among older generations. Indeed, many interpreters still tend to prefer either traditional approaches with no interference from CAI tools, reliance on their own skills acquired through professional experience (Prandi 2020), or technological resources that can be used for preparation rather than during the task itself (Corpas Pastor/Fern 2016). The factors underlying this stance include the cost of CAI tools, their lack of functionality and adaptability to the specific needs of individual interpreters, and their influence on the interpreting process, which is already cognitively demanding (Tripepi Winteringham 2010, Corpas Pastor/Fern 2016). On the other hand, the use of technologies seems to be almost unavoidable outside the interpreting booth, namely during preparation, when interpreters usually make use of non-specific resources, such as the Internet, or office tools such as Word or Excel (Moser-Mercer 1992; Berber-Irabien 2010; Corpas Pastor/Fern 2016). The doubts and criticisms voiced by professional interpreters partly account for the delay in research on new interpreting technologies. However, an interest in these new resources and more in-depth research are essential for future developments, not only because the working reality is constantly evolving, but also because other stakeholders – such as RSI and software providers – may contribute to changing the interpreting market and its practices without considering the profession's standards and code of ethics. If interpreters do not take an interest in technologies that genuinely meet their needs, "it is possible that interpreters may soon have to use tools they have not personally selected" (Prandi 2023: 51).

## 1. Consecutive interpreting and Automatic Speech Recognition

Unlike simultaneous interpreting, which has been the focus of numerous studies on CAI tools, few contributions have been devoted to technology-assisted (or technology-enhanced) consecutive interpreting. One reason assisted consecutive interpreting has been studied less than assisted simultaneous interpreting could be its lower diffusion in the present working reality, as well as the fact that the practice itself is realised in two steps. Another reason is likely to be that in communicative settings the use of a CAI tool is logistically less practical than a CAI tool for simultaneous interpreting, which is used in an isolated booth. Furthermore, in formal and high-stakes settings where consecutive interpreting (CI) is generally needed – such as a meeting of two delegations, a meeting of a working group, inaugural ceremonies, or welcoming a foreign delegation (Riccardi 2003: 110) – interpreters seem to prefer the traditional pen and notepad approach, mainly for logistical (outdoor venues, battery life, inadvertently hit buttons on the tablet screen, screen brightness) and confidentiality reasons (Goldsmith 2018).

The first contributions on technology-based CI were focused on the use of tablets for note taking (Drechsel/Goldsmith 2016; Goldsmith 2018; Altieri 2020), which does not however represent a true automation of CI by means of a CAI tool. In the last few years, particular attention has been devoted to the use of Automatic Speech Recognition (ASR) tools as a support for CI (Wang/Wang 2019; Chen/Kruger 2023; Ünlü 2023). ASR-based consecutive interpreting aims to overcome typical problem triggers in note taking, such as numbers, proper names and acronyms. The availability of a transcript is supposed to ensure additional accuracy and completeness in terms of content, since the risk of missing or misunderstanding numbers or names is theoretically reduced compared to traditional note taking, which involves the activation of a number of cognitive efforts. However, questions may arise on the impact that the use of ASR, or rather an overreliance on ASR, may have in terms of reformulation and transparency of the message in the target language, communicability, ergonomics, and reliability and readability of the transcript. Besides, if transcribing a speech were key to conveying a message correctly, interpreters would only need to use shorthand writing. This is not the case, as the drawback of shorthand is that "aucune sténographie ne fixe une conversation, ni même un discours improvisé"[1] (Seleskovitch 1975: 84). In fact, in addition to the gap between oral and written language, decrypting shorthand writing entails not only a process of deciphering, but also a process of converting the message into the target language, which will inevitably slow down the interpreter.

### 1.1 Automatic Speech Recognition

Automatic Speech Recognition (ASR) is the process of converting human speech into a sequence of written words via a computer programme (Jurafsky and Martin 2009 in

1 "There is no shorthand writing that fixes a conversation, nor an impromptu speech" (Seleskovitch 1975: 84). The quotation was translated into English by the author of this article.

Fantinuoli 2017: 26). Until the late 2000s, the application of ASR to interpreting was largely the subject of theoretical speculation (Tripepi Winteringham 2010). However, since the advent of new computational approaches, the quality of ASR technologies has improved to the point where such tools can achieve an error rate comparable to that of human listeners (Cavallo/Ortiz 2018: 23). In particular, advancements in neural networks and deep learning have significantly enhanced the accuracy of ASR and have therefore driven increasing interest in applying it to interpreting (Fantinuoli 2017: 28). However, ASR is still far from being a perfect transcribing tool for interpreters, as there are both intrinsic accuracy/quality limitations and external linguistic elements that pose potential obstacles to its effectiveness, such as nuances, linguistic variation, non-verbal communication, accents, emotions, metaphors, intonation and irony (Cavallo/Ortiz Schild 2018). Additionally, ambiguity, background noise, speech rate, and body language can also affect the quality of the ASR output (Fantinuoli 2017: 28).

### 1.2 Speaker-dependent and speaker-independent speech recognition

An important criterion to categorise speech recognition systems is their adaptability to the speaker's voice. *Speaker-dependent* speech recognition systems can adapt to the characteristics and pronunciation of an individual speaker, thereby improving over time. By contrast, *speaker-independent* systems are not designed to adjust to a specific voice and are generally more effective in contexts where the message is predictable. However, such systems do not achieve the same level of accuracy as *speaker-dependent* ones. An example of a *speaker-dependent* system is *Dragon Naturally Speaking*[2], primarily used in respeaking, while *Sight-Terp* (§2.2) is a *speaker-independent* speech recognition system.

