# Usability research for interpreter-centred technology

## The case study of SmarTerp

Francesca Maria Frittella

Translation and Multilingual Natural Language Processing 21

language science press

## Translation and Multilingual Natural Language Processing

Editors: Oliver Czulo (Universität Leipzig), Silvia Hansen-Schirra (Johannes Gutenberg-Universität Mainz), Reinhard Rapp (Hochschule Magdeburg-Stendal), Mario Bisiada (Universitat Pompeu Fabra)

In this series (see the complete series history at https://langsci-press.org/catalog/series/tmnlp):

11. Fantinuoli, Claudio (ed.). Interpreting and technology.
12. Nitzke, Jean. Problem solving activities in post-editing and translation from scratch: A multi-method study.
13. Vandevoorde, Lore. Semantic differences in translation.
14. Bisiada, Mario (ed.). Empirical studies in translation and discourse.
15. Tra&Co Group (ed.). Translation, interpreting, cognition: The way out of the box.
16. Nitzke, Jean & Silvia Hansen-Schirra. A short guide to post-editing.
17. Hoberg, Felix. Informationsintegration in mehrsprachigen Textchats: Der Skype Translator im Sprachenpaar Katalanisch-Deutsch.
18. Kenny, Dorothy (ed.). Machine translation for everyone: Empowering users in the age of artificial intelligence.
19. Kajzer-Wietrzny, Marta, Adriano Ferraresi, Ilmari Ivaska & Silvia Bernardini. Mediated discourse at the European Parliament: Empirical investigations.
20. Marzouk, Shaimaa. Sprachkontrolle im Spiegel der Maschinellen Übersetzung: Untersuchung zur Wechselwirkung ausgewählter Regeln der Kontrollierten Sprache mit verschiedenen Ansätzen der Maschinellen Übersetzung.
21. Frittella, Francesca Maria. Usability research for interpreter-centred technology: The case study of SmarTerp.
22. Prandi, Bianca. Computer-assisted simultaneous interpreting: A cognitive-experimental study on terminology.

ISSN: 2364-8899

# Usability research for interpreter-centred technology

## The case study of SmarTerp

Francesca Maria Frittella

language
science
press

Francesca Maria Frittella. 2023. *Usability research for interpreter-centred technology: The case study of SmarTerp* (Translation and Multilingual Natural Language Processing 21). Berlin: Language Science Press.

This title can be downloaded at:
http://langsci-press.org/catalog/book/366

ISBN: 978-3-96110-403-1 (Digital)
978-3-98554-061-7 (Hardcover)

ISSN: 2364-8899
DOI: 10.5281/zenodo.7376351
Source code available from www.github.com/langsci/366
Errata: paperhive.org/documents/remote?type=langsci&id=366

Cover and concept of design: Ulrike Harbort
Typesetting: Mario De Florio, Francesca Maria Frittella, Sebastian Nordhoff, Felix Kopecky, Yanru Lu
Proofreading: Amir Ghorbanpour, Jean Nitzke, Elliott Pearl, Jeroen van de Weijer
Fonts: Libertinus, Arimo, DejaVu Sans Mono
Typesetting software: XeLaTeX

Language Science Press
xHain
Grünberger Str. 16
10243 Berlin, Germany
http://langsci-press.org

Storage and cataloguing done by FU Berlin

Freie Universität Berlin

# Contents

# Acknowledgements

This book was written without any financial support; nonetheless, I would like to express my gratitude to SmarTerp, especially Susana Rodriguez, for allowing me to conduct my research on their software. I am indebted to all the friends, colleagues, reviewers, and editors who contributed to this work. A special thanks goes to my cousin Dr Mario De Florio, who took care of the LaTeX typesetting; without you, I wouldn't have made it (surely not in this timeframe). Publishing a book does take a village, and I am grateful to you all for making this happen.

# 1 Introduction

While all professions are affected by technological innovation, the impact has been particularly profound on the language professions of translation and interpreting (T&I). New information and communication technologies (ICTs) have shaped the "*micro*- and *macro*-systems" (O'Hagan 2013) of T&I, which is to say, both the way the core tasks of these professions are performed and the social and physical contexts in which translators and interpreters operate. While new technologies threaten to replace human interpreters and translators, on the one hand, they also afford new T&I services and support the work of translators and interpreters, with the ambition to increase professionals' productivity and service quality, on the other hand.

The impact of ICTs on written translation became evident in the noughties, leading scholars to speak of a "technological turn" (Chan 2007, Cronin 2010, O'Hagan 2013), although the work on ICTs with revolutionary power (such as machine translation) began in the 1950s. In interpreting, it is not until the last decade that the "technological turn" (Fantinuoli 2018b) was announced by scholars, even though previous technological breakthroughs had already shaped the developments of the profession, for instance leading to the birth of simultaneous interpreting (SI). Compared to previous phases of change following technological developments, today, interpreting technologies are increasing in number and penetrating the profession at a pace that became exponential a few years ago (Fantinuoli 2018b) and was further sped up by the Covid-19 pandemic. It is likely that ICTs will soon become so deeply entrenched in the interpreter's workflow that we will start viewing all of interpreting as a form of human-computer interaction (HCI), as scholars already argued of translation (O'Brien 2012).

Among the technologies that have the potential to shape the ways in which interpreting is performed and alter the very nature of the underlying cognitive sub-processes, computer-assisted interpreting (CAI) tools and their use during SI are of particular significance to the profession. Like computer-assisted translation (CAT) tools for translators, CAI tools are designed with the intent to ease and enhance the work of interpreters, increasing interpreters' efficiency (in the preparation phase) and delivery accuracy (in the interpretation or in-booth phase).

In the *in-booth* phase, the latest generation of CAI tools is aimed at supporting interpreters in the rendition of particularly demanding and error-prone linguistic items. These are named entities, numbers, specialised terms, and acronyms. Previous research on these dreaded "problem triggers" (Gile 2009) has shown both the staggering error rates in interpreters' delivery when these items occur in the speech to be translated (cf. Desmet et al. 2018 and Frittella 2017, 2019a on numbers, for instance) and reported interpreters' feelings of mental effort and stress associated with interpreting these items (e.g. Alessandrini 1990). The increased difficulty forces interpreters to adopt coping tactics during SI in the presence of problem triggers, such as manually searching for named entities, terms and acronyms in the interpreter's glossary or external sources such as electronic dictionaries or databanks, writing down numerals, or asking the colleague working in the booth with them to help them perform these tasks accurately. Because all these processes entail considerable risk for human error and must be attended to by the interpreter, while s/he is performing a task that is in itself complex and cognitively demanding, they risk leading to errors and a disruption in the interpreting activity. The recent integration of *automatic speech recognition* (ASR) and *artificial intelligence* (AI) technology into CAI tools offers the opportunity to help interpreters better cope with problem triggers. By automatically displaying named entities, numbers, specialised terms and acronyms on the interpreter's laptop screen in real-time, ASR- and AI-powered in-booth CAI tools have the potential to provide interpreters with a reliable "virtual boothmate", relieving them from the stress and mental effort in dealing with problem triggers and increasing delivery accuracy.

However, the complexity of the mental task that CAI tools aim to support constrains the potential effectiveness of the artificial boothmate. The synchronized execution of multiple mental operations, which partly compete for the same limited cognitive resources, makes SI a sensational stunt of mental acrobatics. During SI, human interpreters receive a "source speech" in one language, mentally process it and turn it into a "target speech" in a different language. All this happens in real-time, under time pressure and at a pace that cannot be controlled by the interpreter. It is possible that the introduction of yet another source of information into the already cognitively taxing activity of SI may cause interpreters to reach a mental saturation point and disrupt the task. While details matter in all user interfaces and even seemingly minor particulars may disrupt the user, this may be even truer of CAI tool interfaces. Maximum *usability* seems to be an imperative for CAI tools to support rather than disrupt the delicate balance of SI. This assumption is shared amongst leading scholars in this area, who postulated that a CAI tool should "offer ergonomic ways to present extracted

information" (Fantinuoli 2017: 26), if the tool is not to exert the opposite effect and "slow down delivery and place a burden on interpreters' mental processing" (Defrancq & Fantinuoli 2021: 2). However, so far, the set-up of technical requirements for CAI tools and their UI design has been mainly driven by developers' (usually interpreters themselves) intuition, without any "experimental support in design decisions" (Fantinuoli 2018a: 160). The overall process has lacked "test and improvement phases, which are crucial to achieve software maturity and find the golden standard in terms of functionalities desired by the professional group" (Fantinuoli 2018b: 164).

The importance of research as an instrument to increase the usability of CAI tools becomes apparent at a comparison with CAT tools and their development. Also in the case of CAT tools, scholars extensively argued the importance of usability: "The link should be quite obvious: While CAT tools exhibiting a high usability should enhance the translator's cognitive performance, tools exhibiting a low usability will probably tend to decrease it" (Krüger 2016a: 115) and "If CAT tools are easy to use, then more time and cognitive capacity should be available for the decision-making and problem-solving processes integral to translation work. If such tools or certain features are complicated and/or non-intuitive, then human-computer interaction can be compromised, which usually results in less than optimal use and dissatisfaction with tools" (Kappus & Ehrensberger-Dow 2020: 2). However, the development of the first CAT tools, which later became and today still represent the market standard, was market-driven rather than translator-driven, i.e., emerging from the initial disillusionment with "unmanned" machine translation and motivated by the need for larger translation volumes at reduced time and costs. Therefore, translators' needs were not adequately accounted for in the development of the first CAT tools (cf. Moorkens & O'Brien 2017, O'Brien et al. 2017 inter alia). While virtuous examples of systematic, translator-centred CAT tool development do exist, already existing solutions are difficult to change for economic reasons, even where their UI features are found to be "irritating" (O'Brien et al. 2017) and inefficient for translators, hence decreasing their productivity. Several studies identified poor usability as a major reason for translators' resistance to using CAT tools (e.g. LeBlanc 2013, O'Brien et al. 2017). Nonetheless, because it is often the agency or the client who decides which CAT tool is to be used, translators may feel obliged to use a solution that they find too expensive and unusable just to access jobs (Garcia 2015). Forced to adopt solutions that do not fully consider their needs may contribute to translators' feeling of "dehumanisation" and "devaluation" (O'Brien 2012) following the introduction of CAT tools into their workflow.

Given that the economic case for developing CAI tools was much less compelling than CAT, the initial development of these solutions has emerged "from the bottom", i.e. from the initiative of interpreters themselves, rather than being imposed from above. However, this situation appears to be changing. Large remote-simultaneous interpreting (RSI) providers[1] who own the largest share of the RSI market are beginning to integrate CAI tools into their platforms. Large institutional clients such as international organisations[2] are increasingly integrating CAI tools into their work processes. The fact that CAI tool design today still "reflects more the ideas and habits of the respective developer, generally an interpreter himself, than the needs of the interpreter community" (Fantinuoli 2018a: 164) and that a strand of usability-focused CAI research is still missing threaten to lead to unsustainable technological developments. What has been missing from the development of CAI tools is the involvement of the community, the validation of design assumptions and the incorporation of data into the development process to improve the solutions. A recurring slogan under which CAI tools are marketed is "developed *for* interpreters *by* interpreters". What is missing from the equation is "*with* interpreters" – the engagement of the community in conceptualising and creating the tools intended to be used by professionals in their everyday work. The distance of the interpreting community from the development of CAI tools may be a factor contributing to the traditional disengagement and mistrust towards these tools. Time and again, practitioners express concerns about the possible negative impact that CAI tools may have on their interpreting process and final output, the changes that these may cause on their professional status and remuneration, and even that the tools may take over, confusing CAI with the application of ASR- and AI in machine interpreting to replace, rather than support, human service provides. Like translators perceived CAT tools as a "black box" (O'Brien 2012) also interpreters seem to fear a possible dehumanisation and devaluation following the introduction of CAI tools into their work. While some have interpreted professionals' aversion to technological change as "defense of old accrued power, dressed in the guise of quality" (Pym 2011: 4 cited in Fantinuoli 2018b: 155), it may also be interpreted as a consequence of the lack of community engagement in the development of CAI tools.

This book presents a case study of interpreter-centred development of an ASR- and AI-powered CAI tool–*SmarTerp*, developed within the European Union funding line EIT Digital. It represents the first case in which a usability engineering (or "user-centred design") approach was used to develop a CAI tool, starting from

[1] See the introduction of Interpreter Assist (Fantinuoli et al. 2022a) into Kudo, currently the world's number one RSI provider.

[2] For instance, the CAI tool InterpretBank (Fantinuoli 2016) is currently being used for interpreters' terminology management at institutions such as the OECD (https://www.interpretbank.com/site/).

the analysis of users' needs to develop requirements, define UI features, and test and refine the solution iteratively to integrate users' feedback and data into the design of the tool at an early stage of its development. The process exemplifies the role of research as an instrument for the inclusion of the interpreting community into the development of technological solutions.

The focus of the book will be on the empirical evaluation of the CAI tool through the usability testing method. Usability testing consists of the observation of representative users performing representative tasks with the product or a high-fidelity functioning prototype. In HCI and usability engineering processes, it is regarded as a fundamental method in the development of interactive systems because it makes it possible to identify the challenges that users encounter during usage and change the UI features that cause those impasses in the interaction. Usability testing, especially at this initial stage of in-booth CAI tool development, may represent a crucial method in the empirical evaluation of these systems, to identify those seemingly minor details that may make a major difference during SI and gain a better understanding of interpreters' needs.

By making the knowledge and materials developed through this research experience available to a wider audience, I hope that this book may pave the way for the development of a strand of usability-focussed empirical CAI research, contribute to innovation in the field and promote high standards of methodological and scientific rigour. At a time when "interpreting is about to go through a transformation phase driven by socio-technical change" (Fantinuoli 2018b: 8), it is a priority to ensure that technology advances *with* interpreters.

The book is structured as follows. Chapter 2 on usability engineering provides the conceptual and methodological framework of the inquiry. Chapters 3 and 4 respectively review the history of translation and interpreting technology (particularly, CAT and CAI) and the methods that were used to evaluate the usability of these tools. Chapter 5 presents the background of the SmarTerp project. Chapter 6 opens the empirical section of the work with a detailed presentation of the study design and materials. Chapters 7 and 8 present the results of the two usability tests – the pilot study (involving 5 conference interpreters as study participants) and the main study (10 conference interpreters) – leading to design recommendations and the improvement of the CAI tool prototype. The following discussion (Chapter 9) summarises the broader implications of the study findings in terms of gains in the field's scientific understanding of interpreter-CAI interaction and usability principles. The discussion also addresses the limitations of the present work, points to possible future research trajectories and provides methodological recommendations for future researchers wishing to build on this study. The most salient aspects of the work and the future outlooks are summarised in the conclusion (Chapter 10).

# 2 Usability engineering

This chapter is an introduction to usability engineering – a set of practices used in the development of interactive systems to ensure their adequacy to users' needs and their usability. It provides the rest of the work with conceptual and methodological instruments that will then be applied to the analysis of previous research on translation technology (Chapter 3) and interpreting technology (Chapter 4), as well as to design of the study on the CAI tool SmarTerp (Chapter 6) and the analysis and interpretation of findings (Chapters 7 to 9). The present chapter first defines the concept of "usability" and how this abstract concept may be specified into measurable components. It then characterises usability engineering as the different research activities that are conducted in the process of designing and developing interactive systems by this approach. It then reviews different evaluation methods that are used at different development stages to validate and refine concepts and prototypes. Finally, the chapter presents the key methodological tenets of usability testing, a crucial method in usability engineering.

## 2.1 The concept of usability

Usability is a crucial concept in the study of human-computer interaction (HCI) and the development of the concept of usability and that of the discipline have gone hand-in-hand. HCI is fundamentally "the study of people, how they think and learn, how they communicate, and how physical objects are designed to meet their needs" (Lazar et al. 2017: 7). It is a discipline concerned with the theoretical aspects of product design, with the procedural aspects of how to achieve a good design and with the practical aspects of informing specific products. In this respect, Lazar et al. (2017: 10) distinguish between *technical HCI research* (focused on interface building) versus *behavioural HCI research* (focused on cognitive foundations). Irrespective of the specific focus of different strands, HCI is concerned with the *practical relevance* of research outputs: "HCI research must be practical and relevant to people, organizations, or design. The research needs to be able to influence interface design, development processes, user training, public policy, or something else" (Lazar et al. 2017: 7).

The birth of HCI as a discipline is believed to coincide with the first conference on Human Factors in Computing Systems in Gaithersburg (Maryland, United States) in 1982 (Lazar et al. 2017: 1). In the previous decade leading up to the conference, computers had been slowly moving from research laboratories into the home and the office. With this move, the use of computers was no longer the exclusive realm of technicians and, consequently, interfaces had to be designed to be used by laypeople: "The interaction between the human and the computer was suddenly important. Nonengineers would be using computers and, if there wasn't a consideration of ease of use, even at a basic level, then these computers were doomed to failure and nonuse" (Lazar et al. 2017: 2).

At that time, "computer vendors first started viewing users as more than an inconvenience" (Nielsen 2010: 4). They realised that excessive complexity put people off and represented a major barrier to the reception of a product. Nielsen convincingly explains the problem in the context of web design: "If a website is difficult to use, people *leave*. If the homepage fails to clearly state what a company offers and what users can do on the site, people *leave*. If users get lost on a website, they *leave*. If a website's information is hard to read or doesn't answer users' key questions, they *leave*" (Nielsen 2012, emphasis in the original). This realisation led to the emergence of the concept of "user-friendliness", later "usability" (Bevan et al. 1991, Nielsen 2010), and to intensified efforts to make products "usable". The launch of the Apple Macintosh made a strong case for the importance of usability:

> Introduced to the public during the 1984 Super Bowl, the Macintosh seemed to make the case that ease of use sells. It was the symbol of a well-designed product. While usability was only one of the factors that made the Macintosh a household word, it was the most often mentioned factor in the growing, but small, market share that the Macintosh captured. In my view, the Mac changed the level of the argument for investing in usability. Before this period, I had to argue with clients that usability itself was important. Afterward, the high-tech world simply assumed that usability was essential. (Dumas 2007: 56)

The term usability seems to have first emerged in the 1980s as a substitute for the term "user friendly" which engineers realised had "a host of undesirably vague and subjective connotations" (Bevan et al. 1991). It was deemed inadequate for mainly two reasons:

> First, it is unnecessarily anthropomorphic – users don't need machines to be friendly to them, they just need machines that will not stand in their way when they try to get their work done. Second, it implies that users' needs

> can be described along a single dimension by systems that are more or less friendly. In reality, different users have different needs, and a system that is "friendly" to one may feel very tedious to another. (Nielsen 2010: 4)

Bevan et al. (1991) distinguish different views of what usability is and how it is measured. Based on the product-oriented view, usability can be measured in terms of the ergonomic attributes of the product. This view implied that "usability could be designed into the product, and evaluated by assessing consistency with these design guidelines, or by heuristic evaluation" (Bevan et al. 2015: 143). Consistent with this view, major research efforts in the 1980s and 1990s were invested in identifying the attributes that made a product usable. According to the *user-oriented view* (Bevan et al. 1991), usability can be measured in terms of the mental effort and attitude of the user. In the *user performance view* (Bevan et al. 1991), usability can be measured by examining how the user interacts with the product, with particular emphasis on either ease-of-use (how easy the product is to use) or acceptability (whether the product will be used in the real world). The user-oriented and user-performance views emphasise that usability is contingent upon who used the product and with what goals (Bevan et al. 2015: 143). Representatives of this approach advocated that usability should be operationalised and evaluated in terms of actual outcomes of people interacting with the product and their satisfaction (Whiteside et al. 1988). These views are complemented by the *contextually-oriented view* (Bevan et al. 1991), that usability of a product is a function of the particular user or class of users being studied, the task they perform, and environment in which they work. Consistent with a contextually-oriented view of usability, Lewis highlights that usability is "an emergent property that depends on the interactions among users, products, tasks and environments" (Lewis 2012: 1267). Hence, usability should not be considered as a fixed concept. It emerges from the interaction among different agents, called *product* (i.e. the design element, also called *interactive system*) and *user* (i.e. the person for whom the product is designed) and it is contingent upon the physical as well as the extended context in which the interaction occurs (the *environment*) and goal-directed behaviour (the *task*) that the user attempts to enact via the product.

