9 Measuring the Impact of Automatic Speech Recognition on Number Rendition in Simultaneous Interpreting

Elisabetta Pisani and Claudio Fantinuoli

Introduction

Simultaneous interpretation (SI) is the process of translating speech in real-time, with a partial overlap between the listening and the production phase. From a cognitive perspective, SI is a highly demanding task that requires active listening, translating, and monitoring processes. Because of the high cognitive complexity that is involved while performing this task, many co-occurring aspects can contribute to reduced interpreting performance.

The aspects that may constitute a challenge during interpreting have been the focus of research for many years. Among others, much effort has been devoted to understanding detrimental aspects, such as the knowledge gap between speaker and interpreter (e.g. Will, 2007), the speed of delivery (e.g. Meuleman & Van Besien, 2009), and memory constraints (e.g. Moser-Mercer, 2000; Liu et al., 2004), to name just a few. In this context of cognitive complexity, computers have been proposed lately as an instrument that may support interpreters in alleviating some of the pre-process tasks (e.g. Stoll, 2009; Fantinuoli, 2012). Given the fact that computers are ideally suited for the task of recalling items and that they have the ability to store large amounts of information and process it very quickly, it has been hypothesised that, if properly integrated into the interpreting workflow, they may prove helpful not only for the activities that precede the interpreting process proper, such as event preparation (e.g. Fantinuoli, 2018; Xu, 2018), but also as an instrument to be used while interpreting, i.e. for in-process tasks. In particular, computer tools equipped with automatic speech recognition have been proposed as an ideal companion to support the interpreter with real-time suggestions for several problem triggers identified in the literature, such as numbers, terminology, and named entities (Fantinuoli, 2017).

The development of such tools is still in its infancy, and consequently, little is known about how they may perform in real-life applications or, more importantly, how well or badly they will fit into the complex cognitive process of SI. Many questions need to be answered. For example, will the integration of the visual suggestions require too much cognitive load and, consequently, will it lead to a deterioration of the overall performance? Or will the alleviation and the reduction of other cognitive tasks thanks to the suggestions rebalance the added cognitive load introduced by this technology, and lead to an improvement of performance? How will training and experience, i.e. the development of advanced strategies in how to use such suggestions, change such performance?

This research sets out to give some preliminary answers to some of these questions. In particular, the research empirically tests the influence of ASR in the performance of interpreters, analysing a particular kind of problem, the simultaneous interpretation of numbers. The experimental study compares the performance of two groups of participants interpreting a speech dense in numbers, one with the aid of a supporting tool, the other without any form of technological support. Contrary to similar studies conducted to date that were based on a simulated transcription (e.g. Desmet, 2018; Canali, 2019), in our experimental design, we use a real-life ASR-enhanced CAI tool. The experiments performed with a mockup system have the clear advantage of granting better control of the variables at stake since they can simulate a controlled and stable transcription, both in terms of error rates and latency. However, the use of a real-life tool has the advantage of improving the ecological validity of the experiment. In this context, the typical issues of ASR, such as mistranscriptions, latency, etc., which are not eliminated by the experimental design, allow us to draw some conclusions considering both the product of the interpreting process as well as the potential and limits of the state-of-the-art of ASR technology. The present experiment is similar to the one described by Defrancq and Fantinuoli (2020), except for the way suggestions are presented to the user. In Defrancq and Fantinuoli, they were presented as embedded in the complete transcription; in our experiment, they are shown isolated with no embedded context. Since the way results are presented is central in terms of usability in a cognitive-intense activity, this study adds on the findings of Defrancq and Fantinuoli and allows for a more complete picture of user–machine interaction in the context of real-time CAI support. For the analysis of the results, we use a mixed-mode approach, combining both quantitative data and introspections by means of a questionnaire.

Computer-Assisted Interpreting Tools and Automatic Speech Recognition

Computer-assisted or computer-aided interpreting (CAI) is commonly known as a form of human speech translation in which some aspects of the interpreting task are supported by a computer programme (Fantinuoli, 2018). The software used to augment the interpreter's work is named a CAI tool. In this context, CAI tools can be all sorts of programmes and mobile applications specifically designed and developed to support interpreters in at least one of the several sub-processes of interpretation, for example, knowledge acquisition and management, lexicographic memorisation, terminology lookup, and so forth.

One of the most peculiar features of CAI tools is the ability to support the search for specialised terminology in the booth. This functionality is generally designed as a backup strategy when other interpreting strategies, such as paraphrasing or the use of synonyms, are not viable and would lead to miscommunication and to a general degradation of the interpreter's performance. Interpreters may look up a term, generally in an event-specific database, while interpreting, while helping the boothmate or simply during the pauses, perhaps to find the translation of a recurring term used in a previous speech.