Under ideal conditions, *speaker-independent* speech recognition systems can reach an accuracy rate of 95-99% regarding term recognition. Yet, factors such as noise, music, and spontaneous oral speech affect the quality of recognition, thus representing a limitation for this technology (Fantinuoli 2017). Especially spontaneous oral speech may pose obstacles for ASR systems. Digressions, false starts and incomplete sentences, inferences and implicit references, changes in the speaker's intonation, which typically appear in such speech, are often difficult to be transcribed in an easily readable fashion. For an ASR system to be competitive and reliable for an interpreting task, three features are especially crucial:

1. Speaker-independence, enabling the ASR to adapt to a range of speakers and accents.
2. A low word error rate (WER rate), thus ensuring reliability without imposing an excessive cognitive load on the interpreter.
3. Low latency, allowing the ASR to fit in with the average delay of interpreters (3-5 seconds; Fantinuoli 2017: 28-29).

2 See https://dragon-naturally-speaking-premium.en.download.it/.

## 2. Research methodology

The aim of the present study is to investigate the interpreter-machine interaction in ASR-assisted consecutive interpreting to explore how and to what extent the ASR in this case study, Sight-Terp (§2.2), can enhance consecutive interpreting. Based on the assumption that the ASR can produce excellent results with clear and well-structured speeches (e.g. Defrancq/Fantinuoli 2020; Ünlü 2023), it was decided to investigate the effectiveness of the ASR tool in Sight-Terp when coping with spontaneous oral speech (§1.2), which is common in the interpreter's working life. Secondly, it was assumed that the use of ASR for CI reduces the cognitive load in the first phase, since the effort of note taking is partly replaced by the transcript produced by the ASR. The last assumption is that an automatic generated transcript is a useful source to enhance the interpreting accuracy regarding numbers and proper names. However, it leads the interpreters to stick closer to the text and/or to longer renditions in the target language. In fact, a transcript has a higher word density compared to traditional interpreting notes. If using an ASR tool in CI, the interpreters can choose to use the transcript either as a source for a sight translation or as a backup text for the retrieval of individual elements (such as figures and named entities, see §2.2).

Three hypothesises can be postulated as follows:

1. Interpreters can concentrate more during the source language (SL) listening and comprehension phase thanks to the transcript being produced in real time.
2. The interpreter's ASR-assisted delivery requires more time than in the traditional consecutive mode, especially when the interpreters tend to stick close to the text.
3. While the ASR can assist the interpreter in retrieving specific elements, such as numbers and proper names (or named entities, NE), it is less reliable when attempting to transcribe some features that are typical of spontaneous oral speech (such as repetitions, digressions, false starts, etc.).

### 2.1 Methodology

For this study, data triangulation was employed with the aim of observing the same phenomenon from various perspectives, thereby identifying analogies or differences in the interpreters' behaviour and their interaction with Sight-Terp. Moreover, this approach allowed us to assess the same aspects in the most objective manner possible (see Bowker 2022: 403).

To minimise potential bias, as many variables as possible were kept under experimental control. The two speeches interpreted by the participants (§2.3) contained several dependent variables, such as numbers, proper names and figurative language. These elements were distributed as evenly as possible in the texts (e.g. Seeber 2011; Prandi 2018).

## 2.2 Sight-Terp[3]

Sight-Terp[4] is the CAI tool that was used for this study, is a prototype of *speaker-independent* ASR-based CAI tools for consecutive interpreting. It was developed by Cihan Ünlü (2023) for his experimental MA dissertation at Ankara University. As a prototype of a CAI tool, Sight-Terp was employed exclusively as a case study. Therefore, the results obtained (see §4) do not attempt to come to any prescriptive conclusions about the tool itself.

The interface of Sight-Terp, which can be used on personal computers, tablets and smartphones, is divided into two text areas: on the left hand-side of the screen, Sight-Terp provides the interpreter with a full transcript of the original speech, running in real-time through continuous speech recognition, while on the right hand-side of the screen a translation in the target language is provided, running with a delay of few seconds. The aim is to replace traditional note taking with a transcript and its translation, which can eventually be used as a support for sight translation.

The production of the ASR output and the Machine Translation (MT) output is made possible by the Speech translation function (ST), also known as Automatic speech-to-text translation, which is the distinguishing trait of Sight-Terp. Such an innovative model unifies the ASR tool and the MT tool and therefore goes beyond the conventional cascade approach[5], where ASR and MT are used separately. After pressing the start recognition button, "[Sight-Terp] initiates an automatic speech recognition session using Microsoft Speech Translation API, which is based on an end-to-end system using deep neural network-based modelling" (Ünlü 2023: 67).

The average latency of the transcription has been observed as around two seconds, but it is strongly influenced by the stability of the Internet connection. The result displayed on the screen is a text segmented into utterances that are usually made up of three or four sentences. A full stop is placed at the end of an utterance, when a unit of meaning (or semantic unit) is recognised as such by the AI-based speech model incorporated in Sight-Terp. In addition to full stops and text segmentation, the *TrueText* function by Microsoft initiates a text-normalisation and automatic punctuation process. Such a process "enhances the readability of the text […] and therefore increases the accuracy rate of the neural machine translation" (Ünlü 2023: 68), on which the MT tool is based. Each utterance is then numbered and aligned with the corresponding utterance in the translation, preceded by the same number. This is how the app facilitates the retrieval of the various speech passages and their translation. Clearly, inappropriate segmentation of the speech may have significant repercussions for the

3 For a more detailed description of this tool, see Cihan Ünlü's unpublished MA Thesis "Automatic Speech recognition in consecutive interpreter workstation: computer-aided interpreting tool 'sight-terp'".