Foregrounding the utilitarian aspect of using a product, usability may be considered as a component of a product's acceptability, as proposed by Nielsen (2010), a reference figure in the field both as a scholar and as a practitioner and founder of the Nielsen Norman Group (NNG). Nielsen's conceptualisation links usability to other components of the product and highlights the considerations that usability must trade off against in a development project (Nielsen 2010). In Nielsen's view, "usability is a narrow concern compared with the larger issue

of system acceptability, which basically is the question of whether the system is good enough to satisfy all the needs and requirements of the users and other potential stakeholders, such as the users' clients and managers" (Nielsen 2010: 4). Based on this conceptualisation, usability is linked to a product's acceptability, usefulness, and utility (Figure 2.1). The overall *acceptability* of a product is a combination of its social acceptability and its practical acceptability (Nielsen 2010). *Social acceptability* concerns issues such as ethics and safety that need to be fulfilled before we can even consider other aspects (Nielsen 2010). For instance, there is not much point in worrying about the usability of an app for firearm trafficking. Provided that a product is socially acceptable, we can further analyse its *practical acceptability* within various categories, including cost, support, reliability, compatibility with existing systems, as well as the category of usefulness (Nielsen 2010). *Usefulness* is the issue of whether the system can be used to achieve some desired goal and it can be further broken down into the categories of utility and usability (Nielsen 2010). *Utility* is the question of whether the *functionality* of the system in principle can do what is needed and *usability* is the question of how well users can use that functionality (Nielsen 2010). We could say that utility is a matter of concepts and functionality and usability of how that concept is translated into design features.

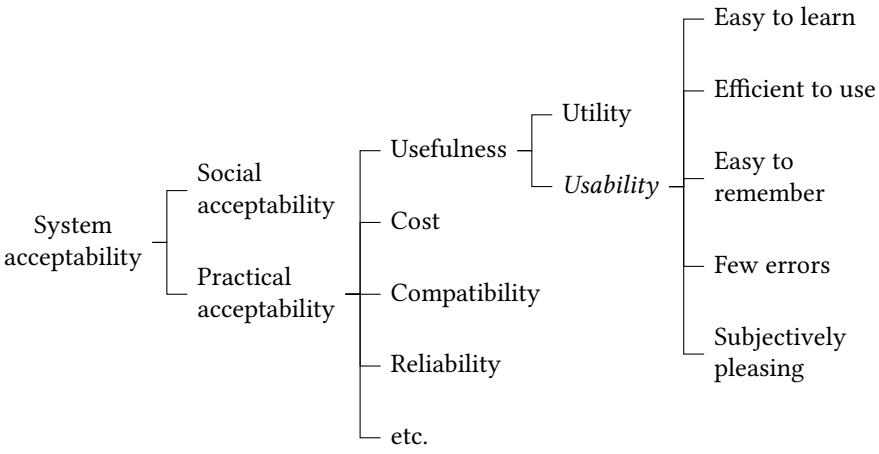


Figure 2.1: A model of the attributes of system acceptability (from Nielsen 2010: 6)

Usability itself is further defined by specific *attributes*, which represent the starting point for operationalising the concept, identifying measures and indicators, hence allowing to evaluate the system. Different sources propose different

attributes to define the usability concept. For instance, Nielsen (2010) proposes that a product's usability is given by its *learnability* ("the system should be easy to learn so that the user can rapidly start getting some work done with the system"), *efficiency* ("the system should be efficient to use so that once the user has learned the system, a high level of productivity is possible"), *memorability* ("the system should be easy to remember so that the casual user is able to return to the system after some period of not having used it without having to learn everything all over again"), *errors* ("the system should have a low error rate so that users make few errors during the use of the system, and so that if they do make errors they can easily recover from them. Further, catastrophic errors must not occur"), and, finally, users' *satisfaction* ("the system should be pleasant to use so that users are subjectively satisfied when using it").

The norm 9241-11 on the Ergonomics of Human Computer Interaction (ISO 2018), previously ISO (1998), attempts to combine the user-oriented view of usability with the performance-oriented and the contextually-oriented views. In the norm, usability is defined as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO 2018). Hence, to be defined as "usable" a product must support *specified users* (i.e. not just any user but those for whom the product was designed) in accomplishing *specified goals* (i.e. not just to perform any task but those tasks that the product aims to support). The outcome of using the product should be both *objectively* successful (i.e. effective and efficient) and *subjectively* successful (i.e. satisfactory). Within the same standard, more specific definitions of the effectiveness, efficiency and satisfaction with which a goal is achieved are provided. *Effectiveness* is defined in terms of the accuracy and *completeness* of users' performance and the protection from severe consequences in case of negative outcomes (inaccurate and incomplete performance). *Efficiency* is defined in terms of the resources (time, human effort, costs and material resources) that are expended when users attempt to achieve the goal (e.g. the time to complete a specific task). *Satisfaction* is defined as "the extent to which attitudes related to the use of a system, product or service and the emotional and physiological effects arising from use are positive or negative" (ISO 2018). Measures for the components of effectiveness, efficiency and satisfaction defined in ISO 9241-11 are provided in ISO/IEC 25022 on the measurement of quality in use (ISO 2016), as summarised below in Table 2.1.

As explained by Bevan et al. (2016), the inclusion of the satisfaction dimension in the revised version of the ISO standard acknowledges the increasing call from several authors to go beyond usability in purely utilitarian terms and consider the complexity of goals that people aim to achieve through products, not just of

Table 2.1: Measures of effectiveness, efficiency and satisfaction (from Bevan et al. 2016)

| Effectiveness | Efficiency | Satisfaction |
|---|---|---|
| Tasks completed | Task time | Overall satisfaction |
| Objectives achieved | Time efficiency | Satisfaction with features |
| Errors in a task | Cost-effectiveness | Discretionary usage |
| Tasks with errors | Productive time ratio | Feature utilisation |
| Task error intensity | Unnecessary actions | Proportion of users complaining |
| | Fatigue | Proportion of user complaint about a particular feature |
| | | User trust |
| | | User pleasure |
| | | Physical comfort |

practical but of emotional and social nature (e.g. Burmester et al. 2002). Proponents of this view encouraged an increased attention for users' motives for using products and their emotional response to the product. This led to the introduction of the term *user experience* (UX) alongside usability which may be defined as "a person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service" (ISO 2010). Some authors incorporate usability within UX. For instance, in Hassenzahl's (2003) UX model, usability is a *pragmatic attribute* of the user experience alongside utility. By contrast, *hedonic attributes* (stimulation, identification, and evocation) emphasise individuals' pleasure in the use of the product.

## 2.2 Usability engineering

At the origins of HCI, the question of how to ensure that a product is usable was first addressed with the methods of experimental psychology (Lewis 2012) – suffice it to say that Shneiderman's (1980) *Software Psychology* is considered to be one of the first books on the topic of HCI. The first graduate programmes in HCI emphasised the use of research methods from the behavioural sciences (Dumas 2007: 55). HCI later grew to incorporate methods and characteristics of other disciplines, such as ethnographical methods derived from the social sciences (Lazar et al. 2017). This evolution characterises HCI as a multifaceted and interdisciplinary field until today. However, at its onset, the discipline lacked an

identity of its own: "HCI was viewed, I believe, as a new area in which to apply traditional methods rather than a one requiring its own unique methods" (Dumas 2007: 55). In Dumas' reconstruction, the "leap forward" in the discipline's development happened when a series of publications argued for a new approach to product design and evaluation, which they called *usability engineering* (Whiteside et al. 1988).

The main driving force that separated HCI, and usability engineering specifically, from the standard protocols of other research traditions was the need to inform the design of products rapidly, at an early stage and throughout the development process (Lewis 2012: 5). As Buxton and Sniderman recount:

> Faced with a problem whose solution can not be derived from the literature, the designer of the human interface is confronted with two alternative strategies. On the one hand, a scientific approach can be followed, in which formal experiments are run, in an attempt to fill in existing gaps in our knowledge. On the other hand, an engineering approach can be taken, in which some ad hoc strategy is followed in order to deliver a system which was typically needed "yesterday". Due to pragmatics, the latter approach is by far the most prevalent. (Buxton & Sniderman 1980: 2)

There were several arguments in favour of an engineering approach to product design and development. One pragmatic argument was related to costs: "Under typical resource constraints, modifications will be feasible only in the prototyping stage. It is much too expensive to change a completely implemented product, especially if testing reveals the need for fundamental changes in the interface structure" (Nielsen 2012: 13). Furthermore, once the product-oriented view of usability (see above) revealed its limitations, HCI experts realised that usability could not simply be designed into a product by following a set of pre-defined and universally valid rules (Bevan et al. 2015: 143). Awareness developed that "design is complex, and there simply is no cookbook approach to design that can rely on general principles and guidelines alone" (Mayhew 2007: 918). While some fundamental principles and analogue precedents could inform the design, they could not alone ensure that a product was usable. Each product had to be treated as a unique piece and considered in its context to evaluate and improve its usability. What makes each project unique is that the design fundamentally involves human beings "who are, to put it mildly, complex" (Lazar et al. 2017: 1). Unlike in other research areas, in usability engineering the complexity of individuals cannot be discarded as a confounding variable. The designer is guided in his/her design choices by that very complexity.

The reasons why a usability engineering approach is needed are convincingly explained in Nielsen (1993) "usability slogans" – of which I will report and discuss the most relevant for the present discussion. A basic reason for the existence of usability engineering is that (1) *Your Best Guess Is Not Good Enough*: "it is impossible to design an optimal user interface just by giving it your best try. Users have infinite potential for making unexpected misinterpretations of interface elements and for performing their job in a different way than you imagine" (Nielsen 1993: 10). The designer should be aware that any initial attempt at a user interface design will include some usability problems and acknowledge the need to modify the original design to accommodate the user's problems. In other words, (2) *The User Is Always Right*: "The designer's attitude should be that if users have problems with an aspect of the interface, then this is not because they are stupid or just should have tried a little harder" (Nielsen 1993: 11). If the user is always right, then why not ask them what they want to see in an interface, use the most recurring requests as a bottom line and provide plenty of customisation flexibility to accommodate the variability in users' wishes? The reason is that, unfortunately, (3) *The User Is Not Always Right*. Users "do not know what is good for them ... Users have a very hard time predicting how they will interact with potential future systems with which they have no experience ... Furthermore, users will often have divergent opinions when asked about details of user interface design." (Nielsen 1993: 11–12). Another reason why the interface design cannot be left up to the user is that (4) *Users Are Not Designers*:

> Studies have shown, however, that novice users do not customize their interfaces even when such facilities are available [Jergensen and Sauer 1990]. One novice user exclaimed, "I didn't dare touch them [the customization features] in case something went wrong." Therefore, a good initial interface is needed to support novice users. Expert users (especially programmers) do use customization features, but there are still compelling reasons not to rely on user customization as the main element of user interface design. (Nielsen 1993: 12)

At the same time, designers cannot just trust their intuition because (5) *Designers Are Not Users* and because they are not users, they will not understand users' needs and struggles with the interface:

> When you have a deep understanding of the structure of a system, it is normally easy to fit a small extra piece of information into the picture and interpret it correctly. Consequently, a system designer may look at any given

> screen design or error message and believe that it makes perfect sense, even though the same screen or message would be completely incomprehensible to a user who did not have the same understanding of the system. Knowing about a system is a one-way street. One cannot go back to knowing nothing. It is almost impossible to disregard the information one already knows when trying to assess whether another piece of information would be easy to understand for a novice user. (Nielsen 1993: 13)

To understand why "knowing about a system is a one-way street", think of the picture puzzles where an animal picture is hidden. It may take a long time to figure out where the animal hides, but once you recognise it, you can't help seeing it. In the same way, designers know what is the best way to use the product they designed. They know what actions lead to a desired result and how to solve problems when they occur. Hence, they will not encounter the problems that a real user would. Unfortunately, when designing interactive systems, (6) *Details Matter* (Nielsen 1993: 15), which means that even seemingly minor interface details can drastically alter the overall usability of the interface, a further reason why a systematic usability engineering approach is needed to ensure that the product meets users' needs and is usable.

The fundamental traits of a usability engineering process were first articulated by Gould & Lewis (1985). The first trait is the *early focus on users and tasks*: the designer first seeks to understand who the users are, in terms of their cognitive, behavioural, social and attitudinal characteristics, and the nature of the work to be accomplished. The second is *empirical measurement*: intended users, i.e. representative of the users for whom the product is designed, should use product simulations and prototypes early in the development process to carry out real work, and their performance and reactions should be observed, recorded, and analysed. The third trait is *iterative design*: when problems are found in user testing, they must be fixed improving the solution. This means that the development process takes the form of several cycles of design, testing and improvement. Through each cycle, the solution is progressively refined.

We can hence see that usability engineering consists of a series of what we may call "design-focussed research activities" spanning throughout the whole product development process. The process for the "design of usable systems" was later conceptualised by Gould (1988) as consisting of four phases, which he called the "gearing-up phase", the "initial design phase", the "iterative development phase", and the "system installation phase". Today, the traits of the engineering approach may be found in popular design models such as the *Usability Engineering Lifecycle* (e.g. Mayhew 2007, Nielsen 1992, 1993), *User Centred Design* (e.g. Still & Crane

2017) and *Design Thinking* (e.g. Brown 2009). Although they operationalise the design and development process differently, they all roughly present phases similar to Gould's conceptualisation. For the sake of simplicity, I will call these phases *analysis*, *design*, *development*, and *implementation*. In the *analysis* phase, research activities focus on users and their tasks, as well as other aspects of the context in which the product is developed, such as existing competitor products. In the *design* phase, a concept, serving as the blueprint for the development, is created and validated. In the *development* phase, prototypes of increasing complexity are developed, tested, and refined iteratively. In the phase of *implementation*, the finished product is evaluated. Although these phases are presented as sequential, a substantial difference exists between processes in which these phases are executed in sequence (i.e. the Waterfall or Liner Sequential Life Cycle Model) or in an iterative, cyclical fashion (i.e. Spiral or Agile models; cf. van Kuijk et al. 2017).

The research activities in each phase have a different focus ad require different methods. The analysis phase draws on methods such as survey, interview, and contextual inquiry to understand users, their tasks and needs that the product should fulfil. Other activities such as review of relevant standards and product requirements are conducted to establish initial design principles. The evaluation of the product concept and its prototypes during the design and development phases may be described as *formative evaluation* whereas evaluation of the finished product, in the implementation phase, may be described as *summative evaluation* (cf. Pyla et al. 2005) – concepts imported from educational science (Scriven 1967). In the context of usability engineering, formative evaluation is explicitly aimed at informing design decisions and improving the solution *during* product design and development. Summative evaluation is conducted to ascertain some quality aspect of the product or its impact *after* it has been developed. Different types of evaluation at different stages of the development process call for different methods.

## 2.3 Empirical evaluation in usability engineering

Empirical evaluation during and after product development is the fundamental tool to ensure the usability of a product. For all the reasons explained above, each product is unique. Because designers do not know what will make that specific product satisfactory for the target users, constant empirical evaluation is needed to bridge the gap between designers and users.

Methods are usually broadly classified as either inspection or testing (e.g. Holzinger 2005). During *usability inspection* (Nielsen 1994), an evaluator inspects a

functioning or non-functioning prototype of the product. Inspection is aimed at generating results quickly and at low costs. Usability inspections are generally used early in the design and development phase and may involve non-functioning prototypes, such as paper prototypes. Paper prototypes capture the essential features of what the product will look like and allow for the evaluation of the initial concept before resources are invested in the development. Inspection methods generally involve experts and do not require them to accomplish real tasks with the functioning prototype. Typical methods include *feature inspection*, in which the evaluator reviews the list of product features intended to accomplish tasks, the *cognitive walkthrough*, in which an evaluator progresses through the steps in accomplishing the task through the product and simulates the problem-solving that the user engages in, and *heuristic evaluation*, in which evaluators compare the prototype against some pre-defined requirements (called heuristics).

By contrast, *testing* requires that the system be used by real users. Users should be representative of the target group not only in broad demographic terms (such as age, educational experience, technical experience, etc.), but also in terms of the task domain; for instance, that means that to study interfaces designed for lawyers, you must actually have practising lawyers taking part in the research (Lazar et al. 2017: 6). Testing offers higher reliability than inspection methods, but it is more time consuming and costly. Testing is used to identify and correct usability problems or to optimise the design.

Based on the stage of development and the specific aims, testing methods may vary in characteristics such as the setting (whether it is artificial or naturalistic), what is being evaluated (whether they are aimed at capturing users' performance or perception), how it is evaluated (whether predominantly qualitatively or quantitatively). The test *setting* may be either artificial, such as the company's premises or a laboratory, or naturalistic, i.e. the real context of use. Studies in fully naturalistic settings are only possible at the very final stage of development or in the implementation phase. An example of a fully naturalistic test is a *field study*, in which users are observed naturally interacting with the product after its release. An artificial setting usually also calls for a higher degree of experimental control in the test tasks. Examples include *usability testing*, which requires test users to use the functioning prototype to accomplish a set of tasks, *A/B testing* when two design alternatives are tested in controlled conditions (or A/B/C testing if three alternatives are involved), *think aloud*, where users are required to think aloud as they perform real tasks with the functioning prototype.

In line with the definition of usability being related both to how users perform on tasks through the product (the dimensions of efficiency and effectiveness)

and what they think of the product (the dimension of satisfaction), testing may gather data both related to users' *performance* with the product and *perception* of the product. Both types of data have their own limitations and both represent necessary evidence for the product's usability (Bevan et al. 1991). Performance data is more reliable as it shows what actually happens when users interact with the product. For instance, while a user may claim to find the product "helpful", an analysis of his/her performance may reveal that s/he performs better without the product. At the same time, performance data, especially of quantitative nature, can only be used to describe what happens but it can hardly clarify why. The internal state of users is important in providing further evidence for a products' usability. For instance, in some circumstances, accurate performance may be achieved by the user only at the cost of considerable mental and physical effort. Furthermore, especially qualitative perception data may help develop an explanation of why certain phenomena occur in the interaction with a product. However, perception data alone can hardly be a reliable and accurate descriptor of a product's usability.