While CAI tools have been designed ergonomically and aim at reducing the cognitive effort needed to start a search and retrieve the results, one of the main limits of such tools is that they require the interpreter to allocate a specific amount of cognitive capacity to perform this task and to integrate the result of such operation into their delivery. Considering the fact that simultaneous interpretation is a cognitively demanding task that is in general performed at the limit of cognitive saturation, a technological means to automate the lookup mechanism could have the potential to reduce the cognitive load, benefiting the whole interpreting process. In this context, the integration of automatic speech recognition to automatise the lookup process may increase the usability of CAI tools. ASR has been regarded as a technology "with considerable potential for changing the way interpreting is practiced" (Pöchhacker, 2016, p. 188). Different to classic CAI tools that require manual input to get a translation for a given terminological unit, an ASR-enhanced CAI tool is able to automatise this process, with obvious advantages at the level of human–machine interaction.

InterpretBank ASR (Fantinuoli, 2017) is a prototype of a web-based ASR-enhanced CAI tool that transcribes in real-time a speech delivered by a speaker and automatically provides the interpreter with translation candidates of terminology as well as with numerals and their units of measurement. The tool's workflow is straightforward: firstly, the acoustic signal that the interpreter receives in the headset is sent to the sound card of the computer equipped with the ASR-CAI tool. The audio signal is then sent to the server with the speech recognition engine which returns the real-time transcript of the speech. For the ASR service, InterpretBank uses the Google Cloud Speech-to-Text API, since experimental tests have shown that, compared to other competitors, it provides the best transcription quality for features useful for CAI integration, such as specialised terminology and numbers (Brüsewitz, 2019). Secondly, the transcription stream is processed. This phase involves chunking the text stream into units of n-words of a fixed size and normalising it. Thirdly, for each n-words window, the units of interest (UI) are extracted: single- and multiword grams are looked up in the terminological database loaded in the tool or translated by means of machine translation and numbers and their units of measurement are detected. In this phase, predictive algorithms could be used to intelligently select the UI and increase the usability of the tool (cf. Vogler et al., 2019). Finally, the extracted data is visualised on the computer's monitor.

In order to empirically study different approaches to human–machine interaction, InterpretBank ASR has been designed with two main models of data visualisation. The first model spots terminological units and numbers and displays the entire speech transcript with highlighted UI. The rationale behind this is that the informational context of the UI may help the interpreter to disambiguate the information, for example, in terms of co-references. The second model, instead, suppresses the text stream and visualises only the extracted UI (terminology and numbers + units of measurement) in a vertical prompt, with the newest information highlighted on top. As a tradeoff for this loss of contextual information, the user is presented with less information load, which can potentially decrease the distraction factor and the risk of cognitive overload.

The quality of an ASR-enhanced CAI tool depends on the quality of the transcription provided by the ASR and the ability of the CAI tool to retrieve and identify the necessary information (Fantinuoli, 2017). Ideally, the system should be characterised by high recall, i.e. it should be able to recognise all transcribed units of interest (regardless of spelling variants, for example, in the case of terminology), and by high precision, i.e. it should have a low rate of wrong or unsolicited results. These factors aim to avoid overloading the interpreter with unnecessary or incorrect results which could be a potential source of distraction and of error.

Related Work

There is general consensus on the fact that numbers can cause performance losses in simultaneous interpreting and that interpreters require specific strategies and specific pedagogical approaches to cope with them (e.g. Gile, 2009; Setton & Dawrant, 2016; Frittella, 2019). In the past, several studies have been conducted on interpreters' performances in the context of number rendition, involving both professional interpreters and students (e.g. Braun & Clarici, 1996; Lamberger-Felber, 2001; Mazza, 2001; Pinochi, 2009; Timarová, 2012; Desmet et al., 2018; Collard, 2019). All seem to confirm that numbers are poorly rendered, with error rates ranging between 21% and 70%. The range of empirical results is wide and may depend on several factors: different experience level of the interpreter population (the worst performances are measured, as expected, with students, while the best performances with professional interpreters); text types and grade of difficulty; differences in the methodologies used to evaluate errors and evaluation discretionality (approximations, for example, may be counted as errors or not); the fact that studies are based on an experimental setup or on a corpus-based analysis of real-life data (i.e. interpreting conditions are intrinsically different), and so forth.

A notable factor that seems to play a crucial role in interpreters' performance with numbers is the possibility to rely on some sort of visual support during interpretation. For example, when interpreters are allowed to take notes while performing the interpretation, the number of errors decreases by about 10% (Mazza, 2001); when they are given documents in the booth, the error rate decreases by 50% (Lamberger-Felber, 2001); finally, when they can see the numbers synchronised with the original speech displayed on a screen, interpreters make 70% fewer errors than without any form of support (Desmet et al., 2018). These three studies seem to indicate two things. First, the availability of visual numerical input clearly improves interpreters' performance. Secondly, the degree of improvement is directly proportional to the reduction of the cognitive load needed to retrieve the numerical information, which is higher in the case of taking notes and lower in the case of synchronised support on-screen.