4 See https://www.sightterp.net/.

5 Current Speech translation systems are based on two approaches: the cascade and the innovative end-to-end approach. If the cascade approach is applied, the ASR tool and the MT tool work separately; once the ASR has generated its output, this serves as an input for the MT, which eventually produces the final output (the translated text). If the end-to-end approach is applied, the separation into two stages is substituted by a unified process where the source language speech is directly translated into the target language.

readability of the transcript and therefore on the interpreter's comprehension effort, as other studies have already highlighted (cfr. Wang/Wang 2019, Chen/Kruger 2023). For this reason, the *Automatic Text Segmentation* function provided by Sight-Terp "is deployed with the aim to display the reference text in an easy-to-read fashion and allow the user (interpreter) to follow up the source segment with its target MT output thanks to the enumerated style" (Ünlü 2023: 69). It must be noted that other ASR systems convert a speech into written words without paying attention to punctuation and the layout of the text (for example Dragon Anywhere[6], where punctuation marks need to be dictated by the speaker, or the ASR incorporated into Word or Google Drive). Nevertheless, there remains a drawback in the transcription by Sight-Terp: a default text segmentation setting which splits the text whenever the speaker takes a pause of two or more seconds, which can pose a potential obstacle to the correct interpretation of the text and be a potential source of error for the MT tool.

Finally, Sight-Terp offers two more features: the *Named Entity Recognition* (NER) function highlights proper names, acronyms, numbers and time references, and the *Digital Notepad* enables users to take notes in the lower part of the tablet.

### 2.3 Participants and speeches

The participants were six students[7] at the end of their two-year master's degree in Conference Interpreting at the University of Trieste. They had all accumulated more than 72 hours of practice in German-Italian consecutive interpreting during their regular classes and passed their exam "Consecutive from German into Italian 2". Two of them had German as a B-language, while the other four participants had German as a C-language[8]. As the participants were not familiar with Sight-Terp, they took part in an 11-hour seminar composed of one theoretical session (two hours) and three practice sessions (nine hours, three hours each session). The first theoretical session was focused on the presentation of Sight-Terp and on preparing for an assignment using AI-based tools, while the other three sessions gave the student-interpreters the opportunity to develop their own approach to interacting with Sight-Terp. The interpreters could consolidate their approach during these three practice sessions focused on several speeches.

6 See how Wang/Wang (2019) employed Dragon Anywhere in their study.

7 All six students took part at the present study voluntarily, they did not receive any remuneration. Before they were recruited, they had all been informed that the recordings of their deliveries would only be analysed within this study and that their names would be anonymised to guarantee the confidentiality of their personal data. Moreover, they all signed a Declaration of Consent for their deliveries to be processed for the purposes and procedures of the present study.

8 The B language is a language which the interpreters master at a level close to mother tongue. This is the language into which the interpreters can provide a fluent and accurate interpretation. The C language is the language which the interpreters can fully understand and from which the interpreters can interpret into their mother tongue. The definitions above refer to the European language profiles described on the Internet page of the European Union (see https://europa.eu/interpretation/freelance.html).

The data collected refers to two speeches, which the interpreters translated from German into Italian. The former, entitled "*Niassa Naturreservat*", focuses on the Niassa natural reserve in Mozambique, and the latter, "*GAFAM-Unternehmen*" focuses on the biggest tech companies worldwide. Both speeches were prepared by interpreting students, reviewed by an interpreting student who was a German native speaker and delivered by a native speaker from Germany, who was asked not to stick to the written words and to deliver the speech naturally and vivaciously.

| | Length (words) | Duration (mins) | Speech rate (words/min) |
|---|---|---|---|
| Speech A | 621 | 5'35" | 110 |
| Speech B | 610 | 6'02" | 101 |

Tab. 1. Features of the interpreted speeches.

| | Figures | Named entities |
|---|---|---|
| Speech A | 13 | 9 |
| Speech B | 10 | 27 |

Tab. 2. Target elements (figures and named entities) of the interpreted speeches.

## 2.4 Data collection procedure

During the practice sessions leading up to the data collection, no technique nor strategy was suggested, since one of the aims was to observe how the interpreters would deal with ASR-assisted consecutive interpreting in the classroom.

In the second phase of this study, the interpreters were asked to interpret two speeches from German into Italian, one using traditional note taking and one using Sight-Terp. The data collection process took place on the 30th of November and the 1st of December 2023. The data collection was spread over two days so that the participants would have to provide only one interpreting performance per day, thereby decreasing the risk that certain variables (such as stress or fatigue) could affect the quality of the second performance.

Since it was decided to reproduce a typical interpreting assignment, the participants were provided with a topic briefing for each speech. The briefings consisted of an introduction on a fictional conference setting, an abstract of the speech in Italian and in German, some in-depth insights on the main topic, and a thematic glossary (35 entries for speech A and 42 entries for speech B). The interpreters had ten minutes to read the briefing and prepare for the assignment. Before the interpreting task, the interpreters were asked to use Sight-Terp by adopting the approach they personally developed during the practice sessions. The above conditions remained unchanged on both days.

The group of six student-interpreters was randomly divided into two subgroups: the first subgroup interpreted speech A with Sight-Terp and speech B with the help of the notes taken by hand, and the second group did the opposite. The random distribu-

tion of participants into two subgroups made it possible to counterbalance the variable related to the intrinsic features of the speeches and/or of the interpreters and look at the data in most rigorous way possible (Gile 2016: 220). Moreover, the random distribution of independent variables is one of the aspects of experimental design that makes it possible to obtain "maximum efficiency [and] reduce the effect of confounding variables and bias" (*Ibid.*).