The analysis of performance may focus on two main aspects (Lewis 2012: 1): one is *measurements* related to the accomplishment of task goals, i.e. of quantitative nature, and the other is related to the identification of "impasses" (such as signs of discomfort and confusion) in the interaction with the product, called *usage problems*, i.e. of rather qualitative nature. Examples of performance measures that are commonly used to evaluate the usability of a product are those provided in ISO/IEC 25022 on the measurement of quality in use (ISO 2016), as explained earlier. The identification of usage problems requires the observation of users' performance and an interpretive process on the researcher's side, as will be explained in greater detail in the following section on usability testing. *Quantitative performance* data may be used for (semi-)summative purposes to identify to what extent the product or the prototype meets usability goals. Furthermore, this type of analysis is only applicable to those tasks that can be separated out, quantified, and measured (Lazar et al. 2017: 8). *Qualitative performance* data is used for formative purposes and it is deemed as essential to improve the product: "Dr White's team knows that they cannot determine if they've met the usability performance goals by examining a list of problems, but they also know that they cannot provide appropriate guidance to product development if they only present a list of global task measurements" (Lewis 2012: 3).

Data related to users' perception, also called "attitude", is most commonly gathered through post-task questionnaires and interviews. Questionnaires may include both open and closed questions. Most commonly, pre-designed questionnaires like the User Experience Questionnaire (UEQ, Laugwitz et al. 2008), the

Systems Usability Scale (SUS, Brooke 1996), the NASA Task Load Index (NASA TLX, Hart & Staveland 1988) are used to obtain a quantitative description of users' perception. These instruments were developed and validated through research and are commonly used both in scientific work and in the evaluation of industrial products. Quantitative perception data may be used for formative purposes, to observe how users' evaluation of the products' usability changes as some design elements are modified through iteration cycles. It may be used for summative purposes by comparing the results against users' evaluation of analogous products of the same class or by matching them against industry benchmarks. Qualitative perception data gathered through interviews is used to gain a deeper understanding of what drives users' perception, thereby yielding a deeper understanding of target users, which may feed back into the analysis phase.

## 2.4 Usability testing

Among all methods that may be used to improve a product's usability, usability testing is often described as the royal road to usability. For example, Nielsen qualified it as "the most fundamental usability method" and claimed that it is "in some sense irreplaceable, since it provides direct information about how people use computers and what their exact problems are with the concrete interface being tested" (Nielsen 1993: 165). The major goal of usability testing is "to be practical and have a major impact" (Lazar et al. 2017: 266), or in other words to identify problems and propose data-driven solutions capable of improving the design of the product.

Since the term is used with some ambiguity, a terminological clarification is required before talking about its specific characteristics. Some authors have used the term in a broad sense to indicate any process or activity that aims to improve the ease of use of an interface, hence as a synonym of usability engineering (Lazar et al. 2017: 267). Other authors have referred to usability testing as "user testing" (e.g. Nielsen 1993), which may refer to any type of testing involving real users (as opposed to inspection methods). Another alternative has been "user research", although user research, as explained in Goodman et al. (2012), actually has a broader meaning and includes a family of inquiry methods and elements of design and development, such as personas, user profiles, card sorting, and competitive research which are not regarded as testing. Furthermore, in usability testing, the primary research focus is the product, not the user, although as Lazar and colleagues point out it "can be used to learn more about how people interact with interfaces, even when the goal is not fixing the interface, but in-

stead learning more about users and interactions. So, in usability testing, maybe we are researching the user?" (Lazar et al. 2017: 263)

Having clarified what it is not, usability testing may be defined as a test method conducted on representative users attempting representative tasks, on early prototypes or working versions of products (Lewis 2012). This definition entails the key tenants of usability testing. First, usability testing involves representative users. Given that usability is established in relation to the relevance of a product for a particular user and aim, study participants should be representative of the target users for which the product was designed. During the test, users work with a working prototype or the product in action to accomplish representative tasks. This means that the test tasks should be real and meaningful to them, and they should represent as closely as possible the tasks that users typically accomplish with the product in real life. Furthermore, the test task should lead participants to interact with all product functions for which testing is required.

Seewald & Hassenzahl (2004) define usability testing as the systematic observation of users in a controlled environment to ascertain how well they can use the product. Hence, usability testing possesses methodological characteristics both of experimental research (cf. also Rubin & Chisnell 2008: 23), because test tasks should be adequately designed to be representative, and of ethnography, in the use of observation. Furthermore, the analysis of performance is often corroborated with perception data to tap into participants' perception and interpretation of their experience with the product, which may be gathered both through post-task questionnaires and interviews (Lazar et al. 2017: 267, Rubin & Chisnell 2008: 65). Verbal protocols, such as think-aloud, may also be used to uncover the mental processes inherent to task execution. However, thinking aloud impacts some aspects of performance (such as time on task) and therefore it excludes the possibility to use such metrics in the analysis. Usability testing may hence be defined as a type of convergent mixed-method research (Creswell & Creswell 2017).

Users' performance on test tasks may be analysed *quantitatively* – on usability metrics discussed earlier such as time on task, success rate, etc. – to ascertain the extent to which a certain product feature leads to desired performance. However, usability testing usually has a strong *qualitative*, *observational* component (Seewald & Hassenzahl 2004). The observation of users' interaction with the product is aimed at identifying "flaw areas of the interface that need improvement" (Lazar et al. 2017: 264). An interface flaw is some aspect of the interface which is confusing, misleading, or generally suboptimal (Lazar et al. 2017: 264). Interface flaws should cause a problem for most people. It is not about font style or colour preferences, but rather about an element that most people stumble upon

with consequences on performance. The manifestation of problems during the interaction is called a *critical incident* (Seewald & Hassenzahl 2004). A critical incident is constituted of a task-related goal, an "activity" of the user (an action as well as a verbal or non-verbal expression) which is discordant with the goal, and a consequence (Seewald & Hassenzahl 2004). An error in the execution of the task, users' facial expressions (e.g. frowning, squinting at the screen, etc.) and verbal reactions (e.g. "what?!" or "where is the button?") all point to them experiencing some form of problem or discomfort with the product and conflicts with the goal of effective and efficient interaction. The interpretation of a common class of critical incidents as being caused by a specific design element leads to the identification of usage problems (Seewald & Hassenzahl 2004).

The identification of usage problems lies at the centre of usability testing. According to Seewald & Hassenzahl's (2004) model, the usability testing process is constituted of the following five steps:

1. *Experience* (data collection): critical incidents are identified through observation and questioning of study participants.

2. *Construction* (data analysis): the critical incidents are coded or classified.

3. *Interpretation*: a possible cause for a common class of incidents, representing a usage problem, is identified.

4. *Prioritisation*: usage problems are ranked based on the researcher's judgment of the "severity" or "impact" of the problem or more subjective measures such as the "frequency" of the problem. An indicator of the stability of the occurrence ("Auftretensstabilität", Seewald & Hassenzahl 2004: 145) can be the number of participants who made a certain mistake, as well as quantitative performance indicators such as task completion rate, success rate, error ratio, etc.

5. *Recommendations*: based on the interpretation of usage problems, the researcher develops recommendations on how to improve the product, often collaboratively with the design team.

Given that usability tests are often conducted with observational techniques and data are collected combining multiple sources, the studies may be very time intensive and therefore allow for only a limited number of participants. The ideal number of study participants for a usability test is still a matter of debate in the scientific community. Nielsen & Landauer (1993) developed a mathematical model suggesting that a test on five users can uncover 85% of usage problems.

More recent studies have suggested that the ideal sample size is 10 ± 2 users (Hwang & Salvendy 2010) or even more than ten users (Schmettow et al. 2013) for high-risk systems. It is generally accepted that qualitative usability testing can be based on around five users if the primary aim is to improve a product and/or to deepen understanding of the user through in-depth interviews. Quantitative usability testing, by contrast, aims to find metrics that predict the behaviour of a larger population and requires a larger sample size.

# 3 Translation technology

This chapter analyses translation technology, computer-assisted translation (CAT) tools in particular, and related research from a usability engineering perspective (discussed in Chapter 2). The analysis will serve as a means of comparison for the review of interpreting technology, computer-assisted interpreting tools in particular, and related research in the following chapter (Chapter 4). After a brief definition of *translation*, the chapter unravels the history of translation technology and the development of computer-assisted translation tools. It then provides a succinct review of CAT tool research, with a focus on studies dedicated to informing the design of CAT tools through the analysis of users' needs and requirements as well as through empirical testing. The chapter concludes with a discussion of the extent to which a usability engineering approach is reflected in the development of CAT tools.

## 3.1 Developments and current landscape

Translation may be defined as "the result of a linguistic-textual operation in which a text in one language is re-produced in another language" (House 2014: 1). Theories of what translation is and how to translate began to emerge as early as in the writings of Cicero and Quintillian, whose debate on translation practice pertained mostly to *how* a text should be translated – either word-for-word or sense-for-sense. Despite these ancient origins, the actual birth of *Translation Studies* (TS), as the discipline that studies the theoretical and practical aspects of translation, goes back to Holmes's (1975) seminal work "the name and nature of translation studies". Through time, theorising on translation has been informed by other disciplines and different worldviews, often referred to as turns in TS (e.g. Snell-Hornby 2006). Different perspectives have led to a substantial redefinition of the concept of translation (Cheung 2011) and through time TS has differentiated into a plethora of approaches and lines of research (Baker & Saldanha 2019). For society at large, translation is a technology – a means to communicate across language differences:

> The infinite extension of the written symbol through time requires the good offices of the translator. Similarly, when the written symbol is considered in terms of spatial range, the default multilingualism of the planet means that there can be no extension in space without the work of the translator. The text needs translation to travel. So the afterlife of the text, dependent on elements of the third system, the artefacts of "ink" and "paper", relies also on the tool of language, and, by extension, translation, for its ability to reach "readers" in a different time and space. (Cronin 2012: 22)

At the same time, translation is, by its own nature, dependent on writing technology, be it stone inscriptions, ink quills on parchment, typewriters, or the personal computer. Because of this dependency on writing technology, the activity of written translation came after that of oral interpreting (see Chapter 5).

When translation scholars discuss the relationship between translation and technology, they commonly refer to the introduction of digital technology into the profession (cf. Chan 2015). Digital technologies changed "how translations are commissioned, produced, distributed and consumed or used" (Jiménez-Crespo 2020: 2). The impact of digital technologies has been profound and affected the translational *micro-* and *macro-systems* (O'Hagan 2013). ICTs created new forms of translating, such as crowdsourced translation and localization. Digital tools came to play a central role in all phases of the translator's work, so much so that by the beginning of the 21st century no translator could expect to work without digital tools any longer (Bowker 2002). ICTs have become so deeply entrenched into translation that all of translation may be seen as a form of *human-computer interaction* (O'Brien 2012). The deep influence that technologies had on translation made scholars refer to it as a technological turn (e.g. Chan 2007, Cronin 2010, O'Hagan 2013). This turn not only led to significant shifts in the way in which translation is carried out in the contemporary world, but also led to a systematic re-examination of the "conventional understandings of what constitutes translation and the position of the translator" (Cronin 2010: 1). Hence, the technological turn, like previous turns, has led to a re-definition of the very concept of translation. From a disciplinary perspective, technology permeates translation studies across its different subdisciplines, both in their theoretical apparatus and/or in their research methodologies (cf. Jiménez-Crespo 2020).

The digital technologies that have so shaped the translator's work are varied and may be summarised as five broad categories: the translator's computer equipment, communication and documentation tools, text edition and desktop publishing, language tools and resources, and translation tools (Alcina 2008). Translation tools, which are the main focus of scholarly debates on technology

and translation, may be further divided into two categories: technologies aimed at *replacing human translators* and those aimed at *supporting human translators* (Alcina 2008).

Technologies aimed at replacing human translators have the ambition to generate fully automatic translation. Efforts to develop such technologies began to emerge in the 1950s and they are referred to as *machine translation* (MT) (Hutchins 2015). The initial enthusiasm for machine translation, which started from the second half of the 1940s, led to the realisation around the first half of the 1960s that machine translation needed to be revised by humans, leading to twice the costs of human translation and a raw output that was unable to meet expectations; therefore, fully automatic machine translation was not going to be realized for a long time (Chan 2015: 3–4). Today, MT produces large volumes of translation and is commonly used for people to get access to content: "Google translation services are used by more than 500 million people every month, producing more than 1 billion translations a day. Every day, Google translates more text than all human translators translate together in a whole year" (Carl & Braun 2017: 377). Despite the uncertainty about whether MT will ever attain sufficient quality to make it a stand-alone solution, the spread of MT has had a psychological impact on translators:

> Over the last decade, speaking to audiences in different parts of the world, the same questions keeps returning: Is there a future for translators? In the age of Google Translate, is the human translator condemned to large-scale extinction, or to the quaint peripherality of the Sunday hobbyist? The demand for translation keeps growing apace in the contemporary world, but will humans continue to be asked to service this need, or will it be our machines that will do the bidding? (Cronin 2012: 1)

This initial disappointment with MT in the 1960s led to interest in developing technology to support human translators, increasing its efficiency and productivity and reducing its costs (Chan 2015: 4). The use of technology to support the translator's work is known as computer-assisted, or -aided, translation (CAT), Chan (2015: 4) reports the United States' Automatic Language Processing Advisory Committee's (ALPAC) 1966 report as a turning point in the evolution of translation technology. After establishing the failure of MT, the report suggested that MT should shift to machine-aided translation, which would be "aimed at improved human translation, with an appropriate use of machine aids" (ALPAC 1966: iii), and that "machine-aided translation may be an important avenue toward better, quicker, and cheaper translation" (ALPAC 1966: 32). Specialised CAT

tools began to be developed for this purpose in the second half of the 1960s (Chan 2015).

Hutchins' (1998) seminal paper attributes the origin of CAT tools (specifically, of the *translation memory* concept, discussed later) to Arthern (1978), Head of English Translation Division at the European Commission (at the time, Council of the European Communities). In his paper, Arthern observed that translators working at the European Commission were wasting valuable time by retranslating (parts of) texts that had already been translated. Hence, he proposed to develop a digital storage of source and target texts for translators to re-use in current translations, speeding up the process. This solution may be referred to this as "translation by text-retrieval" (Arthern 1978: 94).

Already the year before, in 1978, Melby of the Translation Research Group of Brigham Young University conducted research on machine translation and developed an interactive translation system, ALPS (Automated Language Processing Systems) based on "Repetitions Processing", which aimed at finding matched strings (Melby 1978 cited in Chan 2015: 4).

Another milestone was laid by Kay in his paper titled "The proper place of men and machines in language translation" (Kay 1980). He argued that MT and translators should work together and proposed the development of a software split in two windows: the upper window for the source text to be translated and the bottom window as a workspace where the translator could edit the translation with aids for word selection and dictionary consultation. Chan points out that "In view of the level of word-processing capacities at that time, his [Kay's] proposal was inspiring to the development of computer-aided translation and exerted a huge impact on its research later on" (Chan 2015: 5).

## 3.2 Computer-assisted translation (CAT) tools

Computer-assisted translation tools may be defined as "a specialized set of tools that are designed specifically with translation in mind" (O'Hagan 2013: 504). Since MT is now commonly used as a support to human translation, although it was not necessarily designed with the human translator in mind, another possible definition of CAT tools is: "tools that are (relatively) specific to the translation process" (Krüger 2016a: 121). There are several "specialised instruments" aimed for specific aspects of the translator's work. The label "CAT tool" may be used to denote both the specialised instruments in isolation and their integration into a comprehensive solution – which may also be referred to as a "translation environment" (Coppers et al. 2018: 1).

CAT tools can support different phases of translators' and organisations' work, for instance as described in the *Cologne model* (Krüger 2016a,b). Paraphrasing and summarising the main phases of the model (cf. Krüger 2016a,b for a detailed discussion), CAT tools can facilitate the phases of client acquisition, work preparation and coordination among translators through project management functions – i.e. the *project initiation* and (*general* and *translator*) *preparation phases*. CAT tools assist translators in the *actual translation phase*, where the text is translated from the source to the target language. After the translation is completed, CAT tools support the *quality control phase*, in which the translation is revised, the *final administrative work* and the *follow-up work*. In the present discussion, I will focus on the CAT tool components that are aimed to support the actual translation phase. The use of CAT tools during the translation process has been of particular interest for scholars, who maintain that it may have a profound impact on the translator's cognition. In the words of Pym, CAT tools (or their components) used in the translation phase "far from being merely added tools, are altering the very nature of the translator's cognitive activity" (Pym 2011: 1). Furthermore, these may be seen as the equivalent of in-booth CAI tool use, which is the focus of this work.

CAT tools (components) for the translation process were initially developed based on the assumption that translators' work would become more efficient if it could be provided with automated translation aids. This idea originated *translation memory* (TM) systems: "a translation memory stores sentences and their human translation from a specific domain. Given a source segment, the translation memory provides the user sentences that have the same or a similar vocabulary and/or grammar (Coppers et al. 2018: 1). Another more basic type of translation aid is provided by CAT tools through the *term base*, which stores terms and their metadata. Finally, MT itself is integrated into CAT tools as a form of translation aid. The MT engine pre-translates the source-text which is then edited by the translator – an activity commonly referred to as "post-editing" (e.g. Flanagan & Christensen 2014). In some cases, the machine translation engine can adapt its output during translation based on the translator's corrections (Coppers et al. 2018: 1). In professional settings, MT engines are typically trained using in-house TM databases and term bases – a system also referred to as "MT-assisted TM" (Christensen et al. 2017). Since the translator edits the MT output which feeds back into the TM databases and term bases, Mt-assisted TM blurs the traditional distinction between MT, post-editing and TM-assisted translation (Christensen et al. 2017: 9).

## 3.3 CAT tool design, development and reception

While the quest for adequate solutions that could make the machine-translator symbiosis possible started in the 1980s at the ideational level, commercially viable CAT tools first became widely available in the early 1990s. The need to increase the efficiency of human translation was stimulated by the "globalisation turn" (Snell-Hornby 2006) and improved ICTs:

> CAT systems were developed from the early 1990s to respond to the increasing need of corporations and institutions to target products and services toward other languages and markets (localization). Sheer volume and tight deadlines (simultaneous shipment) required teams of translators to work concurrently on the same source material. In this context, the ability to reuse vetted translations and to consistently apply the same terminology became vital. (Garcia 2015: 63)

The first commercial CAT tools were developed by the German-based Trados and the Swiss-based STAR AG. Since then, the growth in number has been exponential:

> Before 1993, there were only three systems available on the market, including Translator's Workbench II of Trados, IBM Translation Manager / 2, and STAR Transit 1.0. During this ten-year period between 1993 and 2003, about twenty systems were developed for sale, including the following better-known systems such as Déjà Vu, Eurolang Optimizer (Brace 1994), Wordfisher, SDLX, ForeignDesk, Trans Suite 2000, Yaxin CAT, Wordfast, Across, OmegaT, MultiTrans, Huajian, Heartsome, and Transwhiz. This means that there was a sixfold increase in commercial computer-aided translation systems during this period. (Chan 2015: 8)

CAT tools were first equipped with basic components, such as translation memory, terminology management, and translation editor. With the growing sophistication if ICTs, more functions were developed into CAT tools and more components were gradually integrated into these systems (Chan 2015: 12).