The first strain of empirical research in CAI support during the process of interpreting has focused on a relatively similar problem trigger, i.e. specialised terminology. Experimental studies have aimed at measuring how interactive visual support in the booth may improve interpreters' performance. The first set of empirical studies focused on manual terminology lookup (e.g. Prandi, 2015; Biagini, 2016). In these experiments, the probands were asked to interpret a speech and use a CAI tool to perform manual searches in an event glossary. The results of such studies seem to indicate that terminology lookup can help to increase the quality of the rendition, especially in terminology-dense texts. However, the level of the reported improvement seems modest. Reasons for this are the added cognitive load needed for the lookup operation, the perceived distracting influence of this dynamic activity and, on the most general term, the lack of familiarity and absence of training for such an added operation in the already demanding setting of simultaneous interpreting.

The introduction of the first prototype of an ASR tool for interpreters (Fantinuoli, 2017) has moved the interest of experimental research towards the analysis of the interpreter–machine interaction by means of automated lookup tools. Experiments on number rendition using mockup systems (Desmet et al., 2018; Canali, 2019) have indicated a clear reduction of error rates among probands, from 43.5% to 13.5% (gain 30%) in Desmet et al. (2018) and from 64% to 25% (gain 39%) in Canali (2019). In both experimental setups, the number transcription has been displayed either on a screen in the conference room or on a computer in the booth. The numbers were displayed without any information about context, reference, etc., following its delivery by the source speaker. In the first experiment, the speech was given live and the prepared numbers advanced manually, in this way generating a variable but minimal latency between spoken word and transcription. In the second, the text was recorded and the video edited with the transcriptions shown exactly after the term has been completely pronounced by the speaker. Another experiment by Defrancq and Fantinuoli (2020) used the same ASR system as in the current chapter, but implemented a different visualisation approach and highlighted numerals embedded in the complete transcription (see the second section of this chapter). With a complete transcription, a decrease of error rates from 32.3% to 9.8% (gain 22.5%) was reported.

Similar experiments with automated suggestion of other types of problem triggers, for example, terminological units, are underway (e.g. Prandi, 2018). Because of the high number of uncontrolled variables at stake in such experimental settings and the complexity of the simultaneous interpreting process, scholars have been animated to elaborate rigorous theoretical frameworks for the design of empirical studies in the area of ASR-supported CAI tools. Prandi (2018), for example, expands the cognitive load model for the "standard" simultaneous interpreting proposed by Seeber (2011) to accommodate the allocation of cognitive resources during the querying and retrieving of lexical information. Different from the traditional manual lookup, the integration of automatic speech recognition in a CAI tool would require no manual-spatial resources, thus lowering the total interference score (Prandi, 2018). This, again, seems to suggest that ASR may have the potential to integrate better into the interpreting workstation than a traditional CAI tool.

Another strain of research aims at improving CAI tools' suggestions by means of machine learning techniques. One of the main limitations of ASR-enhanced CAI tools, in fact, is that they show suggestions in a non-selective way. In the case of terminology, for example, they show all terminological units contained in the event database that match the transcription. In the case of numbers, no distinction is made between complex and simple ones (100 and 153.867 have different levels of difficulty), nor are different strategies applied based on the number density in a given speech segment, etc. This has several disadvantages. The most prominent one is that the interpreter will be prompted with an abundance of (unfiltered) suggestions, with the possible consequence of being distracted, experiencing a cognitive overload, and, ultimately, decreasing the overall quality of the rendition. Users' feedback stressing this shortcoming is introduced in the analysis of results later in this chapter.

With this in mind, Vogler et al. (2019) have proposed to use machine learning to anticipate the textual units that may cause difficulties for the interpreter and limit the number of suggestions only to these cases. The proposed approach is based on the comparison of a parallel corpus of translated and interpreted speeches in order to automatically identify the text features that led to a terminological issue (for example, omission) in the interpreted rendition. This approach is based on an ML-augmented corpus-based analysis and represents one of the first attempts at process-oriented research in technology-mediated interpreting.

Users' Study on the Use of ASR in SI

The goal of the experimental study was to analyse the impact of using automatic speech recognition on the simultaneous interpretation of numbers. Based on the empirical evidence described in the previous sections, the experiment hypothesised that the visualisation of numbers and their units of measurement by means of ASR would improve the rendition of numbers in simultaneous interpretation. The study aimed therefore to establish the extent to which participants benefitted from the support provided by a real-life CAI tool during the simultaneous interpretation of a text that was dense in numbers. In our experiment, we used the tool InterpretBank ASR with the visualisation of the units of interest without full transcription.