Since Sight-Terp is based on a *speaker-independent* ASR and its performance strongly depends on a stable Internet connection, the speeches were pre-recorded and played in a quiet room with adequate internet access. This enabled Sight-Terp to transcribe them with maximum efficiency. The transcription and the MT output by Sight-Terp were also pre-recorded on a tablet screen. Finally, the source language recordings and screen recordings were synchronised. The screen recording allowed to avoid potential differences in the transcription or possible disruption of the app, which would compromise the entire data collection: for instance, a suddenly slower Internet connection or sudden background noises might have resulted in different ASR outputs if Sight-Terp had been used in real-time with each individual participant. This approach allowed all participants to be exposed to the same stimuli and therefore to use the same transcript (i.e. screen recording) as an aid to their consecutive performances.

### 2.5 Focus group discussion and questionnaire

As the number of participants did not warrant quantitative generalisations, a focus group discussion was organised at the end of the data collection phase (e.g. Frittella 2023: 67). This method proved extremely relevant within this study, since the focus group discussion highlighted qualitative aspects of ASR-assisted consecutive interpreting, such as strategic and technical elements. The discussion was moderated by thought-provoking questions (for instance, what was your approach to the use of Sight-Terp in assisted consecutive interpreting? What did you use the transcript for? How did Sight-Terp influence your performance? What type of interpreting strategies did you use in assisted consecutive interpreting?). The discussion also provided insights into the psychological impact of using a transcript for consecutive interpreting. The main topics concerned the strategies adopted to look at notes and to check ASR transcript when delivering the target language speech, as well as strategies to cope with the limitations of ASR and difficulties arising from its use in consecutive interpretation.

After data collection and the focus group discussion, the interpreters were asked to evaluate their experience through a questionnaire. The questionnaire was distributed via Google Forms and consisted of 31 questions of a technical, interpreting and strategic nature – these were thematically divided into five sections:

1. Preparing for an assignment with AI.
2. Sight-Terp-assisted consecutive interpreting.
3. Interaction with Sight-Terp.
4. The performance of Sight-Terp.
5. Future prospects.

Both open and closed-ended questions were adopted to find out how Sight-Terp was used and what strategies were deployed by the interpreters, and to investigate to what extent they were satisfied with their own performance, the transcript produced by Sight-Terp, and Sight-Terp's functionalities.

### 2.6 Limitations of the study

The first limitation of the present study is the number of participants, which is often one of the major limitations of experimental interpreting studies. Moreover, to collect reliable data on ASR-assisted consecutive interpreting, the participants were given the opportunity to practice with this interpreting approach during three sessions, for a total of nine hours (§2.3). The findings obtained must therefore be interpreted with reference to the present study and may not be applied to the whole category of professional interpreters, especially because the participants were six final-year students (§2.3). In addition, the number of professionals who use such tools in their work is still believed to be very limited.

A second limitation is the participants' limited preparation. More targeted training would potentially yield better results in the assisted consecutive approach.

The third limitation is the very fact that Sight-Terp itself was used. Hence, the findings may not be applied to all ASR systems, since at this stage Sight-Terp is still a prototype of a freely accessible *speaker-independent* ASR system. The results of this study cannot be compared to those obtained via a specialised *speaker-dependent* ASR system (such as Dragon Naturally Speaking). However, within this study the key aspect is not really the quality of the ASR transcript, but rather its suitability for CI.

## 3. Results and Analysis

### 3.1 Modes of use

The interpreters' behaviour during the data collection phase was observed at the same time as they interpreted both speeches and subsequently investigated through the answers collected in the questionnaire. Three different modes of using Sight-Terp stood out.

Interpreters 1 and 5 took their notes on a traditional notepad for CI and looked up names, time references or numbers in the transcript produced by Sight-Terp. Their rendition was then generally based on the ASR transcript for short extracts and on their notes for longer extracts.

Interpreters 3 and 4 took their notes on a paper notepad to analyse and process the speech. They then used the transcript to sight translate the text, which was more comprehensive than their notes. Occasionally, they looked at their notes to decipher extracts which were not clear in the transcript.

Interpreters 2 and 6 based both the first and second phase of CI on Sight-Terp without taking any traditional interpreting notes, although they used a paper notepad to note down key concepts and/or transcription errors, such as segmentation or formatting errors (see §3.3).

The final questionnaire confirmed these three approaches, as it revealed that Sight-Terp was used by each interpreter, albeit in different manners and proportions. Out of six student-interpreters, four used Sight-Terp frequently, while two made little use of it. Moreover, five interpreters reported that they mostly used Sight-Terp in the rendition phase of CI (two of them also used it in the phase of listening and analysis). By contrast, one student mainly used Sight-Terp in the first phase; namely, to compare the notes with the transcript or to retrieve missing information. Regarding the use of the transcript and the MT output, five students clearly preferred to use the transcript either to look up individual elements (figures, named entities, terminology) or to perform an on-sight translation. This trend is also confirmed by the fact that the interpreters prevalently used Sight-Terp to retrieve numbers and information they had missed, followed by proper names and lastly by terminology. Only one student also tried to integrate the MT output in the rendition phase. The interpreters evaluated the interaction with Sight-Terp as, on average, 5.5 on a scale out of ten, where ten stands for "easy interaction". One interpreter evaluated the interaction as "one", representing "too difficult" on the scale.

In terms of human-machine interaction, Sight-Terp appears to have been a distracting factor for four interpreters out of six. This was also confirmed in the focus group discussion; indeed, interpreters 2 and 6 stated that the awareness of having a complete transcript at their disposal distracted them during the listening and analysis phase. Similarly, interpreter 5 described a sort of "psychological burden". In the focus discussion she described a feeling of nervousness before the traditional consecutive performance, as is often the case, but this feeling disappeared during the performance itself. By contrast, the opposite occurred in the ASR-assisted consecutive performance: the knowledge of having a full transcript, which is more comprehensive than the interpreting notes, was comforting before the beginning of the performance. However, the effort of looking at Sight-Terp and the worry of watching out for possible transcription errors resulted in feelings of nervousness during the consecutive performance itself.