TRADOS[1] (Translation & Documentation Software) was developed by the German engineers Jochen Hummel and Iko Knyphausen. It soon became the industry standard, partly thanks to successful tender bids to the European Commission in 1996 and 1997 (Garcia 2015: 70). From the late 1990s, TRADOS' technology was integrated into other CAT tools too and its products became the most popular in

[1] https://www.trados.com

the industry (Chan 2015: 13). The acquisition by SDL further supported its commercial success. SDL Trados, the new name of Trados, was the first company to integrate a translation memory into a CAT tool. In 2009, the release of SDL Trados Studios 2009 marked the shift towards an integrated CAT tool with all translation aid components integrated in a single interface. A survey conducted by Proz.com, the largest networking website for professional translators, in 2013[2] suggests that Trados remains the market leader among CAT tools. It is estimated that TRADOS today holds over 70% of shares in the global language translation software market (Cheng 2021).

The UIs of CAT tools present a series of features aimed at enabling translators to incorporate translation aids into their translation. The system front-end that translators use to create the translation is called *editor*. In the editor, translators open the source file for translation, query the TM systems and databases for relevant data and/or post-edit the MT output (Garcia 2015: 71). A CAT system editor segments the source file into translation units to enable the translator to work on individual segments separately and the program to search for matches in the memory (Garcia 2015: 72). Inside the editor window, the translator sees the active source segment displayed together with a workspace for the target text, where matches are shown for the translator to review, edit or translate from scratch (Garcia 2015: 72). The workspace can appear below (*vertical presentation*, as in Trados and Wordfast) or beside (*horizontal* or *tabular presentation*, as in one visualisation option in Déjà Vu) the currently active source-text segment. *Segmentation* and *orientation* (vertical or horizontal) are two of the most commonly debated UI parameters of CAT tools. The similarity between text segments that must be translated and previously translated segments in the TM is expressed as a *matching score* which can be of three main types (Bowker & Fisher 2010: 61): *exact match* (100% correspondence), *fuzzy match* (about 60–70% correspondence), *sub-segment match* (if a portion of the segment is recognised by the TM), *term match* (if a term in the source segment is present in the term base), *no match* (if the TM fails to detect previously translated material).

Today, CAT tools are commonly used in professional translation. However, scholars and professionals have expressed mixed views concerning their impact on human translation. Positive views contend that CAT tools increase the efficiency of time-consuming and error-prone tasks, such as translating repeated text portions in a consistent fashion (O'Brien 2012). Because of this, CAT "has contributed to increasing speed, improving consistency, reducing costs" (Ehrensberger-Dow & Massey 2014b). At the cognitive level, CAT tools "have

[2]https://go.proz.com/blog/cat-tool-use-by-translators-what-are-they-using

extended translators' memory by externalizing it, thus decreasing the load on working and long-term memory" (Pym 2011).

Negative views claim that CAT tools may change the process and the very nature of translating. Pym (2011) contends that CAT tools impose the paradigmatic constraints on the syntagmatic nature of the practice as they cause translation to lose its linearity and drift further away from the perception of translating as an act of translator-mediated communication between people. The constant use of translation aids transforms the translator into a "de-facto post-editor" (Christensen et al. 2017). Some hypothesised disadvantages are of psychological nature: they emerge from translators' perception of and relationship with technology, with several factors potentially contributing to a (perceived or actual) downgrading of the professional status of the translator and subsequent feelings of "dehumanisation" and "devaluation" (O'Brien 2012). These challenges are ascribed to the view of the translator as a fixer of MT errors, as well as translators' lack of knowledge of MT which is seen as a black box, "something they do not quite understand and which removes them further from the task of translation" (O'Brien 2012: 109). This feeling of disconnection and distrust manifests on social media, where professional translators express doubts about the usefulness of technologies like MT (Läubli & Orrego-Carmona 2017). Research carried out at the European Commission's Directorate-General for Translation suggests translators felt uncomfortable with MT technology despite having used it for many years (Cadwell et al. 2016). From a cognitive perspective, the complexity of many of the newer CAT interfaces may increase the cognitive demands on their users (Hansen-Schirra 2012).

A major point of controversy, denounced by scholars and professionals alike is that translators were not sufficiently involved in the development of CAT tools. Moorkens and O'Brien, for instance, concluded that users' needs where insufficiently accounted for in the design of CAT tools: "prior to the current research there has not been a focus on what functionality users would like to see in a tool for post-editing as the MT research community has had a tendency to "focus on system development and evaluation" rather than considering end users (Doherty 2019: 4)" (Moorkens & O'Brien 2017). Observing the lack of a user-centred design, scholars plea for iterative design processes in which users' feedback and data is constantly used to validate and refine the solution (Läubli & Green 2019) – i.e. they call for a usability engineering approach (see Chapter 2) to the design and development of CAT.

Läubli & Green 2019: 381 report of the development of Intellingo (Coppers et al. 2018) as an out-of-the-ordinary example of a user-centred development approach to the development of a CAT tool. Intellingo is an intelligible CAT tool

showing contextual information to the translator, such as metadata about where the translation aids originated from (term base, translation memory, the machine translation engine, or a combination of these resources). The rationale for the development of such a system is that increasing the intelligibility of translation groups could increase users' trust in the tool: "CAT tools offer some form of intelligibility, by showing matching scores, for example. However, the inner logic of the algorithms is rarely shown, and there is plenty of potential to enhance their intelligibility" (Coppers et al. 2018: 3). The authors report in the paper that the whole software was developed in successive iterations and started from surveys and contextual enquiries to develop the tool based on translators' requirements. They then tested two UIs of the tool to check whether the intelligibility function served the translator and to what extent it was used. The development team started by surveying 180 translators, and then iterated over mock-ups and functional prototypes, involving translators from the outset. Also the project by Teixeira et al. (2019) present an iterative development process of a web-based translation editing interface that permits multimodal input via touch-enabled screens and speech recognition in addition to keyboard and mouse. Two usability studies were conducted between iterations and reported in the paper. These latter virtuous examples, however, arose within research projects with a limited impact on the CAT tool market. An example of a translator-centred CAT system with a commercial application is Predictive Translation Memory (PTM) developed by Green et al. (2014). PTM is a type of interactive translation memory (ITM) system. It aims to facilitate the interaction between translators and CAT tools as well as the integration of TM suggestions into the translator's work through a UI reducing gaze shift. In their paper, Green and colleagues explain that translators' discontent with existing systems led to the development of the PTM and detail the lare-scale evaluation study. PTM was later integrated into Lilt's CAT tool editor. Lilt[3] is a translation and technology business founded by John DeNero and Spence Green. It provides translation and localisation services to clients worldwide, supporting their pool of freelance translators with a dedicated CAT tool, based on ongoing research[4] such as the aforementioned paper and others.

## 3.4 CAT tool research

CAT tool research may be positioned within the TS subdiscipline of translation technology research. Given the pervasiveness of technology in the trans-

[3] https://lilt.com/

[4] https://lilt.com/research

lation profession, "the study of translation in one-way or another requires acknowledgement of [the] interrelationship [between translators and technologies]" (Jiménez-Crespo 2020: 2) and technology may be seen as a "connecting thread across sub-disciplines of TS and diverse research areas" (Jiménez-Crespo 2020: 3). I will define translation technology research as research that is specifically focused on translation technology, technology-dependent phenomena as well as the impact of technology on the translator and translating – as opposed to TS research using technology as a necessary medium of translation (e.g. a word processor) without focusing on the impact of the medium on the translation process/product. CAT tool research, in turn, may be defined as a line of research concerned with "the design and adaptation of strategies, tools and technological resources that make the translator's job easier as well as facilitating the research and teaching of such activities" (Alcina 2008: 90).

In TS, translation technologies in general, and CAT tools, in particular, have been studied from a variety of perspectives, drawing explanatory concepts and frameworks from other disciplines (cf. Olohan 2019). Although they may not have an explicit design focus, studies of predominantly cognitive or descriptive nature may yield implications for CAT tool development (e.g. Mellinger & Shreve 2016). Research with an explicit focus on the extent to which CAT tools are adequate to translators' needs and what UI features can support or inhibit the work of translators are most commonly aligned to either an *ergonomics approach* (cf. Ehrensberger-Dow 2019) or a *human-computer interaction* (HCI) *approach* (cf. Läubli & Green 2019).

The ergonomics approach to the study of translation technology encompasses cognitive, physical, and organizational ergonomics to investigate the impact of various factors on the situated activity of translation, including the use of CAT tools (Ehrensberger-Dow & Massey 2014a: 63). This approach was informed, on the one hand, by the cognitive paradigm in TS (cf. Walker 2021: Chapter 2), which called for explorations of issues such as cognitive load and mental processes inherent to translators' use of CAT tools as well as for methods of the translation-process research (TPR) tradition such as key-logging, screen recording, and eye-tracking (cf. Jakobsen 2017). On the other hand, research on the ergonomics of CAT tool was informed by the theory of situated translation (cf. Risku 2002, 2004), grounded in situated cognition theory (e.g. Brown et al. 1989), which called for naturalistic workplace studies aimed at exploring the role of technology in the translator's cognitive ecosystem (cf. Ehrensberger-Dow 2019, Ehrensberger-Dow & O'Brien 2015).

Scholars aligned with the ergonomics approach typically draw on the International Ergonomics' Association's (IEA) definition of *ergonomics* as "the scientific

discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data and methods to design in order to optimize human well-being and overall system performance" (IEA, cited in Ehrensberger-Dow 2019). They also distinguish, with the IEA, between physical, organisational, and cognitive ergonomics. *Physical ergonomics* is defined by the IEA as "human anatomical, anthropometric, physiological and biomechanical characteristics as they relate to physical activity". In the context of translation, this essentially means exploring the impact of the characteristics of the translators' workplace, such as the tables, chairs, computers, noise level etc. (Ehrensberger-Dow 2019: 38). *Organisational ergonomics* "is concerned with the optimization of sociotechnical systems, including their organizational structures, policies and processes." In the context of translation technology research, this has meant a concern with the extent to which the increasing technologization of the profession has impacted the translator's professional status and agency (Ehrensberger-Dow 2019: 39). *Cognitive ergonomics* refers to the "mental processes, such as perception, memory, reasoning, and motor response, as they affect interactions among humans and other elements of a system". O'Brien and colleagues explain the primary concern of CAT tools' cognitive ergonomics research as "understanding of cognitive friction arising from translators' interactions with non-optimal CAT tool features" (O'Brien et al. 2017: 147). Cooper's concept of *cognitive friction* (Cooper 2004) has been used as a primary explanatory concept in the study of CAT tools' cognitive ergonomics. It is defined as "the resistance encountered by a human intellect when it engages with a complex system of rules that change as the problem changes" (Cooper 2004: 19). The concept was first introduced to the study of CAT tools by O'Brien 2012: 110, who paraphrased it as "the tension between translators and computers." It was then further explored by (Ehrensberger-Dow & O'Brien 2015: 102) who defined it as a disturbance to the translation *flow*. Flow is understood as a psychological state of being fully immersed in a task such that this immersion is energising (Nakamura & Csikszentmihalyi 2002, cited in Ehrensberger-Dow & O'Brien 2015). The interruption of the state of flow by sub-optimal CAT tool features is thought to manifest as the translator's *irritation* with those tools and to increase the cognitive load in the translation task (O'Brien et al. 2017: 146). The assumption is hence that "since being irritated can affect negatively performance, improvements in the cognitive ergonomics of translator tools could contribute to better decision-making, creativity, and efficiency" (Ehrensberger-Dow 2019: 43). Therefore, CAT tools "should be designed in such a way that they aid cognition and do not become a potential source of cognitive friction" (Teixeira & O'Brien 2017: 81).

The ergonomics approach to translation technology research, particularly the cognitive ergonomics approach, has contributed to increasing the field's awareness of the importance of CAT tool usability, shedding light on translators' needs and the features of CAT tools that might decrease their usability. However, as noted by Kruger, "the issue of CAT tool usability is not addressed specifically in cognitive ergonomics research, and the investigation of translation technology remains at a rather coarse-grained level, being mostly concerned with shortcomings of user interfaces and the possibility to customise tool settings according to individual preferences" (Krüger 2016a: 128).

More fine-grained usability research has been conducted within what Läubli & Green (2019) refer to as the human-computer interaction (HCI) approach to the study of translation technology. While the ergonomics approach has been mostly influenced by translation theory and led by TS scholars, the HCI approach draws explicitly on concepts and research methods from the field of HCI and usability. The major contribution of studies within the HCI approach is the development of requirements for CAT tool interface and design recommendations through rigorously designed empirical studies (e.g. Läubli 2020, Läubli et al. 2021). Such data may provide compelling arguments for the adaptation of CAT tools to interpreters' needs:

> However, the impact of poor usability on translator performance has rarely been tested empirically, and since the motivation for using CAT tools is primarily economic – saving time by leveraging translation suggestions rather than translating from scratch – the design of these tools is unlikely to change until measurements show that alternative designs speed translators up or cause them to make fewer mistakes. (Läubli 2020: 1)

Within this approach, we also find examples of translator-centred, iterative development of CAT solutions (Coppers et al. 2018, Green et al. 2014, Teixeira et al. 2019), although these were developed within research projects without a direct influence on market realities.

The review below addresses previous research on CAT tools that has explicit implications for their design. The focus will be on the most common methods used to develop recommendations for the design and their further improvement. I will first consider research providing input on users' needs and general CAT tool requirements. Then, I will examine research focused on the empirical evaluation of CAT tools via tool performance, users' performance, and users' perception.

### 3.4.1 Need analysis and tool requirements

The analysis of translators' needs in the use of CAT tools has been the focus of a wealth of studies, the majority of which were conducted by TS scholars. Through the years, several explorations elicited information about the extent to which CAT tools are employed, which tools are most used and users' perceptions of their usability. Several methods were used, with surveys and interviews followed by contextual enquiry being the most common methods. Other studies used user research methods, mixed-method designs and literature review/theoretical modelling to develop requirements for CAT tool design.

Survey was one of the first methods to be used to explore translators' needs and the usability of CAT tools and today remains a popular method to develop requirements for CAT tool design (e.g. Schneider et al. 2018). Surveys concerned specifically with the usability of CAT tools began to emerge in the noughties, about a decade after CAT tools began to appear on the market. Today, surveys remain The survey by Lagoudaki (2006) is considered one of the first surveys dedicated to the usability of TMs. One of the conclusions of the survey, which collected the responses of 874 professionals, was that respondents felt that their needs and usability issues had not adequately been accounted for in the design and development of TMs. The author hence recommended that "user engagement be pursued in all stages of software development" (Lagoudaki 2008: 205, cited in Vargas-Sierra 2019). Among the most renowned studies that employed survey methods, Moorkens & O'Brien (2013) asked posteditors to describe the features of their ideal TM UI. The usability of CAT tool UI was further studied in an international survey within the *ErgoTrans* project (O'Brien et al. 2017), which aimed to identify the specific features of CAT tools that translators found "irritating" or that they felt were missing. Among the most common irritating features, the complexity of the UI and text segmentation were mentioned as disturbing elements. Users reported that they felt overloaded by the crowded UI. The authors also reported that much of such irritation could potentially be reduced by customising the interface through basic tool settings but less than half (44%) of CAT users in the study used customisation options.

Interviews are another method that has been largely employed to investigate translators' needs and define CAT tool requirements. For instance, following up on survey results, Moorkens & O'Brien (2017) conducted an in-depth exploration aimed at identifying user requirements for post-editing tools and possible UI features capable of fulfilling those requirements. To accomplish this aim, they interviewed experienced post-editors. They found that a major source of frustration was the non-responsiveness of MT output, which forced users to make the same

changes, again and again. Another key point that emerged was posteditors' need to know the provenance of TM and MT data, which justified the development of intelligible systems (e.g. Intellingo in Coppers et al. 2018).

Other scholars proposed creative alternatives to the use of surveys and interviews drawing on user research methods. Koskinen & Ruokonen (2017) asked their study participants ($N = 102$ professional translators) to write either a break-up or a love letter addressed to a tool of their choice to gain insight into tool usability but also the broader perception and needs of users. 70% of the collected letters addressed translation technology, mostly search tools and databases rather than TM or MT. The authors mapped translators' comments onto the usability dimensions of learnability, efficiency, memorability, errors and satisfaction. Läubli & Orrego-Carmona (2017) analysed translators' comments related to translation technology and, in particular, MT on Twitter to gain insight into their attitude and the relationship between practitioners' perception and scientific knowledge.

Ethnographic research has been another approach of choice for the development of CAT tool requirements. For example, Asare's (2011) PhD thesis examined the workflow translators working at an agency as a case study for how CAT tools were perceived by users, which features were being used and which ones were not being used. The study concluded that "a number of features in the translation tools were not being used because their purposes were not understood by the tool users" (Asare 2011: 138), pointing out a discrepancy between the CAT tool designer's intention and actual use. Another renowned ethnographic case study was conducted in Canada by LeBlanc (2013) in three medium-sized translation agencies. The methods used were (1) semi-directed interviews with translators (as well as informal conversations, accounts and testimonials) and management, (2) participant observation of translators at work at their workstations (shadowing), and (3) contextual information on the texts, the clients, the service or the firm (LeBlanc 2013: 3). The aim was to explore the perceived advantages and disadvantages of using TMs at the workplace. One of the conclusions was that part of translators' dissatisfaction with TMs "revolves around the tool's conception or design" (LeBlanc 2013: 10).

More elaborate mixed-method designs were also used to gain a deeper understanding of translators' needs, particularly within the framework of PhD theses work. For example, Bundgaard et al. (2016) collected data concerning the revision of MT-assisted TM translation with a group of in-house translators at Text-Minded Danmark. *Micro-level translation processes* were investigated through an experiment with eight in-house translators using SDL Trados Studio 2011 integrated with a client-specific MT engine. (using screen capture, keystroke logging, observation, retrospective interviews, and a post-experimental question-

naire) and *macro-level translation processes* were studied primarily through ethnographic methods, namely participant observation, semi-structured interviews and document collection (Bundgaard et al. 2016: 111). Another example is the doctoral dissertation of Zaretskaya (2017), which investigated translators' needs by means of a user survey, the evaluation of existing CAT systems, and the analysis of the process of post-editing of machine translation.

Differently from these studies, Krüger (2016a, 2019) developed a model of CAT tool usability starting from usability concepts rather than from data. Krüger defined CAT tool usability based on ISO's (2016) usability definition and added elements from ISO (2011) and the *Quality in Use Model* developed by this standard, to define CAT tool usability criteria. He considers CAT tool usability to be dependent on the context of use (and the context coverage of the tool) and defined by the *effectiveness* and *efficiency* of use as well as translators' satisfaction when using the tool to complete translation tasks. He also defines CAT usability by users' perceived learnability. Finally, also the tool's *data security* is included as a usability dimension because of the pressing issue of data security in the translation industry, which, in Krüger's view, restricts the context coverage (Krüger 2019: 113). To my knowledge, however, this model has not yet been used as an analysis instrument in empirical explorations.

### 3.4.2 Evaluation research

CAT tools have been evaluated empirically from different perspectives. In his review, Doherty (2019) distinguishes between *product-oriented*, or *linguistic*, evaluation and *process-oriented*, or *performance-based*, evaluation. Studies in the first category are concerned with the evaluation of the output of translation technology, for instance, the output of MT or TM. Studies in the second category focus on the translation product and process using a variety of methods and metrics. Läubli & Green's (2019) review focuses on evaluation methods testing tools on human translators and distinguishes between the evaluation of *translation speed*, *translation quality*, and *user experience*. In the review below, I will divide studies into the categories (1) evaluation of tool performance, (2) evaluation of users' performance, and (3) evaluation of users' perception.