A quantitative analysis of the performance of interpreting students was carried out to verify the hypothesis. The results were then triangulated against the feedback given by the participants at the end of the experiment.

Participants

The experimental study analysed the performance, with or without the help of ASR, of 20 students (all Italian native speakers) at the end of the first and second year of the master's programme in Conference Interpreting at the University of Trieste. In order to take part in the experiment, participants needed to have passed at least the first-year exam in simultaneous interpretation from English into Italian. The participants were equally divided into two groups of ten people: one group interpreted using InterpretBank ASR, while the control group interpreted without support. The participants in the experiment did not undergo any targeted training and had never used CAI tools before.

Data

We set out to use a text dense in numbers. An English speech was selected from the European Commission website in order to keep the text typology as close as possible to the ones used during the interpreting classes. In order to achieve the set goal of number density, more sentences containing figures were added, after some research was conducted on the subject using the European Investment Bank website. The speech was then recorded by an English native speaker. The speech rate reached 123.8 words per minute, which is close to the ideal speed for the interpretation of improvised speeches (Riccardi, 2010).

Experimental Setup

The participants were divided into four batteries. The recorded video was played four times, one for each battery of probands. Before entering the booth, each group received a briefing on the topic of the speech and the group that used ASR was shown how the software worked. The participants did not actively try InterpretBank before the experiment or take part in any practice experiment beforehand. They were only introduced to the speech recognition function by means of a short explanation by the authors of the experiment and by watching a live demonstration with a video fragment taken from a different speech, also dense in numbers. This helped them get to grips with the way in which numbers were transcribed, their colour, and how the order of magnitude was shown. The control group was asked to bring a notepad into the booth as the use of traditional tools was allowed. The performances of the participants were audio-recorded. No information on gaze, fixations, and other features was taken into consideration in this experiment.

At the end of the assignment, the participants were asked to fill in a questionnaire including questions on their note-taking habits as well as on their perception of the speech and of their performance, and their opinion of the ASR functionality. The participants were used to working alone in the booth most of the time and usually without technological support. The interaction with a CAI tool could therefore have led to a perceived increase in cognitive effort for the participants and a feeling of distraction. These perceptions were then triangulated against the measured performance in terms of numbers rendition. Furthermore, they were supposed to give first insights on how the use of CAI tools is perceived by undergraduate users that have not undergone a specific training phase in their use.

Error Annotation

We applied the error categories defined in similar studies on the interpretation of numbers (Braun & Clarici, 1996; Pinochi, 2009). The error typologies considered for the study were the following: *omissions* (the numeral is left out or replaced by a generic expression such as “some,” “a few”); *approximations* (the number is rounded off and sometimes a phrasal element such as “approximately,” “about” is added); *lexical errors* (the order of magnitude is correct, but one or more figures were wrongly interpreted; i.e. 243 instead of 244); *syntactical errors* (the number is of a wrong order of magnitude); *errors of phonetic perception* (the error can be ascribed to a phonetically wrong perception of the number; i.e. 30 instead of 13); *errors of inversion* (the figures are correct, but in the wrong order; i.e. 1.7 instead of 7.1); *other errors* (all errors that do not completely correspond to any of the other typologies). In particular, for each stimulus (the number pronounced by the speaker in the source language) the following information was recorded: numerical class, whether the number was correctly interpreted, the actual number the interpreter pronounced, and, if applicable, the error typology. The total correct/wrong/partially wrong outputs and the distribution of every error typology were computed. For the data gathered from the group using InterpretBank, the following information was recorded: number transcribed by the software, whether the transcription was correct, and temporary versions of the transcriptions. Finally, the total amount of correct/incorrect transcriptions was determined.

It should be noted that, in order not to alter the results of the data analysis, all cases where two errors coexisted were counted as one error when computing the total number of errors. However, in order to produce a more complete picture of the error distribution, double errors were split during the analysis of the error typologies. Consequently, due to the phenomenon of double errors, discrepancies emerged between the sum of incorrectly interpreted numbers and the sum of errors by typology.

ASR Results

The source text included 56 numbers. The overall precision of the transcription was of 81.43%. This score also takes into account the cases where numbers were erroneously transcribed without any digit present in the original speech (for example, in the case of speaker hesitations recognised as numbers or pronouns transcribed as numbers).

Considering only the transcription quality of the numbers in the original speech, the precision reaches 82.6% which is well below evaluations conducted in similar experiments (e.g. Brüsewitz, 2019; Defrancq & Fantinuol, 2020). This result was mainly caused by a particular type of error that occurred repeatedly during the course of the text, namely lists of numbers. With adjacent numbers, such as “in 2000: 6.7%,” the absence of a clear stop signal between the digits made the tool first transcribe “2000” correctly, then correct itself by changing the “2000” into “2006.” This was even more true for long sequences of numbers which were present in the speech. Without considering this particular case of number proximity, the ASR performed similarly to the results of the above-mentioned experiments, with precision values around 94%.