Moreover, according to two out of six interpreters, Sight-Terp had an adverse impact on eye contact. According to four out of six, it had an impact that was more negative than positive, with five out of six interpreters claiming that Sight-Terp adversely affected the fluidity of their rendition. However, the interpreters had a positive attitude towards the use of ASR for CI, as they stated that more practice with it could eventually lead to better performances in assisted consecutive interpreting. Out of six interpreters, only one had a negative attitude to the use of ASR in consecutive interpreting. In addition, a deeper understanding and awareness of the strengths and limitations of ASR could potentially have a positive impact on the performance.

### 3.2 Interpreters' renditions

The following table compares the duration of the original speeches with that of the interpreters' renditions.

| Interpreter | Original speech (mins) | Rendition (mins) | Mode | Time difference (mins) |
|---|---|---|---|---|
| Int. 1 | 5’35” | 5’13” | ST | -0’22” |
| Int. 2 | | 5’16” | CONSEC | -0’19” |
| Int. 3 | | 4’20” | CONSEC | -1’15” |
| Int. 4 | | 4’22” | ST | -1’13” |
| Int. 5 | | 4’48” | CONSEC | -0’48” |
| Int. 6 | | 7’07” | ST | +1’32” |

Tab. 3. Duration of speech A and duration of interpreters’ renditions. ST = Sight-Terp assisted consecutive, CONSEC = traditional consecutive

| Interpreter | Original speech (mins) | Rendition (mins) | Mode | Time difference (mins) |
|---|---|---|---|---|
| Int. 1 | 6’02” | 5’42” | CONSEC | -0’20” |
| Int. 2 | | 6’22” | ST | +0’20” |
| Int. 3 | | 5’58” | ST | -0’04” |
| Int. 4 | | 4’45” | CONSEC | -1’08” |
| Int. 5 | | 5’14” | ST | -0’48” |
| Int. 6 | | 4’52” | CONSEC | -1’10” |

Tab. 4. Duration of speech B and duration of interpreters’ renditions. ST = Sight-Terp assisted consecutive, CONSEC = traditional consecutive

In the traditional consecutive approach, the duration of the interpreters’ renditions was always shorter than the original speech. By contrast, two interpreters produced a longer rendition than the original speech when using Sight-Terp for their consecutive performance (interpreter 6 Tab. 3, interpreter 2 in Tab. 4). Interestingly, these were the two interpreters who exclusively relied on Sight-Terp for their rendition, thus performing a sight translation.

| Time difference in CONSEC | Median | Time difference in ST | Median |
|---|---|---|---|
| -0’04’’ | (-0’20’’) + (-0’48’’): 2 = *0’34”* | +1’32’’ | (-0’22’’) + (-0’48’’): 2 = *0’40”* |
| -0’19’’ | | +0’20’’ | |
| *-0’20”* | | *-0’22”* | |
| *-0’48”* | | *-0’48”* | |
| -1’08’’ | | -1’08’’ | |
| -1’10’’ | | -1’15’’ | |

Tab. 5. Calculation of the median of time difference. ST = Sight-Terp assisted consecutive, CONSEC = traditional consecutive

Since the values of time difference noticed in consecutive interpreting performances and in Sight-Terp assisted performances were heterogeneous and the number of participants was limited, a calculation of the average time difference would not be sufficiently representative. Instead, the median of time difference was calculated, which can be considered as the middle value (Tab. 5)

Generally speaking, the use of ASR for CI does not always result in a longer rendition than the original speech. However, it can be observed that the interpreters' renditions in assisted consecutive interpreting tend to be longer compared to their performances in the traditional consecutive mode. The median of time difference is 34 sec. in the traditional mode, while the median of time difference is 40 sec. in the assisted consecutive mode.

### 3.3 The performance of Sight-Terp

| | WER | Figures | Proper names | Incorrect segmentation |
|---|---|---|---|---|
| Speech A | 13,6% | 9/13 | 10/18 | 9 |
| Speech B | 8,2% | 8/10 | 23/28 | 12 |

Tab. 6. The performance of Sight-Terp

In speech A, the ASR correctly transcribed nine figures out of thirteen. The remaining four figures were considered as not immediately readable, rather than errors as such. Indeed, two figures were transcribed in full words (*4.500 Quadratkilometern ✍ viertausendfünfhundert, 168 Milliarden ✍ Einhundertachtundsechzig Milliarden*), while two figures were not immediately readable because of their format (*100.000 Dollar ✍ 100 00 $0, 50.000 Dollar ✍ 50 00 0€*). On one occasion, the ASR correctly transcribed the figure, but not its reference (*50% des Einkommens ✍ 50% des Einkaufs*). However, since the ASR system in Sight-Terp does not include the joint recognition of figures and their references, the latter figure was not considered as an error. In speech B, the rate of correctly transcribed figures amounts to 8/10. Once again, one figure was transcribed in full words (*über 130.000 Beschäftigte ✍ über einhunderdreißigtausend Beschäftigte*) and another was partially transcribed (*über 1 Milliarde Stunden ✍ Milliarde Stunden*).

With regard to named entities (NE), here the ASR system seems to have produced reliable results, particularly when transcribing well-known NE (*Afrika, Mosambik, Google, YouTube, Apple, Airbnb*). However, on two occasions there were errors (*Mosambik ✍ Osnabrück, NGO ✍ MGO*). Conversely, less widely known NE, such as the name *Niassa*, the acronyms NATU and GAFAM were not correctly transcribed. In such cases the transcript generated words which are phonetically similar to the original ones (*Gafam ✍ Gafar, Gafan, NATU ✍ Lato, Niassa ✍ erster, nasser, Iassa*).