#### 3.4.2.1 Tool performance

The evaluation of CAT tool performance has been mostly concerned with some quality aspects of the translation aids using several methods and metrics. Other possible evaluation focuses include cost-effectiveness, ease of implementation

and maintenance, and considerations of training (Whyman & Somers 1999). The evaluation of translation aids has typically been based on the measures of accuracy, precision and recall (Whyman & Somers 1999).

The evaluation of TMs attempted to evaluate the "usability" of translation aids without the involvement of human translators. Whyman & Somers (1999), for instance, report searching for a parameter capable of being used as an optimisation criterion. In their study, they evaluated the accuracy, precision and recall of TM matching and proposed a weighting factor in terms of keystrokes needed to change the proposed target segment into the desired text. Another example is provided by Colominas (2008), who evaluated the accuracy and recall of TM segmentation at the sub-sentential level.

The evaluation of the accuracy of MT output has been a major concern in the industry. Through time, it has shifted from human annotation to automatic evaluation, which was intended to be more objective, consistent, faster and cheaper than human evaluation (Doherty 2019). Human evaluation was typically conducted by asking human raters to express a judgment on Likert-scale items (Doherty 2019: 340). Machine-based evaluation measures are called *automatic evaluation metrics* (AEMs). Their purpose was to "measure the similarity or difference between the output of an MT system and a reference translation or gold standard, typically a human translation, on a set of predefined linguistic criteria" (Doherty 2019: 344). Doherty (2019: 344–345) explains that AEMs originated from speech recognition research and the metric of word error rate (WER) which was adapted into translation error rate (TER) and human-targeted TER. Other AEMs, such as BLUE (Papineni et al. 2002), gained substantial popularity. More complex AEMs later emerged to outperform BLUE in their correlations with human evaluation as well as the complexity of linguistic features they can cover (Doherty 2019: 345).

#### 3.4.2.2 User's performance

Data concerning users' performance has also been used to evaluate several usability aspects of CAT tools. The evaluation has been directed both to the translation product and the translation process. One aspect of translators' performance that has been evaluated to gain insight into the effectiveness/efficiency of CAT tools is the *translation speed*, typically measured as words per hour, seconds per word, seconds per segment (Läubli & Green 2019). This is considered a critical measure due to its direct economic impact (Läubli & Green 2019: 375). Time spent on a translation unit may also be interpreted as a sign of mental effort and measured, for instance, as the number of keystrokes and mouse clicks used to produce a

target text or segment and *keystroke ratio*, i.e. ratio of keystrokes to the number of characters in the final text or segment (cf. Koponen et al. 2012). Läubli and Green explain how these measures may be used to evaluate the clarity or efficiency of the UI or interactive TM (ITM): "In the context of IMT, typing effort is an interesting metric to gauge how often translators make use of suggested translations, e.g., by accepting the completion of a word instead of typing it out" (Läubli & Green 2019: 375). A higher-than-expected keystroke ratio may hence be interpreted as a sign for a lack of visibility of the translation aids, which hinders their identification and incorporation into the translation by the user, leading the translator to type the target text rather than simply accept the aid.

Another important aspect in the evaluation of users' performance is the *quality* of translators' output. This aspect is essential to consider because "it can offset gains in speed and usability. Even when a sentence completion feature allows for a 30% increase in translation speed, if the feature leads users to produce worse translations, then the finding is less (or not at all) meaningful" (Läubli & Green 2019: 375). (Läubli & Green 2019: 341–342) explains that the evaluation has typically revolved around the measures of accuracy and fluency to then expand to the dimensions of readability, comprehensibility, and acceptability. Paraphrasing Läubli & Green (2019: 341–342), *accuracy* (also called adequacy or fidelity) pertains to the extent to which the translation unit carries the meaning of the source into the target. *Fluency* (also called intelligibility) focuses on the extent to which the translation complies with the rules and norms of the target language. *Readability* relates to the extent to which a defined segment of text can be read by a specified reader. *Comprehensibility* (Closely related to the theoretical construct and measurement of readability) measures to what extent a reader has understood and retained the information contained in a text. *Acceptability* refers to the extent to which a system and its output meet users' expectations, essentially as expressed in usability research (e.g. Nielsen 2010). Hence through time, the evaluation of translators' performance has extended from the linguistic level to a broader perspective of the user of translators as a means to evaluate the usability of CAT tools. I call this perspective communicative because it considers translation as an activity of communication with the recipient of translation.

While the evaluation methods above pertain to the translation products, other methods have been used to tap into the translation process and identify problems in translators' interaction with CAT tools (cf. O'Brien 2005). Examples include the use of TPR instruments such as TransLog, eye trackers, think-aloud protocols (TAP) and cued retrospective interviews.

The evaluation of translators' performance when using CAT tools essentially depends on two key variables: the CAT tool and the text to be translated. These

may be considered as the two core test materials that may require manipulation to varying extents based on the study aims. Generally, exploratory studies aimed at a holistic evaluation of CAT tool use in real practice call for naturalistic study materials. On the contrary, studies aimed at a more fine-grained evaluation of the UI to develop design recommendations require some degree of manipulation of the study materials to control for influencing variables and ensure that the conditions for meaningful observations are in place. One control measure is the design of the text to be translated during the test. For example, for their study on SDL TRADOS' AutoSuggest function, O'Brien et al. (2010) designed a semi-technical German text of 424 words in 25 segments from the domain of business to be translated into English by subjects. In their study of the optimal format of text presentation for translation and revision, Läubli (2020) inserted errors into human translations that are unambiguously wrong to measure whether and how quickly study participants could correct these errors within the different UIs. The UI design is another study material that may need to be manipulated to control for influencing variables and zoom in on precise UI design features. For instance, Läubli (2020) aimed to gain insight into the impact of text segmentation and alignment on users' performance in translation and revision tasks. To achieve this aim, they presented test participants with different UIs and measured their performance in terms of speed and accuracy.

#### 3.4.2.3 User's perception

Users' perception of CAT tools is a further source of data commonly used to evaluate the usability of CAT tools. Most commonly, this has been gathered through post-task questionnaires. Authors have used both self-designed questionnaires and previously-existing usability questionnaires, such as the Systems Usability Scale (Coppers et al. 2018) and the Software Usability Measurement Inventory (Vargas-Sierra 2019). In the development and evaluation of Lilt, Green et al. (2014) asked translators to rank the usefulness of translation aids from the most to the least helpful with the aim to identify the preferred features.

## 3.5 Discussion: Usability engineering in CAT

This chapter contextualised the development of CAT tools within translation technology and discussed the research informing the development of such tools. As emerges from this review, the development of CAT tools was driven by market needs for greater translation volumes at reduced turnaround times and costs

rather than translators' needs. The basic features of CAT tools were mostly defined by ICT experts and implemented in the first commercially available CAT tools without preliminary research on translators' needs to define requirements and UI design principles. Since then, these first tools have become the industry standard, and the core UI features have remained relatively stable. Although new tools keep emerging on the market, most often is the agency or the client to decide which CAT tool is to be used, which implies that translators are often forced to choose a tool that they find too expensive and less attractive than another one out of the sheer practical need of accessing jobs (cf. Garcia 2015). Despite the concern that "if tool settings and features do not align with translators' ways of working, then their flow can be interrupted, their cognitive load increased, and their efficacy compromised" (Kappus & Ehrensberger-Dow 2020: 2), the design of CAT tools has been rather "non-user centric" (Moorkens & O'Brien 2017). Indeed, translators' needs began to be systematically examined only about a decade after the commercial release of CAT tools and focussed empirical investigations of the impact of fundamental UI features on the translator's work are still scarce (cf. Läubli 2020).

Although established market realities are difficult to change, virtuous examples of translator-centred CAT tool development do exist and usability-oriented CAT research can draw on a plethora of methods and a wealth of interdisciplinary experience. Given the commercial interest in leveraging technology to increase the efficiency and quality of translation, CAT tools have been the object of a multitude of studies conducted not just by TS scholars but also by HCI experts.

As discussed in the review, the usability of CAT tools has been explored by means of tool performance, users' performance, and users' perception. The studies explicitly aimed at informing CAT tool design are usually characterised by some degree of experimental control, especially in the manipulation of test materials – primarily, the text to be translated, the task (e.g. translating, revising etc.) and the UI that users have to work with during the test. Studies within this line have produced the most fine-grained analyses of the impact of specific UI features on translators' performance (e.g. Läubli 2020). This line of research has the potential not just to advance scientific understanding of translator-CAI tool interaction and the variables inherent to the UI that influence such interaction. While studies approaching CAT tools with this level of detail remain scarce, they are contributing to developing generalisable principles for CAT tools' UI design that can ensure that CAT tools are more adequate to translators' needs. In doing so, they have the potential to pay great service to the profession. As

pointed out by Läubli (2020: 1) "since the motivation for using CAT tools is primarily economic – saving time by leveraging translation suggestions rather than translating from scratch – the design of these tools is unlikely to change until measurements show that alternative designs speed translators up or cause them to make fewer mistakes". This reveals the instrumental function of usability research in the development of translation technology – to increase the inclusion of translators' views in the development of technology.

Since already mature systems are difficult to change for economic reasons, experts recommend involving translators early on in the design of CAT solutions, advancing the development through iterations of prototyping, testing and revision and incorporating data into the whole development process (cf. Läubli & Green 2019), as prescribed by the usability engineering approach (see Chapter 2). Even in the case of already mature systems, research can provide compelling arguments for change. As pointed out by Läubli (2020: 1) "since the motivation for using CAT tools is primarily economic – saving time by leveraging translation suggestions rather than translating from scratch – the design of these tools is unlikely to change until measurements show that alternative designs speed translators up or cause them to make fewer mistakes". This reflection exemplifies the role of research not just to advance scientific understanding but also to drive change. By establishing the heuristic principles of CAT tool design that make UI interfaces maximally efficient and effective, usability-oriented CAT research can contribute to the development of instruments that are adequate to translators' needs.

# 4 Interpreting technology

This chapter analyses interpreting technology with a focus on computer-assisted interpreting (CAI) tools and related research from a usability engineering perspective (discussed in Chapter 2). The analysis offered in this chapter positions the present work within previously conducted research, identifies the knowledge gap, and sheds light on the relevance of usability-focused CAI research. After a definition of "interpreting", this chapter examines the history of interpreting technology and the development of computer-assisted interpreting tools. It then provides a succinct review of CAI tool research, with a focus on studies dedicated to informing the design of CAI tools through the analysis of users' needs and requirements as well as through empirical testing. The chapter concludes with a discussion of the extent to which a usability engineering approach is reflected in the development of CAI tools and the relevance of usability research for the future developments of the profession.

## 4.1 Developments and current landscape

While translation was defined as a written, textual operation (see Chapter 3), interpreting may be defined as a translational activity characterised by its immediacy: "in principle, interpreting is performed "here and now" for the benefit of people who want to engage in communication across barriers of language and culture" (Pöchhacker 2004: 10). It is a form of oral translation enabling direct communication between two groups, especially where a lingua franca does not exist. Interpreting in all its forms is studied in the academic discipline of *Interpreting Studies* (IS) which emerged as a subdiscipline of translation studies around the 1990s and since then has gained an increasingly distinct identity (Pöchhacker 2015a).

There are several modes of interpreting. Liaison interpreting (also dialogue interpreting) is performed in a variety of business, community and diplomatic encounters in which the interpreter is physically and/or metaphorically "the person in the middle" of the communication and performs a connecting function (Merlini 2015). Consecutive interpreting is mainly performed in conference settings

and involves the rendition of a whole speech from the source language (SL) into the target language (TL) *after* it has been fully uttered by the speaker (Andres 2015). Simultaneous interpreting (SI) is the rendition of the message from SL to TL in real-time, *while* it is being uttered by the speaker and typically takes place in conference and court settings (Pöchhacker 2015b). It can be further differentiated based on the supporting equipment being used (Pöchhacker 2015b) which distinguishes simultaneous interpreting (performed through special equipment) from its precursor form of *whispered interpreting* (or *chuchotage*, in which the interpreter whispered the interpretation into the ears of the audience) and from the more recent *remote simultaneous interpreting* modality (also called distance interpreting and defined as "interpreting of a speaker in a different location from that of the interpreter, enabled by information and communications technology (ict)").

It is often said that interpreting is the second oldest profession in human history (Baigorri-Jalón 2014: 10). Interpreting has existed as long as there have been different spoken languages in contact, long before written language and hence written translation. Written accounts of the activity of the interpreter through history are scarce but they may be found as early as 3000 BC in ancient Egypt, where the activity of interpreting was first institutionalised and designated by its own hieroglyphic (cf. Falbo 2016, Kurz 1985). Interpreters are mentioned several times in the Bible, for instance in Corinthians (14:28): "If any man speak in an unknown tongue, let it be by two, or at the most by three, and that by course; and let one interpret. But if there be no interpreter, let him keep silence".

Despite such a longstanding tradition, conference interpreting began to emerge as a profession in Europe in the 20th century (Baigorri-Jalón 2014). The first conference interpreters came to work in government departments after World War I as well as in early international organisations. At the time, the pluralism of languages began to characterise the world of diplomacy, which had previously been monolingual, with Latin in medieval and renaissance western Europe and French in the early 20th century being used by diplomats of different origins as lingua franca. During high-level international meetings, the early conference interpreters enabled participants speaking different languages to communicate with each other by turning their messages (from short statements to whole speeches) from one language into another in the consecutive mode.

Since then, the rapid evolution of the conference interpreting profession has been stimulated by historical events, the emergence of new political structures (especially international organisations), new societal needs and it has been propelled by technological advancement. Scholars define the impact of technology

on the interpreting profession as a series of "breakthroughs" (Fantinuoli 2018b, Tripepi Winteringham 2010).

The first breakthrough in the development of conference interpreting was marked by the birth of simultaneous interpreting. In the 1920s IBM patented a wired system for real-time speech transmission which was adopted by the Sixth Congress of the Comintern in the former Soviet Union, then at the International Labour Conference and irreversibly made popular by the Nuremberg trials (Fantinuoli 2018b). The advantage of this new system was that it curbed conference times and made it possible to host a conference in multiple languages at the same time. SI grew at a rapid speed despite the resistance of early conference interpreters, who feared that being relegated at the back of the room might diminish their prestige (Fantinuoli 2018b). Since then, SI became the standard interpreting mode employed at international meetings and organisations such as the United Nations and the European Union (EU) – the biggest employer of conference interpreters in the world (European Commission, Directorate-General for Interpretation 2018).

The second breakthrough was introduced by the World Wide Web, which changed interpreters' access to knowledge. To comprehend how the internet changed the profession, one must consider the crucial importance of preparation for interpreters:

> Interpreters are knowledge workers constantly moving back and forth between different contexts, languages and conceptual systems. To do this, they rely on their own knowledge base being complemented by external information sources explored in what can be called "secondary knowledge work" in order to properly perform the actual, "primary" knowledge work, i.e. the activity of interpreting work as such ... In-depth preparation is therefore all the more important. It involves specific terms, semantic background knowledge and context knowledge about the assignment. (Rütten 2016)

While prior to the internet interpreters had to acquire terminology and specialised knowledge for each assignment through paper-based sources, the internet made it possible to access a wealth of information at any time. According to some interpreters and scholars, the internet increased the efficiency of interpreters' preparation: "The greatest advantage of new information technology lies in the speedier collection of information from different sources and easier management so that preparation is more efficient and results in improved quality" (Kalina 2010: 81). Others, point to both pros and cons in the "surfeit of data" (Donovan 2006) with which interpreters were suddenly confronted in their

preparation. The risk for interpreters is to be carried away by the information overload in their preparation.

According to Fantinuoli (2018b), recent technological developments brought about a third breakthrough in conference interpreting with even greater transformative potential than the previous two: "Bigger by one order of magnitude if compared to the first two breakthroughs, its pervasiveness and the changes that it may bring about could reach a level that has the potential to radically change the profession" (Fantinuoli 2018b: 3). Given the breadth and depth of these changes, he suggests calling the recent third breakthrough in conference interpreting the *technological turn*.

Today, the interpreting technology landscape presents a plethora of technological solutions that are involved in the technological turn in the profession. Braun (2019) distinguishes four categories of interpreting technologies. The first comprises all technologies used to deliver interpreting services and enhance their reach, such as RSI equipment, giving rise to *technology-mediated interpreting*. The second category includes all technologies (both generic and bespoke for interpreters) that can be applied to support or enhance the interpreter's preparation, performance, and workflow, leading to *technology-supported interpreting*. The third category comprises all technologies that are designed to replace human interpreters, leading to *technology-generated* or *machine interpreting* (MI). Machine interpreting (also known as automatic speech translation, automatic interpreting or speech-to-speech translation) may be defined as "the technology that allows the translation of spoken texts from one language to another by means of a computer program" (Fantinuoli 2018b: 5). An audible version of the SL speech into a TL speech is generated combining automatic speech recognition (ASR), to transcribe the oral speech into written text, machine translation (MT), and speech-to-text synthesis (STT, Fantinuoli 2018b: 5). A fourth category of interpreting technologies in Braun's classification comprises all forms of *technology-enabled hybrid modalities*, such as respeaking.

Technologies for technology-mediated and technology-supported interpreting, in Braun's classification, correspond to setting-oriented and process-oriented technologies in Fantinuoli's (2018a) classification, which further elaborates on the impact that these technologies exert on interpreting. Setting-oriented technologies "primarily influence the external conditions in which interpreting is performed" (Fantinuoli 2018a: 155). In Fantinuoli's view, these have an impact on the interpreter profession, its status, and working conditions, but do not radically change the nature of the mental task performed by interpreters. By contrast, *process-oriented technologies* "support the interpreter during the different

sub-processes of interpreting and, consequently, in the various phases of an assignment, i.e. prior to, during and possibly after the interpreting activity proper" (Fantinuoli 2018a: 155). In Fantinuoli's view, by becoming an integral part of the interpreting process, these technologies are directly linked to and might influence the cognitive processes underlying the task of interpreting. The central process-oriented technologies giving rise to technology-supported interpreting are called *computer-assisted interpreting* (CAI) tools.

## 4.2 Computer-assisted interpreting (CAI) tools

CAI tools may be defined as "all sorts of computer programs specifically designed and developed to assist interpreters in at least one of the different sub-processes of interpreting" (Fantinuoli 2018a: 155). Better specified, the aim of CAI tools is to "improve the interpreters' work experience, by relieving them of the burden of some of the most time-consuming tasks ... both during preparation and during the very act of interpreting" (Fantinuoli 2018b: 4). By doing so CAI tools ultimately aim to increase the interpreters' productivity as well as the quality of his/her performance.

This definition differentiates CAI tools from general software solutions that were not developed for interpreters, but that can be used by interpreters in their works, for instance, word or Excel files used to create glossaries. It also specifies that CAI tools aim to *support* interpreters, not to replace them contrary to MI technology in one or more sub-processes of interpreting or phases of the interpreter's workflow.

For the sake of simplification, these phases can be defined as preparation (*pre-booth*), interpretation (*in-booth*) and follow-up (*post-booth*, EABM 2021a). This definition is based on the breakdown of the interpreting workflow as *pre-*, *peri-*, *in-* and post-process (Kalina 2000). While the follow-up phase is essentially about revising and expanding the interpreter's knowledge based on the information gathered during the assignment, the two most important phases which have been central to the development of CAI tools are the pre-booth and the in-booth phases.