The other errors made by the ASR can be ascribed to phonetic misinterpretations, such as “2” instead of “to,” “6” instead of “success,” and by number proximity. In some cases, the real-time correction ability of the ASR system caused the transcription to be changed several times until the final result stabilised. This is due to frequent phonetic perception errors found in the early stages of transcription, for example, 30 being transcribed as 30, then as 13, and finally as 30 again. For the purposes of this study, all cases in which the last transcription on the screen was incorrect or not displayed were categorised as errors, even when a previous correct transcription had been displayed. On the other hand, even when the final transcription was correct, the time constraints of simultaneous interpretation, however, may prevent the interpreter from making proper use of the suggestion provided or, conversely, create confusion and worsen his or her performance.

In other cases, the live correction of the transcribed number created confusion by quickly and repeatedly changing the displayed digit. In the case of "for the next financial period 2021–2027," for example, "2027" was interpreted as a correction of the previous number "2021," which was immediately substituted by "2027." In some isolated cases, the numbers were correctly transcribed in a temporary version (shown for a relatively long interval), but the final version transcribed on the screen was incorrect. In these cases, the participants often had the chance to read the correct temporary transcription, without letting the following incorrect transcription affect their performance.

An analysis of how wrong transcriptions were interpreted by the students highlighted that, on average, 0.8 errors committed by the student were to be ascribed to an incorrect/inexistent transcription by the software. The fact that participants correctly interpreted numbers even when the transcription displayed was wrong or lacking shows that the students were able to use InterpretBank as a supportive tool, rather than as a replacement of their skills.

Product-Based Results

The objective of this study is to assess whether the use of automatic speech recognition for the rendition of numbers can improve the performances of students who usually interpret without any aid. Overall, the data analysis revealed an error rate of 14.8% for the group using technological support and 39.8% for the control group. Therefore, the participants who used an ASR-enhanced CAI tool correctly interpreted 41.5% more numbers than the other group.

As far as a more fine-grained analysis is concerned, the control group registered a wider distribution of errors over several different typologies and in a more balanced way. In the following, a brief analysis of the single typologies of errors is presented. Overall, the support of the software led on average to 25% more correctly interpreted numbers in every numerical class.

Omissions

In both groups, the majority of errors were classified as omissions. On average, 4.5 numbers were omitted with ASR and 12.9 without technological support. As expected, the control group recorded a higher percentage of full omissions. It has to be noted that omissions, as a typology of errors, can be seen both as an error and a strategy, since interpreters often omit numbers strategically to avoid cognitive overload, preventing in this way a worsening of their overall rendition at the expense of a limited loss in numerical precision. In this context, both groups limited omissions to cases in which they did not entail any actual loss of meaning. Whenever possible they applied strategies to mitigate the omission, for example, by using determiners.

Comparing the two groups, the participants relying on the help of ASR omitted numbers by replacing them with a linguistic expression ("many," "few") on average only 0.3 times, while the control group resorted to this strategy more readily (on average 2.6 times). This seems to highlight the benefit of technological support not only in reducing the number of errors that lead to a complete loss in the meaning conveyed but also in improving the precision of the rendition.

Approximations

The students who used technological support resorted to approximations to a lesser extent (from an average of 2.1 numbers approximated without technological support to 0.3 with ASR), since they most frequently succeeded in retrieving the number transcribed on the screen and correctly interpreting all the digits. On the other hand, the students of the control group frequently rounded off a number in order to avoid a rendition which was either incorrect or too imprecise. To be specific, only two out of the ten students using ASR resorted to this strategy, compared to eight out of ten students of the control group. Hence, it can be inferred that the control group encountered more difficulties in correctly rendering all the digits that constitute a number. This result confirms the effectiveness of the software in providing support for interpreting numbers, including complex or very long ones.

Lexical and Syntactic Errors

Lexical errors – the order of magnitude is correct, but one or more digits are incorrect, i.e. 1998 instead of 1989 – were registered in both groups (on average 0.8 with ASR and 1.6 without), especially when it came to interpreting decimal numbers (for example, 6.7 interpreted as 6.2). Syntactic errors – the order of magnitude is incorrect, but the digits are correct – were found in both groups (on average 1.2 with ASR and 2.7 without). The greatest part of errors occurred while translating "billion" or "million" into Italian. Taking into consideration that the software always correctly transcribed the orders of magnitude and showed their full original form (i.e. "billion" and not "bl") in the source language, these data show that students still have doubts – at least when they are under pressure – about the translation of these words into Italian and often confuse one for the other (i.e. "billion" as "milione"). A solution to this challenge could lie in a translated transcription of the order of magnitude (_billion>_miliardo).