When the accuracy results on figures and NE are correlated with the interpreters' satisfaction, it appears that the aforementioned errors did not particularly destabilise them. Indeed, according to the questionnaire, the interpreters did not feel that the transcription caused them to commit errors they would have not committed in the

traditional consecutive mode. Incorrect numbers and names, as well as omitted words, did not result in errors in the interpreters' renditions.

For the quality assessment of the transcripts, the Word Error Rate (WER) was calculated using *Amberscript*[9], a software programme which also offers an automatic WER tool. The WER values reported in Tab. 6 indicate that the transcript produced by Sight-Terp is definitely not flawless, but not bad either. Yet, the WER itself is not sufficient to determine whether a transcript is suitable for the purpose of interpreting. For this reason, it was necessary to examine how and where the ASR segmented the text. The ASR feature in Sight-Terp generally produces complete sentences. However, the ASR segmented certain utterances incorrectly: for instance, by separating the subject from its verb. Wrong segmentation occurred more often when the speaker paused to plan the sentence or self-correct, made digressions or had false starts, thereby making it harder for the ASR to transcribe spontaneous speech in an easily readable fashion. This drawback, which in itself is one of the current limitations of ASR, stems from the distance between mostly unplanned spoken language and well-structured written language. Although the readability of the text was not significantly compromised, incorrect segmentation may require more effort, and thus more time, to make sense of the transcribed speech.

By comparing the above-mentioned results with the results of the questionnaire, the relatively good quality of the ASR transcript was also confirmed by the participants: four interpreters reported being satisfied with the transcript quality, and two of them very satisfied. Moreover, according to the questionnaire, three interpreters tended to rely on the transcript, while the other three tended not to rely on the same output, and therefore always preferred to double-check their speech comprehension by means of their traditional notes.

Furthermore, sudden changes in the speech rate or in the speaker's intonation seem to have affected not only text segmentation, but also transcription, leading to errors or omissions of single words (e.g. *in meinem heutigen relativ kurzen Beitrag möchte ich auf eine Frage [eingehen]*), which may also explain the incorrect transcription of well-known named entities (e.g. *Mosambik, NGO*). This is an undeniable limitation of current ASR systems, which should always be taken into account, especially when using such systems in authentic work settings, where sound disruptions or background noise could influence the performance of such CAI tools. As regards the interpreters' opinion about text segmentation and incorrectly transcribed or omitted words, it appears that four interpreters felt they did not commit any errors related to the use of the transcript. One interpreter excluded this possibility, and one interpreter got the impression that some errors could have been avoided in the traditional consecutive mode.

## 4. Discussion

This section discusses the results presented in section §3, in an attempt to give an answer to the hypotheses formulated in section §2.

9 See https://www.amberscript.com/en/resources/wer-tool/.

### 4.1 Listening and comprehension skills

As indicated by the questionnaire, the transcript produced by Sight-Terp is mostly employed as an aid in the renditions of the interpreters, who looked up figures or performed an on-sight translation. One student-interpreter used the transcript mostly to integrate information in note taking or to double-check that the notes had been correctly understood. Otherwise, the transcript was generally perceived by the interpreters as a potential distraction, because it often forced them to weigh up and monitor the solutions given by the ASR. More specifically, during the focus group discussion interpreter 2 and interpreter 6 stated that knowing that a full transcript is available may inadvertently disrupt their listening. This, in turn, adversely affects both speech comprehension and the thorough analysis of its logical structure.

By contrast, four interpreters out of six used Sight-Terp as a support for the retrieval of numbers and sometimes names, without giving up their traditional note taking, which was confirmed to be necessary for deeper comprehension. The transcript itself proved to be a reasonably reliable source of content, which was, however, not a product of the interpreter's elaboration; as such, the meaning of a text with a high word density is not as immediately transparent as the interpreter's notes, which are the product of an in-depth internalisation of the speech.

Therefore, the first hypothesis was disproved, as the participants were not able to concentrate more on listening and comprehension, but instead needed to coordinate additional efforts both in stage 1 and stage 2 of CI. This was particularly evident in the interpreters who decided to rely exclusively on the transcript produced by Sight-Terp.

### 4.2 Speech rendition duration

Out of six ASR-based consecutive interpreting performances, two renditions turned out to be longer than the original speech, whereas all the traditional consecutives were shorter. Therefore, the use of an ASR-generated transcript cannot be said to lead to a longer rendition every single time, but ASR-enhanced consecutives were always slightly longer compared to the traditional ones.

The two longer renditions were produced by interpreter 6 and interpreter 2 (see Tab. 3 and Tab. 4), who based their interpretation exclusively on the Sight-Terp transcript. It can be concluded that both did not process the speech as thoroughly as in traditional consecutive interpreting via note taking. Moreover, during the focus group discussion both interpreters mentioned feeling distracted because they knew they would have a full transcript at their disposal. Therefore, it can be concluded that the lack of a deeper analysis of the speech must have affected their listening and comprehension. This was subsequently compensated for by the attempt to reconstruct the sense of the transcript and to rephrase it during the second phase of consecutive interpreting. Indeed, it must be considered that traditional note taking not only enables a thorough analysis of the speech, but also a preliminary translation process. By contrast, the interpreter who chooses to rely exclusively on the ASR transcript to perform a sight translation, shifts the entire translation process to the rendition phase, therefore requiring more effort. The accumulation of additional effort in the second phase of

CI, which should only entail the free production of the target language speech, may result in an increase in the length of the interpreters' rendition (see interpreter 2 and interpreter 6 and their approach to Sight-Terp, §3.1). Thus, similar to simultaneous interpreting, a tightrope effect (Gile 1999) can be observed in the second phase of assisted consecutive interpreting.