In the preparation or pre-booth phase, CAI tools aim to increase the efficiency of interpreters' preparation for their assignments. Interpreters' preparation generally consists of reading materials provided by the speaker and the conference organisers to gain language knowledge (i.e. terminology), content knowledge (i.e. information about the topic that will be interpreted), and context knowledge (i.e. information about the communicative setting in which interpreting takes place;

(Rütten 2007). If preparation material is not provided or insufficient, interpreters have to find relevant material themselves. In this phase of the interpreter's workflow, CAI tools can help interpreters create glossaries, manage them, and share them with colleagues. The extent to which the glossary creation process is supported essentially depends on the degree of sophistication of the CAI tool, as explained below.

In the interpretation or in-booth phase, CAI tools aim to help interpreters access information while interpreting, easing the burden related to the interpreting task and ultimately increasing delivery accuracy. While interpreters perform SI, they retrieve specialised terminology found in the preparation phase from their memory, which may imply a certain mental effort if the terminology is not so consolidated for the retrieval to be automatic. When in doubt, they search for the TL equivalent term in their glossary. If the term is not present in their glossary, they search for it in online databanks, dictionaries etc. In general, these processes are considered to be "time-consuming and distracting in an activity that requires concentration and fast-paced decoding and delivery. The interpreter at work may not have the time or the cognitive ability to look up a word online or in his/her electronic dictionary, or detect and choose the correct translation of a specific term among the myriad of possible solutions that are generally offered by dictionaries" (Tripepi Winteringham 2010: 90). During SI, interpreters have to deal with not just specialised terminology but also other factors of complexity, known as "problem triggers" (Gile 2009). These elements of speech, such as acronyms, named entities, numbers and specialised terms, are generally considered to be highly problematic and are associated with higher-than-normal error rates. Support for problem triggers is sometimes provided by the booth colleague, who may look up an unknown term or jot a numeral down. However, this support is not always reliable and efficient:

> Even the help of the fellow colleague in the booth may sometimes prove useless in real time oral translation, and may even slow down the interpreting process. The interpreter, when hearing something unknown, is often alone and has nothing to resort to but his/her own memory and mind (ibid: 77). A simultaneous interpreter at work cannot wait for more than half a second for a missing word otherwise his/her narrative would sound broken and the short memory be overburdened ... The activity of searching for the right term may result in distraction and loss of concentration for the interpreter. (Tripepi Winteringham 2010: 91)

Depending on their architecture and functionalities, CAI tools have been traditionally divided into *generations* (EABM 2021a, Fantinuoli 2018a, Prandi 2022,

2023). *First-generation CAI tools* are defined as "programs designed to manage terminology in an interpreter-friendly manner" (Fantinuoli 2018a: 164). They were designed to support interpreters in the pre-booth phase of preparation. They developed from the concept of terminology management systems for terminologists and translators but present a simpler entry structure which is more adequate to the interpreter's terminology work. They offer basic functionalities to look up glossaries in the booth which are similar to the look up system in a word or Excel file. To query the glossary, the interpreter types the term or part of the term in the search field and presses the enter key.

*Second-generation* tools offered "a holistic approach to terminology and knowledge for interpreting tasks and ... advanced functionalities that go beyond basic terminology management, such as features to organise textual material, retrieve information from corpora or other resources (both online and offline), learn conceptualised domains, etc" (Fantinuoli 2018a: 165). Other than such advanced functionalities for knowledge acquisition and management, the second generation of tools started bespoke functionalities for the in-booth phase. Advanced search algorithms were introduced to ease glossary query during SI, i.e. taking into account the time constraints and the complexity of the interpreting task. For example, *fuzzy search* acts as an error correction mechanism for misspelling in the word typed by the interpreter or present in the glossary, *stopwords exclusion* reduces the number of matches displayed as a query result, *dynamic search* finds terms without interpreters having to press the enter button, and *progressive search* searches for the term in other glossaries of the interpreter if it is not found in the glossary that is currently being queried (Fantinuoli 2018a: 166).

The introduction of artificial intelligence (AI) into CAI tools paved the way for a *third generation* (EABM 2021a, Prandi 2022). First, this technology has made it possible to automate parts of the preparation task, such as by automatically compiling glossaries starting from interpreters' preparation material. Second, the combined use of AI and automatic speech recognition (ASR) technology has made it possible to automate the query of interpreters' glossaries and provide support for all major classes of problem triggers (c.f. Fantinuoli 2017). Not only are the TL-equivalent of specialised terms automatically extracted from interpreters' glossaries, but also acronyms, named entities and numbers may be displayed in real-time on interpreters' laptop screens. Given the novelty of these tools, they currently present technological limitations. The major problems are the system's *latency*, which is the delay between the uttered word and its appearance on the interpreter's screen, and its *accuracy*, i.e. the possibility that the tool might produce errors in the displayed items or omit some information. Current tools are based on a *cascaded system* of independent ASR and AI modules, which

first recognise and transcribe the source speech, then extract problem triggers and finally convert them into a graphic representation that is displayed to the interpreter. This increases both system latency and the risk of inaccuracies due to the propagation of errors from one module to the next. To overcome these limitations, developers are currently working on an end-to-end system which, they suggest, may dramatically increase the potential of CAI tools and introduce a *fourth generation* of tools (Gaido et al. 2021).

## 4.3 CAI tool design, development and reception

The development of CAI tools has been initiated by conference interpreters with programming skills. in some cases, or by conference interpreters in tandem with developers, in other cases. A comprehensive review of past and present CAI tools is offered by Stoll (2009), Costa et al. (2014), Will (2015), and Fantinuoli (2016) among others.

At the time of writing, the commercially available and actively maintained first-generation CAI tools are *Interplex*[1], *Interpreters' Help*[2], and *InterpretBank*[3] (which was later developed into a second-generation and then a third-generation CAI tool). Two CAI tools, InterpretBank and *LookUp*[4], were developed within doctoral research projects, but only InterpretBank has made it to the commercial stage. Interplex is a glossary builder and management tool for both translators and interpreters first released in 2003. It was designed by conference interpreter Peter Sand and developed by software developer Eric Hartner. After the first release, the software has been "tweaked and re-tweaked" by its creators, "taking on board the many suggestions received from interpreters using it in the booth" (Sand 2015). Interpreters' Help is a cloud-based glossary builder and management tool for conference interpreters first released in 2013. It was designed and developed by software developers Yann Plancqueel and Benoît Werner with the consultancy of two conference interpreters who are part of the team. InterpretBank was first developed as a first-generation CAI tool and then became the first second-generation CAI tool. Its functions were glossary building, management, and look up during simultaneous interpreting. It was designed and developed between 2008 and 2012 by conference interpreter and scholar Claudio Fantinuoli as part of his doctoral research project at the University of Mainz/Germersheim

[1] https://interplex.com/
[2] https://interpretershelp.com/
[3] https://www.interpretbank.com/site/
[4] http://www.lookup-web.de/index.php

(Fantinuoli 2016). In its subsequent development, it was the first tool to introduced advanced search functions for in-booth use, which makes it particularly popular among interpreters. The starting point for the design was a review of the literature modelling interpreters' workflow and analysing interpreters' needs, as presented in the previous section.

Currently, the ASR- and AI-powered "third-generation" CAI tools on the market are *InterpretBank ASR* and *Kudo's InterpreterAssist*[5] – as well as SmarTerp (presented in Chapter 5). InterpretBank ASR (Fantinuoli 2017) is the first-ever third-generation CAI tool. It is the eighth release of InterpretBank, in which ASR and AI technology was introduced to retrieve terms from interpreters' glossaries and automatically display numerals as a cloud-based experimental feature. At present, users can choose between two UIs of InterpretBank ASR. In the first, numbers only are provided in the form of a scrolling list with new numerals placed on top of the list and highlighted (see Figure 4.1).

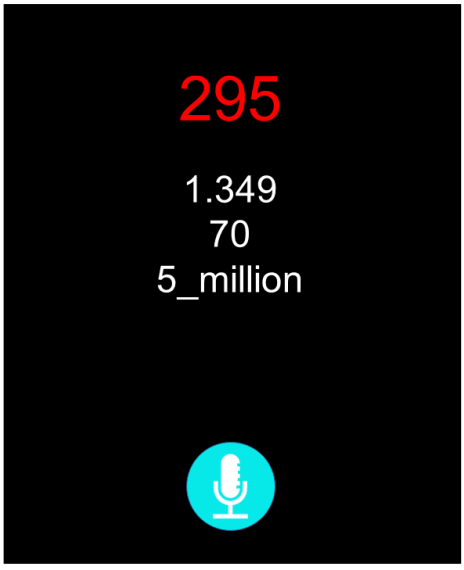


Figure 4.1: Representation of InterpretBankASR (numbers only view)

In the second UI, specialised terms and numerals are provided in two distinct columns. New items appear at the top of the scrolling list and are highlighted (see Figure 4.2). In both visualisations, numerals are displayed in the source language.

[5]https://kudoway.com/kudo-interpreter-assist-artificial-intelligence/

A research paper (Defrancq & Fantinuoli 2021: 5) describes a third possible interface option for InterpretBank ASR, which displays the whole source-language transcript and highlighted units of interest – numbers and terms (see Figure 4.3).

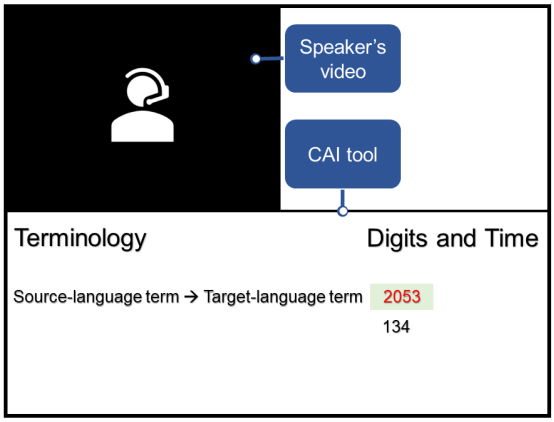


Figure 4.2: Representation of InterpretBank ASR (numbers and terms view)

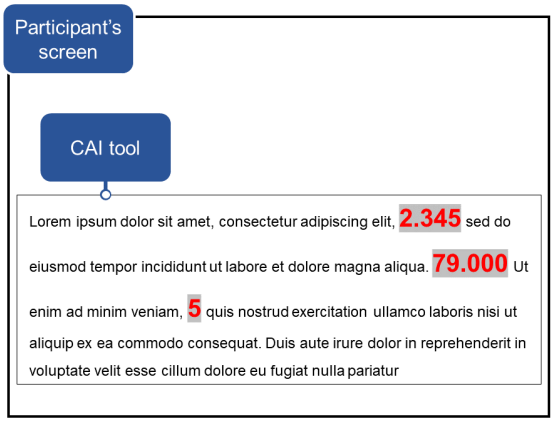


Figure 4.3: InterpretBank ASR transcript interface, from Defrancq & Fantinuoli (2021: 7)

Kudo InterpreterAssist (Fantinuoli et al. 2022a) is a recent CAI tool developed in collaboration with Claudio Fantinuoli to be integrated into Kudo's RSI platform. Currently, no information is available on the UI of this CAI tool.

Compared to CAT tools, CAI tools have been used to a much smaller extent (Tripepi Winteringham 2010: 89). One factor that has certainly contributed to

this more modest spread is the fact that CAI tools, so far, have been meant to be used by individual interpreters based on personal interest. CAT tools, on the contrary, have been adopted by agencies and institutions, which forced the adoption by translators. However, the voluntary nature of CAI tool use may change in the next year, with their growing integration into RSI platforms. Organisations have also shown increasing interest in adopting CAI tools to streamline the management of terminology and preparation materials for interpreters' assignments. CAI tools have been received by practitioners with mixed feelings:

> These opposite views are nothing new: whenever change comes, there are those who feel delighted with it and those whose emotions range from fear or panic to anger and aggression. In the case of ICTs we could say that the delight would come because they can be seen as tools that can make the work of the conference interpreter much easier, in any of the phases of the interpreting, while the fear or anger may come from hurt pride, as many interpreters feel it a part of their self-image as professionals to be able to manage without anything except their extraordinary memories. (Berber-Irabien 2010: 27)

A further reason for the limited reception of CAI tools by the interpreting profession has to do with concerns of different nature. The first concern is linked to the complexity of the simultaneous interpreting task. The use of ICTs in the interpreting process has been defined as "unnatural" (e.g. Donovan 2006). Professionals have maintained that the use of a new technological instrument in the booth may exacerbate the cognitive load in SI (cf. Prandi 2017), disrupt the task and lead to a deterioration of performance rather than quality improvement. Jones (2014), cited in Fantinuoli (2018b), for example, spoke of "alienation due to the use of new technology" and contended that "[using ICTs in the booth] can lead the interpreter to lose sight of the first aim of interpreting as we learn it, namely conveying meaning and facilitating communication". Another concern that emerges in the dialogue with practitioners is of economic nature. Interpreters often raise the question of whether the use of new technologies increasing the efficiency of interpreters' preparation may lead clients to demand a reduction in the price of interpreting services, similar to what happened with the introduction of CAT tools. Finally, a concern that is frequently expressed by professionals may be caused by the lack of knowledge of CAI tools. The frequent question posed by interpreters during public discussions "aren't these tools aimed at replacing us?" shows that they are not fully aware of the difference between technologies aimed at replacing human interpreters and supporting them.

The way in which CAI tools have originally been developed may also be a factor contributing to their limited acceptance by professionals. While the initial development of some CAI tools (in particular, InterpretBank, which is the most used CAI tool, and LookUp) has drawn on scholarly work on interpreters' workflow and preparation needs, the UI design has been "more or less based on the personal ideas of their developers, mostly interpreters themselves, and lack[s] any experimental support in [the] design decisions" (Fantinuoli 2018a: 160). Hence, CAI tool UI design has been predicated upon developers' personal intuition and rather vague heuristics. For example, Fantinuoli (2017: 62) defines a "simple" and "distraction-free" UI as a requirement for an effective CAI tool. Defrancq and Fantinuoli postulate that, in order to effectively support interpreters, the in-booth CAI tools should present the visual input "in an ergonomically suitable format" (Defrancq & Fantinuoli 2021: 4). However, the exact features and characteristics that make a CAI-tool interface "simple" and "ergonomic" are not specified. Furthermore, Fantinuoli (2018a: 164) points out that the development of CAI tools has been based on partial and unsystematic knowledge and has lacked "test and improvement phases, which are crucial to achieve software maturity and find the golden standard in terms of functionalities desired by the professional group."

## 4.4 CAI tool research

CAI tool research may be positioned within the IS sub-field of interpreting technology research, alongside research on remote interpreting and the use of technology in interpreter training. The academic interest in interpreting technology remained scarce until the first decade of the 21st century (Berber-Irabien 2010). Until recent years, research on CAI tools remained underrepresented in the interpreting technology research landscape (Fantinuoli 2018a). It is noticeable that, with the exception of the study by Desmet et al. (2018) and Defrancq & Fantinuoli (2021), most empirical studies on the use of CAI tools in the booth were conducted within the framework of Master's degree theses and PhD theses, in the case of Prandi. Prandi (2022: 87) interprets this as a sign of the increasing interest in CAI tools among younger generations, but it could also be interpreted as a sign of the marginalised role that (in-booth) CAI tools have played in academic research until recent times. Over the past couple of years, the spread of RSI propelled by the Covid-19 pandemic and the introduction of ASR technology into CAI tools has renewed the interest in different ICTs in interpreting and in the in-booth CAI tool use, in particular.

CAI tool research is a nascent research strand and therefore no specific definition has yet been provided. Prandi (2022: 85) points out that "the publications devoted to CAI tools address a number of topics and issues inherent to the tools, ranging from a simple presentation and comparison of the available solutions to the first empirical tests conducted on one or more tools". Since CAI tool research is a relatively new field of inquiry that has been much less prolific than CAT tool research, the research landscape is still largely lacking clear research strands and agendas. We could broadly define CAI tool research as a research strand that is concerned with the design and development of CAI tools and their impact on the interpreting product and process.

The review below addresses previous research on CAI tools that has implications for their design. As in the review of CAT research (cf. Chapter 3), the focus will be on the most common methods used to develop recommendations for the design and their further improvement. I will first consider research providing input on users' needs and general CAI tool requirements. Then, I will examine research focused on the empirical evaluation of CAI tools via tool performance, users' performance, and users' perception.

### 4.4.1 Need analysis and tool requirements

As discussed earlier in this chapter, the initial "first-generation" CAI tools aimed to support interpreters in the preparation of their assignments. Some scholarly work preceded the development of the first CAI tools in the analysis of interpreters' needs in preparation and identification of the pain points that could be addressed by a technological solution. The final aim, which is more or less explicit in the different studies, is to develop a set of requirements for the development of CAI tools.

The first studies preceded and accompanied the development of first-generation CAI tools and focused on interpreters' needs inherent to preparation and terminology management. Because at the time of the first studies the term "CAI tools" was not yet in use, they refer to their object of inquiry with terms such as "interpreting-specific technologies", "electronic tools for interpreters" etc. An important theoretical input to the development of CAI tools in this early stage came from authors such as Gile (1987), Kalina (2000, 2007), and Will (2009), who defined the workflow of interpreters and modelled the processes inherent to preparation (cf. also Kalina 2015). Fantinuoli (2006) proposed a model of Corpus Driven Interpreters Preparation and discussed how a CAI tool can facilitate this preparation process.

Parallel to the theoretical studies, some empirical studies explored interpreters' preparation and terminology management. For example, Moser-Mercer (1992) conducted a survey about how conference interpreters handle terminology documentation and document control to identify some guidelines for the development of CAI tools for terminology and documentation management. Another study that used survey methods to analyse interpreters' needs in the use of technology for terminology management was conducted by Bilgen (2009). The results further shed light on the difference between interpreters' and translators'/terminologists' terminological needs and the requirements that a terminology tool for interpreters should have (e.g. simple entry structure, flexibility in the search process etc.). The studies by Berber-Irabien (2010) and Corpas Pastor & Fern (2016) surveyed the use of ICTs by interpreters to portray the landscape of available solutions and identify possible gaps. A different design was chosen by Rütten (2007), who moved from a field study in which she observed interpreters in their terminology acquisition and management process to derive requirements for CAI tool development.

When it comes to need analysis related to the use of CAI tools in the booth, studies describing interpreters' in-booth behaviour (Chmiel 2008, Duflou 2016, Jensen 2006) may be considered to be precursor work to the design and development of CAI tools. Also research on the impact of "problem triggers" (cf. Gile 2009), such as research on the impact of numbers on SI (e.g. Braun & Clarici 1996, Frittella 2019a, Mazza 2001), is often reported as a justification for the need for a CAI tool in the booth to access terminology and the integration of ASR to support the rendition of numbers (for instance, see the description of research gap and purpose statement in Defrancq & Fantinuoli 2021, Desmet et al. 2018, Fantinuoli 2017).

An in-depth ethnographic study conducted by Rütten (2018) collected and analysed the booth notes of 25 conference interpreters at the European Commission and Council. The study provided insights into which items are usually written down by interpreters during SI, how these are written, how the sheet of paper on which items are noted functions as a "communication platform" to exchange information between colleagues (Rütten 2018: 143). As the author herself suggests, information on how interpreters use non-digital supports during SI could provide valuable input to the design of CAI tools. However, detailed investigations aimed at informing tool development for in-booth support have been scarcer than analyses of preparation needs.