Phonetic Perception

The participants who used the support committed considerably fewer phonetic perception errors (from an average of 2.2 errors for the control group to 0.4 with ASR). Students in the control group not only made recurrent phonetic perception mistakes, such as 30 > 13, but also often wrongly interpreted decimals (i.e. 8.7% > 7%). On the other hand, no participant using ASR made this kind of mistake with decimals because they had the possibility to check the digits on the screen. The CAI tool therefore proved useful and effective for non-complex numbers too.

Errors of Inversion

In both groups, no error corresponding to such typology was identified. Since the language of the speech (English) has a linear numerical system, a low percentage of such errors was expected, even though they were expected to occur at least with decimals. However, regarding the category of decimals, the analysis often detected approximations, phonetic perception errors, or lexical errors, but never errors of inversion.

Other Typology

This additional typology was included to count all those errors that did not completely correspond to any of the other typologies, and all errors to be ascribed to incorrect transcriptions by the software. Some errors categorised as "other typology" turned out to be linked to the expectations of the interpreter, such as the example of "it will have multiplied 15 times over," where "15 times" was often interpreted (in Italian) as "by 15%," even though the software did not transcribe "%," but only "15." Only 7% of the total errors committed by the control group were categorised as "other typology," whereas 15% is the case for the group using the technological support. The higher number of errors for the group using ASR is also due to the fact that this typology included all errors caused by relying on incorrect transcriptions.

Perception-Based Results

At the end of the experiment in the booth, participants were asked to fill in a questionnaire regarding their habits of jotting down numbers on paper during the simultaneous interpretation of numbers, how they perceived the speech and their performance during the experiment, and how they perceived interaction with the automatic suggestions. The questionnaire included both open and closed questions to be answered by choosing the level of agreement to a statement (Likert scale). From these data (averaged for each group), the degree of difficulty perceived during the experiment can be deducted. In the following paragraph, some of the most telling results will be discussed.

Participants affirmed that they usually perceive the simultaneous rendition of numbers as slightly difficult (average levels of agreement of 3.4/5 and 3.8/5 on the Likert scale for the two groups). In both groups, only half of the participants claimed to always jot a number down before interpreting it. Participants who do not have the habit of taking notes of numbers stated that they usually display them mentally and focus only on the number (often to the detriment of the general output in the target language).

Regarding the experiment in question, the number of numbers in the text was considered high (4.2/5) by the group using technological support and extremely high (5/5) by the control group. Hence, all the participants of the control group experienced many difficulties when interpreting this text dense in numbers. The control group considered the high frequency of numbers to be a source of high cognitive load (4.8/5), whereas the group using ASR only agreed to a medium extent (3.5/5). The distance between the numbers in the text was considered acceptable by the group using technological support, but too short by the control group. The statement "*the numerical density affected my output in Italian*" obtained a 3.5/5 level of agreement by the group using ASR and 4.6/5 by the control group. These answers are aligned to the performances measured in the empirical experiment.

Another interesting result is the perceived difficulty in understanding the exact number uttered, which is given 2.8/5 on the Likert scale by the group using technological support but reaches the value of 4.3/5 with the control group. These figures highlight the significantly higher degree of difficulty perceived by participants who did not use the software. The students in the group using ASR also claimed not to have encountered any particular difficulties in identifying the order of magnitude (2.1/5) or what the number referred to (2.9/5), while students of the control group encountered more difficulties (3.4/5 in both cases).

All participants were generally quite dissatisfied with their own performance (2.4/5 for those who used the software and 2.2/5 for the control group), even though the first group, in fact, interpreted more numbers correctly. However, it should be noted that an interpreting performance includes not only the correct rendition of numbers but also their context, the discursive parts, and a good output in the target language. Therefore, the results of the analysis of the rendition of numbers cannot be linearly compared to the satisfaction of the performance during the experiment, since the data analysed only examine one of the several aspects of satisfactory interpreting performance.

Participants also pointed out that when the software made mistakes in the transcription of a number, they often found it difficult to correct their rendition in a timely manner. Participants using ASR stated in the questionnaire that they believed they interpreted discursive parts well and demonstrated a strong ability to close sentences. However, when asked to provide examples of weaknesses in their performance, they listed poor accuracy, references not understood, the output in the target language, and the rendition of segments dense in numbers. The control group considered their strengths to be the application of strategies (approximation or omission) in the case of numbers not fully understood, the output in Italian, and the general understanding of the text. Under the weak points, almost all participants of the control group included difficulties related to quantity and density of figures in the text, especially as regards the rendering of references and of the order of magnitude.

The probands who used ASR during the experiment emphasised that they could have made better use of the tool if they had had the opportunity to practice beforehand. The statement "*the speed at which numbers appeared on the screen was appropriate for the purposes of interpreting*" was given a degree of agreement of 3.6/5. This value may suggest that participants encountered difficulties in interpreting a text dense in numbers, even when relying on the software. Future measurements with different values of latency should shed light on the ideal threshold of latency in order to increase the usability of the suggestions.