As regards the other interpreters, the use of a transcript may have resulted in an attempt to be more exhaustive and to stick to the transcribed text relatively closely (i.e. trying to reproduce all its parts), because of the psychological burden they felt, as was mentioned by the interpreters during the focus group discussion.

In conclusion, the second hypothesis was partly confirmed by the results. The renditions produced in the assisted consecutive mode are not always longer than the original speech, and a key role is played by the way the ASR tool is employed. However, in comparison with the renditions produced in traditional consecutive interpreting, ASR-assisted consecutive interpreting entails a slight duration increase of the speech rendition.

### 4.3 Readability of automatic transcription

As expected, the ASR feature in Sight-Terp produced acceptable transcripts. Indeed, the positive WER rates (Tab. 6) are also reflected in the interpreters' satisfaction with the transcription: four interpreters reported being satisfied, and two of them very satisfied. The same is also valid for the recognition of numbers and proper names, despite some differences in the accuracy rates between speech A and speech B (Tab. 6).

Nevertheless, in terms of its application to interpreting, the transcripts must be examined in their entirety, taking into account the coherence and cohesion of the texts, as well as their readability. The text segmentation function in Sight-Terp aims to spread the text contents into enumerated chunks, thus distributing the information load. Yet, this very function posed an obstacle to readability, especially when syntactic units were segmented in the wrong place, thereby making the retrieval of the logical structure more difficult. Incorrect segmentations can be attributed to long pauses taken by speakers (pauses of reflection or related to sentence planning efforts), digressions or false starts. Furthermore, emphasis must be placed on the linearity of the transcription process. Unlike a human interpreter, the ASR transcription proceeds in a linear sense, and therefore reproduces the original speech as closely as possible, including its irregularities. Interpreters who choose to rely on an insufficiently transparent transcript run the risk of transferring the task of inferring the deep meaning of the speech to the audience, whose understanding will be hindered rather than facilitated.

It is worthwhile mentioning that transcription errors affecting individual (especially well-known) words or their individual omission usually belonging to collocations did not pose an obstacle to the participants. This can be explained in light of the interpreters' language skills and the briefing, which gave the interpreters the possibility to prepare themselves and anticipate the key ideas of the speech, as would be the case in a real setting. It is believed that interpreters with a good level preparation are able to spot general errors committed by the machine immediately, especially if such errors are likely to be anticipated. In the example "in meinem heutigen relativ kurzen Beitrag

möchte ich auf eine Frage [eingehen]"[10], the only verb that is supposed to belong to this collocation is the verb "eingehen"[11].

In conclusion, the third hypothesis concerning the way ASR deals with spontaneous oral speech was confirmed, since ASR is more accurate than a human being in the retrieval and transcription of technical elements, such as numbers and proper names.

## 5. Conclusion

The present study identified three approaches to performing an ASR-based consecutive interpretation. In all these approaches, an ASR tool is applied to transcribe the source language speech in the first stage of consecutive interpreting, while its output (the transcript) is used by the interpreters in the second stage of CI, either as a source text for a sight translation or as a backup text for the traditional interpreting notes. Both the potential and drawbacks of this approach were discussed, including whether the ASR transcript might be a useful aid to enhance consecutive interpreting (i.e. for the retrieval of numbers and named entities), and whether an overreliance on it can inadvertently hinder the interpreter's performance (less fluency, longer rendition, additional analysis and rephrasing effort). However, the use of a transcript produced by an ASR tool can ultimately enhance consecutive interpreting under certain conditions: for instance, if such tool is applied in a consecutive setting with a stable Internet connection and no background noises, or if it is used as a backup for note taking, which in fact enables a thorough speech analysis.

The interpreting profession is undeniably at a turning point, as new technologies, such as AI and Automatic Speech Recognition (ASR), are significantly influencing the way interpreting is performed. While simultaneous interpreting has been the focus of a number of studies on CAI tools, new ways of performing consecutive interpreting potentially involving Automatic Speech Recognition and/or Machine Interpreting are being experimented with. It is thus imperative that interpreters develop a deeper understanding of the new tools available to develop strategic approaches for their use, especially as they still present some intrinsic limitations. New studies in this field can give the opportunity to continue exploring how CAI tools can be strategically integrated into the interpreting workflow, without adversely affecting the process, but rather offering interpreters actual advantages. Only with their active engagement can interpreters uphold quality and working standards in their profession.

10 "In my relatively short speech today, I would like [to address] a question". The translation into English of the original German example was done by the author of this article.

11 For non German-speakers: an important strategy in the interpretation from German into a romance language is the anticipation of the verb at the end of the sentence. An extensive and solid knowledge of German collocations is crucial to anticipate the meaning of the sentence before the speaker concludes it. In the example reported above, the interpreters who work from German do not need to wait too long to understand that the verb belonging to "auf eine Frage …" is almost certainly "eingehen".

References


Altieri M. (2020) "Tablet interpreting: étude expérimentale de l'interprétation consécutive sur tablette ", *The Interpreters' Newsletter* 25/2020, 19-35.

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

Berber-Irabien D. (2010) *Information and communication technologies in conference interpreting*, unpublished PhD Thesis, Terragona, Universitat Rovira I Virgili.

Bowker L. (2022) "Computer-assisted translation and interpreting tools", in F. Zanettin / C. Rundle (eds.) *The Routledge Handbook of Translation and Methodology*, London, Routledge, 392-409.

Cavallo P. / Ortiz Schild L.E. (2018) "Computer-assisted interpreting tools (CAI) and options for automation with Automatic Speech Recognition", *TradTerm*, 32/2018, 9-31, <https://www.researchgate.net/publication/330207613_Computer-Assisted_Interpreting_Tools_CAI_and_options_for_automation_with_Automatic_Speech_Recognition> (28 October 2024).