A recent EU-funded project led by the University of Ghent in collaboration with the University of Mainz/Germersheim attempted to develop recommendations for the UI design of third-generation CAI tools based on a survey in which

they asked practising conference interpreters to express their preferences on a series of design options (EABM 2021b). Participants were asked to express a preference based on graphic representations of the interface options and written descriptions but did not use the options. Although the results of the survey should be treated as hypotheses rather than valid design specifications, this study testifies the growing interest in usability and UI design-focused CAI research.

Other authors attempted to define heuristic principles for the development of CAI tools. Costa et al. (2014) evaluated CAI tools for preparation by self-chosen features related to how glossaries can be compiled and presented. Will (2015) proposed some features that characterise a CAI tool as "simultaneous-interpreting compliant" (simultanfähig) based on theory and logical reasoning. Fantinuoli (2017) proposed technical requirements for an ASR system to be successfully integrated into CAI tools and for a CAI tool to effectively support ASR integration. These requirements were partly derived from commonly used metrics measuring the speed and accuracy of ASR systems and partly based on logical reasoning as well as his practical experience as a developer.

### 4.4.2 Evaluation research

Empirical research aimed at the evaluation of in-booth CAI tools has been conducted by interpreting studies scholars via the analysis of (1) *tool performance*, (2) *users' performance*, or (3) *users' perception*.

#### 4.4.2.1 Tool performance

Thus far, only two published studies, both written by Fantinuoli as sole or first author, reported on the evaluation of CAI tool performance. Such evaluation has focused on the accuracy of the visual aids displayed by interpreters.

Fantinuoli (2017) first proposed metrics for the evaluation of in-booth CAI tool performance and applied them to the evaluation of the first version of InterpretBank with ASR integration. He used the terminology-dense speeches designed by Prandi (2017), run the prototype and recorded the outcomes of term and number extraction. He used *word-error rate*, i.e. the percentage of wrongly displayed items out of all items that should have been displayed, as a measure of the tool's accuracy. He also uses the metrics of precision $p$ (i.e. the number of correct positive results divided by the number of all positive results) and the recall $r$ (i.e. the number of correct positive results divided by the number of positive results that should have been returned) to calculate the *F1 score* (i.e. the harmonic mean of precision and recall expressed as a value comprised between 0 and 1).

After the recent release of Kudo InterpreterAssist, Fantinuoli et al. (2022b) performed a technical evaluation of the tool's performance in the automatic generation of glossaries and the ASR-powered CAI tool. To evaluate the glossary-generation feature, they automatically generated three 100-term glossaries in English > French and English > Italian in three specialised domains. They first asked three conference interpreters to evaluate the relevance and accuracy of extracted terms categorising them as either "specialised term", "general term", or "error" for incomplete or lexically invalid elements. Then, they asked three conference interpreters to evaluate the English > French glossary and three conference interpreters to evaluate the English > Italian glossary by marking the target-language translation of extracted terms as either "acceptable" or "unacceptable". For the evaluation of the CAI tool output, they selected some speeches representative of the speeches interpreted on Kudo. They evaluated both the final output and the intermediate stage of automatic speech transcription in terms of precision, recall and F1.

The evaluation of CAI tool performance is still scarce and rather limited in scope. Other aspects, such as the stability of CAI tool latency (e.g. based on the speaker's accent and speaking pace, sound quality, word length etc.), have not been considered. A further limitation in previous evaluations of CAI tool performance is that the tool was tested in highly controlled conditions only (i.e. speeches characterised by high sound quality and standard native accents) whereas no evaluations have so far been conducted of tools in naturalistic conditions.

#### 4.4.2.2 Users' performance

The evaluation of the tool via users' performance is by far the most frequent method of CAI tool evaluation. Prandi's (2022: 89) research on "the impact of CAI tools usage on the quality of SI" is currently the most prolific type. Empirical research on the impact of CAI tools on users' performance began to emerge in the second decade of the 21st century and gained momentum with the integration of ASR into CAI tools about five years ago. As Prandi (2022: 87) notices, the evaluation of users' performance has mostly focused on the interpreting product.

The assessment of the impact of the tool on users' performance was mostly focused on the rendition of individual items (i.e. interpreted specialised terms or numerals) rather than larger units of analysis (e.g. the meaning of the interpreted sentence or speech passage) and more broadly conceived delivery quality, or even the perspective of the recipient of the interpreting service. For instance, in their study of the CAI tool-supported SI of numbers, Desmet et al. (2018) and

Pisani & Fantinuoli (2021) used the error classification proposed by Braun & Clarici (1996) which is concerned with the interpreted numeral only. Defrancq & Fantinuoli (2021) chose the numeral as their unit of analysis too. This means that all other components of the *numerical information unit*, such as the entity the numeral refers to (referent), its measure (unit of measurement), were left out from the evaluation, which makes the validity of the interpreted numeral a measure of CAI tool effectiveness questionable. In fact, contextualisation errors, i.e. where the numeral was correctly reported but wrongly contextualised in the speech, were incidentally reported in some studies (Canali 2019, Pisani & Fantinuoli 2021). Similarly, studies on the impact of in-booth CAI tool use on terminological accuracy considered the interpreted term as their unit of analysis. In sum, the rendition of isolated problem triggers has been the privileged measure of tool effectiveness. This is a major difference from CAT tool research, where a "communicative approach" was used to evaluate users' performance holistically, considering aspects such as text coherence and consistency, fluency, as well as its clarity and acceptability for recipients (cf. Chapter 3).

As of today, Prandi's (2022) doctoral dissertation may be regarded as the only study on the CAI tool-supported interpreting process. She compared the use of a "traditional" glossary with InterpretBank with manual-look up and an ASR simulation in the interpretation of three comparable speeches. Other than performance data (i.e. participants' deliveries), she gathered eye-tracking data to ascertain whether a variation in cognitive load could be identified in the three conditions. This may be considered as a measure of tool efficiency, i.e. a usability property of different technological solutions and CAI tool generations.

While previous empirical CAI tool research has not focused on UI design or tool development at a fine-grained level, a notable exception is represented by Montecchio's Master's degree thesis (Fantinuoli & Montecchio 2022, Montecchio 2021). This study evaluated participants' rendition of a number-dense speech with CAI tool support at a latency of 1, 2, 3, 4, and 5 seconds with the aim to ascertain what is the ideal and maximum acceptable latency. This study differs from the previous ones reviewed in that it was specifically designed to derive implications for CAI tool development. However, several methodological issues limit the reliability of the findings. First, the latency at which stimuli were presented was not randomised; the gradual increase may have caused a learning effect. Second, the numerals in the source speech were replaced by an acoustic signal ("beep" sound) to ensure that participants would rely on the visual stimulus; while this is an interesting approach, we cannot assume that the outcome would be equal when users are presented with multimodal input. Third, the numerals at each latency level were not comparable, as discussed below. While this

study remains interesting as an example of the growing interest in usability- and design-focused studies on CAI tools, the methodological limitations constrain the translation of findings into design principles.

Considering the design of test materials used in previous studies, like in empirical CAT research, the UI design of the CAI tool and the source speech are the test materials that require manipulation in order to gather data capable of responding to the research question. A further aspect to be added is CAI tool performance: since ASR-powered solutions are still in the prototype stage, their performance is not yet fully stable and may vary even when it is presented with the same speech. Common measures that have been taken to control for the variable of CAI tool performance is using mock-ups simulating peak performance of the CAI tool (Canali 2019, Desmet et al. 2018). Researchers have controlled the performance of second-generation CAI tools (i.e. InterpretBank with manual look-up as in Biagini 2015, Gacek 2015, Prandi 2015, 2022 by developing a glossary for the study and feeding them to the CAI tool, so that all participants could have equal conditions. Live third-generation CAI tools (Defrancq & Fantinuoli 2021, Pisani & Fantinuoli 2021, Van Cauwenberghe 2020) have been used, too, in exploratory studies aimed at evaluating the overall feasibility of the tool.

Coming to the design of test speeches, the strategies used by researchers have been mostly aimed at preventing the impact of confounding variables on the delivery, hence making it impossible to ascertain whether a certain phenomenon (e.g. an error) was caused by CAI tool use or a factor inherent to the source speech (e.g. speed, syntactic complexity etc.) The speech design in Prandi's study on terminology is the most elaborate (Prandi 2017, 2022). She carefully selected the number of syllables and type of terms (e.g. avoiding cognates) to include in her test speeches so that the terms would be perfectly comparable. Following the method proposed by Seeber & Kerzel (2012), she kept sentence length and syntactic complexity constant and alternated "target sentences" (i.e. those containing a problem trigger) and "control sentences" (i.e. providing a "spillover region"). Apart from Prandi's research, the test speech designs of previous studies were not fully controlled and aligned with the research questions. Van Cauwenberghe's (2020) MA thesis at the University of Ghent used a speech with specialised terms without attention to their typology or distribution. Desmet et al. (2018) and Canali (2019) numbers at random intervals equally distributed among simple whole numbers (e.g. 87 or 60 000), complex whole numbers (e.g. 387 or 65 400), decimals (e.g. 28.3) and years (e.g. 2012). Defrancq & Fantinuoli (2021) delegated the preparation of their test speeches to a colleague, which resulted in one of their four speeches presenting nearly twice as many numbers as each of the other three speeches. In Pisani & Fantinuoli (2021), an English speech was

selected from the European Commission website and more sentences containing figures were added. None of these studies took into consideration the distribution of figures in the speech in the design, as well as other factors such as what type of referents are associated to the number. In Fantinuoli & Montecchio's (2022) and Montecchio's (2021) study the numeral type and size is randomly distributed across latency levels.

#### 4.4.2.3 Users' perception

Several studies have evaluated the CAI tool based on users' perception. Typically, users' perception has been gathered through post-task questionnaires designed by the authors (De Merulis 2013, Defrancq & Fantinuoli 2021, Desmet et al. 2018, Gacek 2015, Pisani & Fantinuoli 2021, Prandi 2015, 2022) whereas no study no far used a usability questionnaire commonly used in HCI.

## 4.5 Discussion: Usability engineering in CAI

Differently from CAT tools (see Chapter 3), CAI tools did not emerge as a response to market demands for higher productivity and efficiency. Rather, their development was first prompted by conference interpreters themselves in tandem with developers, or the tools were even developed from scratch by interpreters who were programmers themselves. With the recent integration of CAI tools into RSI platforms, the development of new tools is now being prompted by interpreting businesses with the aim to increase the efficiency of interpreters' workflow and provide interpreters with a better user experience. However, the design and development processes continue to involve conference interpreters. For instance, Kudo's Interpreter Assist was developed by Claudio Fantinuoli, the conference interpreter who developed InterpretBank. While the motivations behind the development of these tools were to serve interpreters, the community has not been much involved in their design and development. While some tools that have been developed within the framework of doctoral research work (i.e. InterpretBank and LookUp) have moved from previous interpreting research and theory, no study reports of dedicated data collection in order to better specify the needs of the community as a starting point for the design. As a consequence, "tool design nowadays reflects more the ideas and habits of the respective developer, generally an interpreter himself, than the needs of the interpreter community" (Fantinuoli 2018a: 164). Furthermore, no previous project reports an iterative process of prototyping, testing and revision, which is essential to develop a mature and optimal solution.

Empirical as well as theoretical work on the workflow of interpreters and the challenges that interpreters face in their work represents the first systematic analyses of interpreters' needs. These studies may be considered as the initial input to the development of CAI tools and provided initial requirements for their development. Initial inquiries emerged in response to the absence of technological tools suitable for interpreters' preparation: before the naughties, tools for terminologists and translators already existed but interpreters found them too complex and ill-suited to their preparation process. These concerns, alongside pedagogical needs, stimulated a series of theoretical and empirical inquiries into interpreters' workflow, preparation strategies, and in-booth behaviour, which we may consider as a "forerunner" of CAI tool research and development.

Once tools began to be developed, the first empirical studies echoed the concerns that using a digital tool in the booth to access information during SI may be unfeasible and undesirable. The first empirical studies on the topic, most of which were Master's degree theses (e.g. Biagini 2015, De Merulis 2013, Gacek 2015, Prandi 2015) aimed to answer exploratory or experimental research questions concerning the hypothesised beneficial or detrimental impact of using a CAI tool during preparation or interpretation on the interpreters' delivery. The same aim may be found in studies testing the feasibility of ASR integration in the booth, which, so far, have mostly focused on numbers (Canali 2019, Defrancq & Fantinuoli 2021, Desmet et al. 2018, Pisani & Fantinuoli 2021). Alongside these studies on feasibility, Prandi (2018), Prandi's (2022) doctoral dissertation paved the way for a cognitive line of research on in-booth CAI tool use. This development seems only natural considering that the use of CAI tools in the booth "is not only challenging the way interpreting is performed, but it may have an impact on the cognitive processes underlying the interpreting task, even on some basic assumptions and theories of interpreting, for example, the cognitive load distribution between different tasks during simultaneous interpreting" (Fantinuoli 2018a: 153). Hence, it may be more appropriate to say that empirical CAI research so far has mostly provided evidence for the tool's *utility* rather than usability.

Only recently, researchers began to express an interest in research that is explicitly aimed at developing recommendations for CAI tool UI design (e.g EABM 2021b) or technical specifications, such as latency (Montecchio 2021). However, previous studies present several limitations. Tool performance was evaluated under controlled and optimal conditions only. Users' performance was evaluated mostly by delivery accuracy. However, accuracy was measured by the rendition of isolated items only (i.e. specialised terms or numerals) without taking into account broader aspects of the interpreted message. Furthermore the impact of

potentially crucial speech variables, such us the density of problem triggers in a given speech unit, has not been explored. When it comes to the evaluation of CAI tools via users' perception, post-task questionnaires have been the preferred method. Since the questionnaires vary across studies, we do not have standardised data collection methods nor benchmark values. We could hence say that, differently from CAT research, which has been enriched by cross-fertilisation with the field of HCI, empirical CAI research is still to develop a usability-focused strand of inquiry.

# 5 SmarTerp

This chapter presents the case study on which this work is based: the CAI tool SmarTerp. The chapter clarifies the background of the project, outlines its design and development process, and details the analysis activities conducted to identify users' requirements and define UI design features. Finally, it clarifies the role of testing within the design and development project, which will be presented in detail in the following empirical part of the work (Chapters 6–8).

## 5.1 Background

The SmarTerp[1] project is an Innovation Activity funded by the European Union in the framework of the EIT Digital BP2021. Its goal is to develop a remote simultaneous interpreting (RSI) platform with an integrated ASR- and AI-powered CAI tool. The project was started by the Spanish conference interpreter Susana Rodríguez. Under her coordination, an interdisciplinary team started working on SmarTerp in the autumn of 2020. The EIT Digital funding period ended in December 2021. The timeline of the research and development activities is depicted in Figure 5.1.

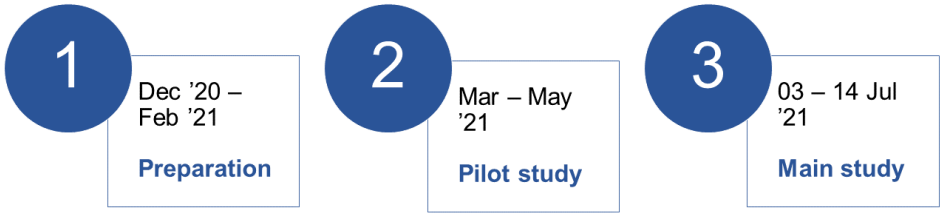


Figure 5.1: Timeline of research activities

The early development of SmarTerp under the EIT Digital grant began in the Autumn of 2020. The Activity Leader Susana Rodríguez organised a series of

[1] www.smarter-interpreting.eu

think tanks with several stakeholders to begin preparing a document on interpreters' requirements for the development of the CAI tool and the RSI platform.

The design, development and research activities began in January 2021 and continued until the end of the year. They were conducted by a consortium of universities, institutes and external consultants working from Spain, Italy, Germany and New Zealand. As for the Spanish side, the teams involved are *Next Generation Internet Group* (GING) and *Ontology Engineering Group* (OEG) from the *Technical University of Madrid* (UPM). The former was responsible for telecommunication and the development of the RSI platform, and the latter was responsible for language technologies and natural language processing. The Italian teams involved are the Trento-based *Fondazione Bruno Kessler* (FBK) and the *University of Bologna/Forlì*. FBK, in particular the SpeechTek Research Unit, was responsible for the Speech Recognition engine and its integration into the AI module. The University of Bologna was responsible for piloting the solution with their students. Within the Spain-based HCI team, Giona Fossati led the user research activities and created a design concept and Iciar Villamayor was responsible for the development.

I began collaborating on the project first as a volunteer in the focus group organised by Susana Rodríguez in the Autumn of 2020, which opened the analysis activities prior to CAI tool design (see below). From January 2021, I began to collaborate as an external researcher. My role within the team was of User Researcher. I was presented with the need to test the CAI tool empirically to improve its usability at an early stage of its development. I was responsible for designing a study that would match this practical need. At the same time, as an academic researcher, it was my interest to ensure a level of scientific rigour required to yield insights capable of informing current scientific understanding of interpreter-CAI interaction and, possibly, making progress towards the development of general principles for the design of CAI tools in line with interpreters' needs. After designing the study, I developed the study materials, procedures and analysis criteria, recruited participants, collected and analysed data and, finally, developed design recommendations which I discussed with the rest of the SmarTerp team in workshops after each round of testing and presented in a final report. I shared my study materials and methods with the other members of the Consortium so that they could be used in subsequent rounds of testing.

At the time of writing, the RSI platform reached the Technology Readiness Level[2] (TRL)9, with the actual system proven in an operational environment and ready to be marketed. The CAI tool reached TRL7 and TRL8 and is expected to

[2]https://www.esa.int/Enabling_Support/Space_Engineering_Technology/Shaping_the_Future/Technology_Readiness_Levels_TRL

reach TRL9 by the end of 2022. In the meantime, the system is being further improved with new research conducted on the development of an end-to-end system replacing the cascaded system behind the CAI tool. At the time of writing, the CAI tool works only integrated into the RSI platform. However, in the future, the development of the CAI tool as a stand-alone solution decoupled from the RSI platform is foreseen. More research is also being conducted to improve the accessibility of the tool for visually impaired interpreters.

## 5.2 Design and development process

The design and development process of the CAI tool SmarTerp was inspired by the usability engineering approach to product development (cf. Chapter 2). The process was envisaged to incorporate users' requirements, developed through various data sources, early into the design of the tool. It was also planned to incorporate data into the whole process to refine the design through cycles of testing and improvement. The detailed design process and all the research activities conducted by the HCI team are described in detail in the Master's degree thesis by Fossati (2021). The core phases of the design and development process and the activities that were conducted in each phase are detailed in Fossati (2021). In the remainder of this chapter, I will summarise the key requirements identified through the research activities I was actively involved in, namely the focus group and the literature review concerning the CAI tool.

## 5.3 Need analysis

### 5.3.1 Focus group and contextual inquiry

The first phase, from September 2020 to March 2021, consisted in the *analysis of users' needs and requirements* through the expert focus group, contextual inquiry, users' personas and literature review methods. The expert focus group was directed by the project coordinator Susana Rodríguez and took place from September to November 2020. The participants were 15 practising conference interpreters and interpreting studies scholars with expertise in interpreting technology and related fields. In four distinct sessions of approximately 1.5 hours each, they provided input on the needs of interpreters and the requirements for an in-booth CAI tool. As a result of the focus group discussions, the product owner drafted a report, in which she detailed the needs expressed by participants and their requirements on the CAI tool. The document was then sent to the focus group participants for them to check its accuracy and completeness. After

this stage, from January to March 2021, SmarTerp's user experience researcher, Giona Fossati, expanded the analysis of users needs through contextual inquiry and cognitive walkthrough methods with expert interpreters (cf. Fossati 2021). The user needs identified from this stage of analysis are reported below.