Regarding the formal aspects of ASR, the transcription of the number by the software was rated as averagely accurate (3.2/5), and the way numbers were shown was considered on average clear and understandable (3.1/5). This value indicates a misalignment between the perception of ASR accuracy and the measured precision of transcription. The probands suggested small improvements such as different demarcation signs between units and decimals and between thousands and hundreds (i.e. for Italian, comma and full stop). The answers to the question "*the use of the software was not a source of distraction*" are a useful indicator of the perceived difficulties encountered by the students in integrating this kind of tool while simultaneously interpreting: 50% agreed with a level of 4/5, 30% with 3/5, and only 20% with 2/5. In general, the probands that used the software described it as helpful, especially in segments rich in numbers (including complex numbers), but less so with dates or numbers with up to three zeros.

The added task of reading the transcribed numbers did not raise any particular difficulty, as the numbers were marked with different colours and the last number would always appear at the top. Nonetheless, reading the transcription on the screen proved to be a potential source of distraction, which is why some probands stressed the importance of getting to grips with the software before using it in a real-life situation. Of the probands who used technological support during the experiment, 40% stated they would use it in a work context, while 50% stressed that the choice would depend primarily on the type of text in question (namely, on its number density).

Finally, the integration of CAI tools into translation and interpretation technology courses was strongly recommended by both groups, since more practice with CAI tools would considerably decrease the distraction factor linked to such tools.

Conclusions

The experiment confirms that automatic speech recognition proved effective in providing support during the interpretation of a speech dense in numbers. The results show a significant reduction of error rate, which drops from 39.8% without technological support to 14.8% with the support of automatic speech recognition (gain 25%). In particular, the support of the ASR led to a reduction in omissions (from an average of 12.9 without ASR to 4.5 with ASR) and helped to improve performance for complex or very long numbers, as shown by the decrease in cases of approximation (from an average of 2.1 errors without ASR to 0.3 with ASR). Furthermore, ASR helped participants avoid phonetic perception errors (from an average of 2.2 errors without ASR to 0.4 with ASR). The experimental group also registered a smaller percentage of double errors (for example, syntactic error and approximation at the same time), especially with long numbers (from an average of 0.6 double errors without ASR to 0.1 with ASR). This phenomenon could be partly explained by the larger quantity of elements to be processed in complex numbers: without technological support, this numerical class could entail more significant memorisation and elaboration difficulties.

Most of the software errors occurred with simple numerals. Being easier to process, they had no serious repercussions in the output, since the probands either did not rely on the ASR for that speech segment or they were able to correct the wrong suggestion.

Quite interestingly, the results of our experiment using state-of-the-art automatic speech recognition are very similar to the results of Desmet et al. (2018) using a simulated transcription and manual synchronisation between suggestions and original speech. In their experiment, the average error rate without technological support was 43.5% (compared to 39.8% in ours) while the average error rate with technological support was 13.5% (compared to 14.8% in ours). This seems to suggest that the quality of today's technological development, at least in a comparable setting and in the context of numbers, is already mature enough to be used in real-life applications.

The analysis of the questionnaire stresses some of the difficulties encountered by the probands, such as the feeling of distraction caused by the added visual stimulus and the need to coordinate it with the other co-occurring subprocesses of simultaneous interpretation. Such drawbacks could be mitigated by specific training in the use of the support. The fact that the probands were already able to do a selective use of the technological support, as demonstrated, for example, by the fact that they correctly interpreted numbers even when they were wrongly transcribed (hence indicating a use of the software as a supportive tool, rather than as a replacement for the listening and comprehension skill), leads us to think that training could develop a strategic approach to the use of CAI tools and, as a consequence, reduce the effect of distraction produced by them. Furthermore, the user feedback indicated some shortcomings in the way suggestions were presented. In this context, further research in data visualisation should be conducted and new methods tested.

The positive outcome resulting from the use of this software on untrained interpreters and the feedback collected stress the need to explore in greater depth the technological tools available in the area of interpreting and integrate them into the curricula.

One limitation of our experiment was that the test population did include only undergraduate interpreters. In order to generalise the validity of the results, a similar study should be replicated with professional interpreters. It is our hypothesis that professionals, especially after being trained on how to use the tool, will improve their performance on numbers too, but that the magnitude of improvement will be smaller because of the lower error rates that characterise professional renditions without technological support.

References

Braun, S., & Clarici, A. (1996). Inaccuracy for numerals in simultaneous interpretation: Neurolinguistic and neuropsychological perspectives. *The Interpreters' Newsletter*, 7:85–102.