Chen S. / Kruger J. (2023) "The effectiveness of computer-assisted interpreting. A preliminary study based on English-Chinese consecutive interpreting", *Translation and Interpreting Studies*, 18/3, 399-420.

Chernov S. (2016) "At the dawn of simultaneous interpreting in the USSR. Filling the gaps in history", in J. Baigorri-Jalón, K. Takeda (eds.) *New Insights in the History of Interpreting,* Amsterdam/Philadelphia, John Benjamins, 135-166.

Corpas Pastor G. / Fern L.M. (2016) *A survey of interpreters' needs and practices related to language technology*, <http://www.lexytrad.es/assets/Corpas-Fern-2016.pdf> (28 October 2024).

Defrancq B. / Fantinuoli C. (2020) "Automatic Speech Recognition in the booth: Assessment of system performance, interpreters' performances and interactions in the context of numbers", *Target*, 33/1, 73-102.

Drechsel A. / Goldsmith J. (2016) *Tablet Interpreting. The evolution and uses of mobile devices in interpreting*, <https://www.academia.edu/36017504/Tablet_Interpreting_The_evolution_and_uses_of_mobile_devices_in_interpreting> (28 October 2024).

Fantinuoli C. (2016), "InterpretBank. Redefining computer-assisted interpreting tools", in J. Esteves-Ferreira / J.M. Macan / R. Mitkov / O.M. Stefanov (eds.) *Proceedings of the 38th Conference Translating and the Computer*, Geneva, Tradulex, 42-52, <https://aclanthology.org/2016.tc-1.5.pdf> (28 October 2024).

Fantinuoli C. (2017) "Speech Recognition in the Interpreter Workstation", in J. Esteves-Ferreira / J.M. Macan / R. Mitkov / O.M. Stefanov (eds.) *Proceedings of the 39th Conference Translating and the Computer*, Geneva, Tradulex, 25-34, <https://www.asling.org/tc39/wp-content/uploads/TC39-proceedings-final-1Nov-4.20pm.pdf> (28 October 2024).

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

Fantinuoli C. (2021) “Conference interpreting and new technologies”, in M. Albl-Mikasa / E. Tiselius (eds.) *Routledge Handbook of Conference Interpreting*, Routledge, 508-522.
Frittella F.M. (2023) *Usability research for interpreter-centred technology. The case study of SmarTerp*, Berlin, Language Science Press.
Gaiba F. (1998) *The origins of simultaneous interpretation. The Nuremberg Trial*, Ottawa, University of Ottawa Press.
Gile D. (1999) “Testing the Effort Models’ tightrope hypothesis in simultaneous interpreting – A contribution”, *HERMES*, 12/23, 153-172.
Gile D. (2016) “Experimental research”, in C.V. Angelelli, B.J. Baer (eds) *Researching Translation and Interpreting*, New York, Routledge, 220-228.
Goldsmith J. (2018) “Tablet interpreting. Consecutive interpreting 2.0”, in N.K. Pokorn / C.D. Mellinger (eds.) *Community Interpreting, Translation, and Technology*, Special Issue of *Translation and Interpreting Studies* 3, 342-365.
Herbert J. (1980/1952) *Manuel de l’interprète. Comment on devient interprète de conférence*, Genève, Librairie de l’université Georg.
Kalina S., Ziegler K. (2015) “Technology”, in F. Pöchhacker, N. Grbić, P. Mead, R. Setton (eds.) *Routledge Encyclopedia of Interpreting Studies*, New York, Routledge, 410-412.
Moser-Mercer B. (1992) “Terminology Documentation in Conference Interpretation”, *Terminologie et Traduction*, 2/3, 285-304.
Prandi B. (2018) “An exploratory study an CAI tools in Simultaneous Interpreting: theoretical framework and stimulus validation”, in C. Fantinuoli (ed.) *Interpreting and technology*, Berlin, Language Science Press, 29-59.
Prandi B. (2020) “The use of CAI tools in interpreter training: where are we now and where do we go from here?”, Special Issue of *inTRAlinea: Technology in Interpreter Education and Practice*, <https://www.intralinea.org/specials/article/the_use_of_cai_tools_in_interpreter_training> (28 October 2024).
Prandi B. (2023) *Computer-assisted simultaneous interpreting. A cognitive-experimental study on terminology*, Berlin, Language Science Press.
Riccardi A. (2003), *Dalla traduzione all’interpretazione. Studi d’interpretazione simultanea* (*From translation to interpretation, studies on simultaneous interpreting*), Milano, Led.
Seeber K.G. (2011) “Cognitive load in simultaneous interpreting: Existing theories – New models”, *Interpreting*, 13/2, 176-204.
Seleskovitch D. (1975) *Langage, langues et mémoire : étude de la prise de notes en interprétation consécutive*, Paris, Lettres Modernes.
Stoll C. (2001) “Neue Technologien im Konferenzdolmetschen: DigiLab, ein Forschungsprojekt des Instituts für Übersetzen und Dolmetschen der Universität Heidelberg” (“New Technologies in Conference interpreting, DigiLab, a Research project of the Institute for Translation and Interpreting at Heidelberg University”), in A.F. Kelletat (ed.) *Dolmetschen*, Frankfurt am Main, Peter Lang.
Tripepi Winteringham S. (2010) “The usefulness of ICTs in interpreting practice”, *The Interpreters’ Newsletter*, 15/2010, 87-99.

Ünlü C. (2023) *Automatic Speech recognition in consecutive interpreter workstation: computer-aided interpreting tool 'sight-terp'*, unpublished MA Thesis, University of Ankara.

Wang C. / Wang X. (2019) "Can computer-assisted interpreting tools assist interpreting?", *Transletters. International Journal of Translation and Interpreting*, 3, 109-139.