Brüsewitz, N. (2019). Simultandolmetschen 4.0: Ist automatische Spracherkennung der nächste Schritt? In: *Proceedings of the Conference on Übersetzen Und Dolmetschen 4.0. – Neue Wege im digitalen Zeitalter*. BDÜ Fachverlag.

Canali, S. (2019). *Technologie und Zahlen beim Simultandolmetschen: Utilizzo del riconoscimento vocale come supporto durante l'interpretazione simultanea dei numeri*. Università degli studi internazionali di Roma, unpublished MA thesis.

Collard, C. (2019). *A Corpus-Based Study of Simultaneous Interpreting with Special Reference to sex*. PhD diss. Ghent University.

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*, 32(2), 73–102.

Desmet, B., Vandierendonck, M., & Defrancq, B. (2018). Simultaneous interpretation of numbers and the impact of technological support. In: Fantinuoli C. (Ed.) *Interpreting and Technology*. Berlin: Language Science Press, 13–27.

Fantinuoli, C. (2012). *InterpretBank—Design and Implementation of a Terminology and Knowledge Management Software for Conference Interpreters*. PhD diss. University of Mainz.

Fantinuoli, C. (2017). Speech recognition in the interpreter workstation. In: *Proceedings of the Translating and the Computer 39*. London.

Fantinuoli, C. (2018). Computer-assisted interpreting: Challenges and future perspectives. In: Durán Muñoz I. & Corpas Pastor G. (Eds.) *Trends in e-Tools and Resources for Translators and Interpreters*. Leiden: Brill, 153–174.

Frittella, F. (2019). "70.6 billion world citizens": Investigating the difficulty of interpreting numbers. *Translation & Interpreting*, 11(1):79–99.

Gile, D. (2009). *Basic Concepts and Models for Interpreter and Translator Training* (2nd ed.). Amsterdam: John Benjamins.

Lamberger-Felber, H. (2001). Text-oriented research into interpreting: Examples from a case-study. *Hermes - Journal of Language & Communication in Business*, 26(26):39–63.

Liu, M., Schallert, D. L., & Carroll, P. J. (2004). Working memory and expertise in simultaneous interpreting. *Interpreting. International Journal of Research & Practice in Interpreting*, 6(1):19–42.

Mazza, C. (2001). Numbers in simultaneous interpretation. *The Interpreters' Newsletter*, 11:87–104.

Meuleman, C., & VanBesien, F. (2009). Coping with extreme speech conditions in simultaneous interpreting. *Interpreting. International Journal of Research & Practice in Interpreting*, 11(1):20–34.

Moser-Mercer, B. (2000). Simultaneous interpreting: Cognitive potential and limitations. *Interpreting. International Journal of Research & Practice in Interpreting*, 5(2):83–94.

Pinochi, D. (2009). Simultaneous interpretation of numbers: Comparing German and English to Italian. An experimental study. *The Interpreters' Newsletter*, 14:33–57.

Pöchhacker, F. (2016). *Introducing Interpreting Studies* (2nd ed.). London: Routledge.

Prandi, B. (2015). The use of CAI tools in interpreters' training: A pilot study. In: *Proceedings of the 37 Conference Translating and the Computer*, 48–57.

Prandi, B. (2018). An exploratory study on CAI tools in simultaneous interpreting: Theoretical framework and stimulus validation. In: Fantinuoli C. (Ed.) *Interpreting and Technology*. Berlin: Language Science Press, 25–54.

Riccardi, A. (2010). Velocità d'eloquio e interpretazione simultanea. In: *Am Schnittpunkt von Philologie und Translationswissenschaft : Festschrift zu Ehren von Martin Forstner*, 281–299.

Seeber, K. G. (2011). Cognitive load in simultaneous interpreting: Existing theories – new models. *Interpreting*, 13, 176–204.

Setton, R., & Dawrant, A. (2016). *Conference Interpreting: A Complete Course*. Amsterdam: John benjamins.

Stoll, C. (2009). *Jenseits simultanfähiger Terminologiesysteme: Methoden der Vorverlagerung und Fixierung von Kognition im Arbeitsablauf professioneller Konferenzdolmetscher*. Trier: WVT Wissenschaftlicher Verlag Trier.

Timarová, S. (2012). *Working Memory in Simultaneous Interpreting*. PhD diss. KU Leuven.

Vogler, N., Stewart, C., & Neubig, G. (2019). *Lost in Interpretation: Predicting Untranslated Terminology in Simultaneous Interpretation*. arXiv:1904.00930v1.

Will, M. (2007). Terminology work for simultaneous interpreters in LSP conferences: Model and method. In: *Proceedings of the EU-High-Level Scientific Conference Series MuTra*, 65–99.

Xu, R. (2018). Corpus-based terminological preparation for simultaneous interpreting. *Interpreting. International Journal of Research & Practice in Interpreting*, 20(1):29–58.