*Numbers:* users need to...

- See the numeral in its final version (and not propose partial renditions of it, i.e. the provisional ASR results, which are seen as a distractor)
- See the numeral with punctuation based on the target language, e.g. 2020 (years) vs. 2,020 (quantity)
- See the numeral and the element it refers to
- See numerals converted in the target-language system, e.g. 1 billion (EN) ⟶ mil millones (ES)
- Have units of measurement converted and rounded, e.g. 54 gallons ⟶ 204,41 litres

*Terms:* users need to...

- See both the source language and target language, so that they can detect potential machine errors (acronyms, homonyms, minimal pairs or a pair of words that differ in only one element of their pronunciation such as pin-bin)
- See the origin of the suggested term (whether it comes from the interpreter's glossary, a specialised dictionary, a database, etc.)
- Review and validate/modify displayed terms record online/at the end of the session and store/download for future occasions

*Named entities:* users need to...

- See further information about the place, person, thing, etc. being referred to, in an expandable link

Further needs expressed by focus group participants were for items to be clearly identifiable and always visible on the screen. As predictable, considering that "users are not designers" (cf. Chapter 2), when it came to the exact design features opinions diverged substantially. Some interpreters hypothesised that they may find the whole running transcript most helpful, others that they would find individual elements most helpful. Some speculated that all problem triggers should be displayed in a single interface field, others would want to see them divided into three separate fields. While all interpreters agreed that some

visual signalling would make the items clearer to identify, some suggested doing so through colours, others through size, brightness, etc. All focus group participants confirmed the utility of the CAI tool in its core function that consists in displaying named entities, numbers, as well as specialised terms and acronyms

### 5.3.2 Design-focussed literature review

While the contextual inquiry proceeded in parallel, from January to February 2021, I conducted a literature review of empirical CAI research. The focus of my literature review was on previously highlighted requirements for in-booth CAI tool use as well as possible design hypotheses emerging from the findings of previous studies. The key issues emerging from the literature review were discussed in a workshop with the UI design team to integrate relevant information into the developing design concept.

A first section of the literature review focused on previous studies on CAI tools and the general requirements for CAI tool development. I reported the general principles defined by Fantinuoli (2017) for ASR-based CAI tools:

> To be used with a CAI tool, an ASR system needs to satisfy the following criteria at minimum:
>
> - be speaker-independent
> - be able to manage continuous speech
> - support large-vocabulary recognition
> - support vocabulary customisation for the recognition of specialised terms
> - have high performance accuracy, i.e. a low word error rate (WER)
> - be high speed, i.e have a low real-time factor (RTF)
>
> [...] As for CAI tools, in order to successfully support the integration of a ASR system, the tool needs to satisfy the following requirements:
>
> - high precision, precision being the fraction of relevant instances among the retrieved instances
> - high recall, recall being the fraction of relevant instances that have been retrieved over the total amount of relevant instances present in the speech
> - if a priority has to be set, precision has priority over recall, in order to avoid producing results that are not useful and may distract the interpreter

- deal with morphological variations between transcription and database entries without increasing the number of results
- have a simple and distraction-free graphical user interface to present the results.

(Fantinuoli 2017)

I also reported the findings of the EABM survey (EABM 2021b). I reported that most respondents to the survey expressed the following preferences:

- A vertical layout where new items are added under previous items (59.24% of respondents).
- Items remain on screen as long as possible and only disappear when there is no longer any room left on screen (82.29%).
- Terms are on the left, numbers on the right (39.62%) or both items are in the same interface section (39.43%).
- New items appear in a bold font (38.11%), a larger font size (23.32%) and/or in a different colour (27.87%).

As predictable, survey respondents' opinions diverge substantially on concrete UI features. Furthermore, because "users are not designers" and "users do not know what's best for them" (cf. Chapter 2), the design team agreed that the EABM survey findings should be used as a starting point to develop hypotheses about users' requirements but not directly applied as design principles.

I then reviewed relevant findings concerning the interpretation of numbers and the inherent sources of error – a topic which I had researched extensively in the past (Frittella 2017, 2019a) and training conference interpreters on Frittella (2019b). The HCI team felt that a dedicated literature review was needed on this specific topic because the design of CAI tool UI for numbers required choosing between a variety of options (e.g. should numerals be displayed as a word or an Arabic numeral? In the source and the target language? What else should we display? etc.)

My literature review provided a brief summary of the causes why errors occur in the SI of numbers (cf. Frittella 2017, 2019a) to shed light on possible design principles. First, the cognitive processing of numerals during SI may be simplistically modelled as the sub-processes of *decoding* (comprehension of the source-language numeral), *transcoding* (turning the mental representation into a graphic numeral, e.g., in the Arabic code), *recoding* (turning the graphic numeral into the

target-language numeral). All these sub-processes are likely to require some degree of cognitive control, i.e., not to be automatic – which, among other evidence, is revealed by the fact that errors may occur in each of these phases. Another example where it is possible to see the non-automaticity of these sub-processes is the fact that interpreters often write down large source-language numerals not as a full Arabic numeral but rather as a mixed code of Arabic numerals, for the digits, and phonological code, for the order of magnitude – most commonly when above "thousand" (e.g., three million ⟶ 3Mio). The mixture of codes is likely to be utilised to simplify the transcoding and recoding processes. In their recent study on the CAI-tool supported SI of numbers, Pisani & Fantinuoli (2021) displayed the numeral as a combination of graphic Arabic code for digits and *source-language* phonological code for orders of magnitude above thousand. The results show errors in participants' recoding of the order of magnitude (e.g., they saw "million" appear on the screen and interpreted it as "billion"). This error pattern is predictable in light of the discussion above. I hence postulated that, to be maximally efficient, a CAI tool should support the interpreter in all phases of item processing. In the case of numbers, this means supporting the whole process until the recoding phase. Therefore, we decided to display numerals as a mixture of graphic Arabic code for digits and target-language phonological code for orders of magnitude above thousand.

Second, interpreting numbers does not only mean transcoding digits. Numerals are only one part of a *numerical information unit* (NIU). To express the meaning of the NIU, the interpreter must render the other components of the information unit as well as the numeral. These are the *referent* (i.e., the entity that the numeral refers to), the *unit of measurement* (the standard measure of the given value), the *relative value* (e.g., increase, decrease, etc.), as well as its *temporal* and *spatial* location. Ideally, a CAI tool should help the interpreter reconstruct the whole NIU since numerals alone out of context do not convey meaning. However, this is not possible at the current stage of technological development. Currently, item extraction is based on word recognition or may take place through a syntactic analysis of the speech. Identifying the components of the NIU would require a real-time semantic analysis, which current algorithms cannot perform. Therefore, I recommended that the CAI tool display at least the numeral together with the following element in the sentence, which is usually either the referent (as in 3 *people*) or the unit of measurement (as in $3m^2$).

## 5.4 Design requirements and features

The user requirements derived from the focus group discussions, along with relevant previous research and the knowledge of the HCI experts within the project, led to the formulation of principles for the design and development of SmarTerp. Below, I will summarise key user requirements extrapolated from the various sources used in the need analysis phase and map the chosen design principles onto those requirements. The principles are divided into the categories (1) general UI features, (2) problem triggers, and (3) technical specifications.

### 5.4.1 General UI features

The principles concerning the general UI features of SmarTerp refer to the ways in which elements are organised on the interface. Before detailing the principles, a brief terminological clarification is necessary at this point. The design team referred to the individual elements (terms, numbers, etc.) as *items* and to the graphic unit containing an item as *item box*. A *module* is a user interface section containing items of the same category. Figure 5.2 identifies these key interface units on the first version of the SmarTerp interface.

The first requirement concerning general UI features for SmarTerp that emerged from the focus group discussions was that as many items as possible should be simultaneously visible. For this reason, the design team decided to structure the interface into three modules. In left-to-right order, the modules were named entities, terms and acronyms, and numbers. A further requirement was that items should be clearly and immediately identifiable. To fulfil this requirement, the design team decided to place items into an item box and highlight new items in a different colour.

Finally, a requirement of users was that the interface should be as unobtrusive as possible. The design team hypothesised that displaying items as a scrolling list, with new items appearing on top and causing already displayed items to scroll down, may attract the attention of users causing unnecessary distraction. They therefore opted for a mode of display where a new item replaces the last one to appear on the screen. For example, if items A, B, and C are already displayed, item D will replace A, E will replace B, etc.

### 5.4.2 Item classes

The design principles concerning item classes led the design team to defining how the problem trigger classes should be presented. The design team decided

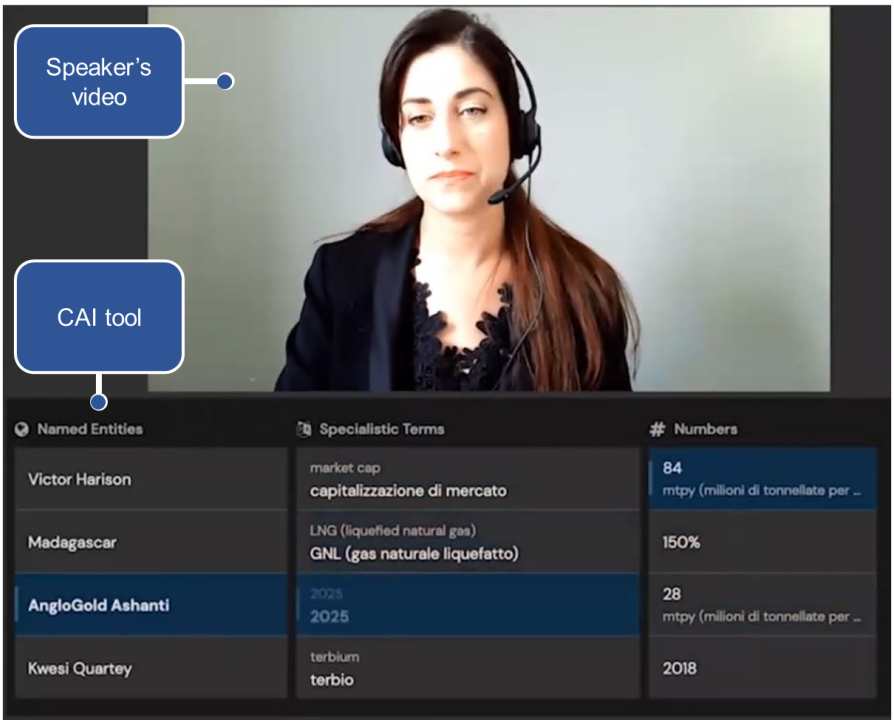


Figure 5.2: SmarTerp's interface – first prototype

that *named entities* should be transcribed using the official transcription in the target language because this emerged as the preferred option from the focus group discussions. Coming to *terms* and *acronyms*, users expressed the need to be able to check the accuracy of the tool's suggestions. For this reason, the design team decided to display terms and acronyms both in the source- and the target language, with the TL version being highlighted to make it more easily identifiable. Users also wished to be presented both with the short and the long version of acronyms, so that they could select which version to use based on the audience's prior knowledge and time constraints in interpreting. When deciding how to display numbers, the requirements were derived both from the focus group and from the literature review (presented above). Because of the complexity of all phases of numerical transcoding, where errors can emerge also in the transcoding from Arabic graphic numeral to the TL numeral, I formulated the design requirement that numerals should be displayed as a mix of graphic Arabic code (for digits) and TL verbal code (for orders of magnitude above "thousand"). Where the SL and TL code do not correspond, the numeral should be recoded

into the TL (e.g., 1 billion should be displayed as "mil millones" in Spanish). Furthermore, because errors in the interpretation of numbers may be caused not just by the numeral itself but, rather, by the whole numerical information unit, I recommended that the numeral be accompanied by the unit of measurement and the referent in the same item box. Because this was technically unfeasible, the developers proposed to display the following element in the sentence in the item box, which could be either the unit of measurement or the referent depending on the sentence structure.

### 5.4.3 Technical specifications

A final set of principles concerned the technical specifications of the tool. Based on the literature review, in particular the requirements formulated by Fantinuoli (2017), we posited that the latency should not exceed two seconds and that precision should be favoured over recall, i.e., items should only be displayed if accurate and complete.

## 5.5 Testing

Based on the requirements and features presented above, the HCI team designed an interface concept, which was reviewed internally. Then, they developed a high fidelity prototype, which was then tested empirically in two usability tests (see timeline in Figure 5.1). Because the tests took place at an early stage of the CAI tool's development, we referred to the testing cycles as *Early Testing*, of which I was responsible. The testing consisted in two cycles. The first cycle, which we called *pilot study*, aimed at gathering initial evidence by testing the prototype on 5 conference interpreters. We expected that this first round of testing would allow us to detect and correct a number of bugs and basic usage problems that we could correct as early in the development as possible. The pilot study also aimed at validating the testing methodology, which was a necessity because the materials, procedures and methods are novel in the field of interpreting technology and CAI tool research. The second cycle, which we called *main study*, aimed to gather evidence on a larger sample (10 conference interpreters) using the CAI tool UI polished from the issues identified in the pilot study. We expected that, working with a more mature prototype and validated methods and materials, we would identify deeper usage problems and focus on more fine-grained design principles.

Before defining the most suitable testing method and designing the study, I discussed the aims of testing with the whole team to make sure that expectations and key objectives would be met. Through the discussions, I realised that

the team had objectives of different types that they hoped to achieve through the early testing. The central objective was of validating or discarding the design principles that guided the UI design to improve it. Another objective of the HCI team was to ascertain the extent to which the tool fulfilled its purpose – that of supporting users in achieving an accurate performance, and whether it was perceived as satisfactory. A further objective was gaining a deeper understanding of users and their needs, revising and expanding our analysis. This aim emerged from the fact that interpreters' needs related to the use of ASR-powered CAI tools in the booth have not yet been the object of dedicated empirical explorations. Therefore, questions related to issues such as users' acceptance (why do interpreters introduce a CAI tool into their workflow? why not?), perceived utility (when is a CAI tool most helpful to interpreters? under which circumstances is it less helpful?), training needs (do interpreters need to be trained on using CAI tools? what type of training?) etc., are still largely unexplored. Finally, my personal objective was to inform scientific understanding of interpreter-CAI interaction and contribute to defining principles for the design of CAI tools' interface adequate to interpreters' needs. In the following chapter, I will discuss the methods of the study in detail.

# 6 Usability test of SmarTerp: Methods

This chapter opens the empirical section of the work by presenting the methods used in the two cycles of prototyping, testing and improvement of the SmarTerp prototype (cf. Chapter 5). The chapter first clarifies the study aims, choice of research approach and study design. It then details the rationales for the development of study materials and their characteristics. After discussing the participant selection criteria and procedures, the chapter presents the data collection procedures and, finally, the analysis criteria.

## 6.1 Aims

As explained in the previous chapter (cf. Chapter 5), the testing of SmarTerp aimed to achieve goals of different nature: improving the UI, assessing the extent to which the tool was useful and satisfactory to users and contributing to scientific understanding of interpreter-CAI interaction. To better define and operationalise the goals of the study, I specified them as follows. Starting from the goal of improving the interface, I defined the research question as follows:

> Do problems systematically occur when interpreters use SmarTerp's CAI tool? Can these problems be attributed to design features or technical specifications of the CAI tool?

I specified the aim of evaluating the impact of the CAI tool on users and their satisfaction as follows, drawing on ISO's (2018) definition of usability

> Evaluating the extent to which the CAI tool SmarTerp may support representative users in completing representative tasks in a way that is effective, efficient and with a high degree of satisfaction.

Because the study aimed to fulfil not just the practical aim of developing recommendations for the improvement of SmarTerp but also a scientific aim, the analysis aimed to identify the usability factors that have an impact on interpreters' performance and, possibly, explain why. In other words, the aim above could be reformulated as the following questions:

> To what extent was SmarTerp effective/efficient/satisfactory to test participants? Why was it (not)?

## 6.2 Choice of research approach

The aims of the study explained above required the collection of both performance and perception data and their analysis with both quantitative and qualitative methods. Performance data was required to evaluate the tool's actual effectiveness and efficiency, since participants' perception may not be regarded as a reliable indicator of these aspects. Perception data was required to ascertain participants' satisfaction with SmarTerp. Effectiveness, efficiency, and satisfaction had to be evaluated quantitatively to answer the research questions concerning their extent, developing an overall description of what happened in the test and obtaining benchmark values against which to compare future testing results. These usability aspects also had to be evaluated qualitatively to identify deeper patterns that could not be grasped at a mere quantitative description and unveil reasons behind users' observable behaviour and reported perceptions.

These requirements are consistent with the double practical and scientific aim of the study. On the practical side, the need was to extrapolate design recommendations from the collected evidence, which required an interpretive process whose robustness increases if multiple data sources are used. This is particularly necessary for the exploration of complex systems, such as the interpreter-CAI tool interaction, in which we assume both human factors and tool functions to exert an influence on the outcome making it necessary to isolate the phenomena caused by tool-related issues from those arising from contingent and idiosyncratic factors. On the scientific aim, contributing to the field's understanding of interpreter-CAI tool interaction required a level of rigour and detail that is usually not achieved in usability tests conducted for purely commercial purposes (cf. Barnum 2020).

In the light of these requirements, I opted for a mixed-method approach. Mixed-methods is a research approach that emerged around the late 1980s and early 1990s informed by the pragmatic worldview that research design should be driven by the study objectives rather than rigid canons, which justifies the combination of data from multiple sources and methods from across the quantitative-qualitative spectrum (Creswell & Creswell 2017: 294–324). This is considered as a particularly adequate methodology to explore complex systems that must be studied from different perspectives to gain a holistic and reliable understanding.

This design does not only offer benefits but also poses challenges for the inquirer, including the need for extensive data collection, the time-intensive nature of analysing both qualitative and quantitative data, the requirement for the researcher to be familiar with both quantitative and qualitative forms of research and the complex flow of research activities (Creswell & Creswell 2017: 298). To

cope with these challenges, I opted for a small sample, in line with the aim to obtain a complete, robust interpretation over statistically generalisable results. I also dedicated great attention to the design of materials and procedures to have a clear and efficient organisation.

## 6.3 Study design

In order to address the aims, the study design had to fulfil some requirements. The first was involving *representative users* reflective of the target group for whom SmarTerp was designed, i.e. practising conference interpreters fulfilling a series of selection criteria. The second was using the *working* prototype of SmarTerp in action during the simultaneous interpretation of a speech. The third was ensuring that during the test the users would perform tasks representative of those that they would normally perform with the support of the tool, which required a special speech design strategy. Because of all these characteristics, I chose *usability testing* as a method.

The individual test sessions consisted of (1) an *interpreting test* in which participants were asked to interpret simultaneously an ad-hoc designed speech with the support of a mock-up of the SmarTerp prototype, (2) a *usability questionnaire*, and (3) a semi-structured *interview*. The interpreted test was aimed at gathering performance data to evaluate the tool's (actual) effectiveness and efficiency. The questionnaire aimed to gather a quantitative measure of participants' perception of the tool and its usability. The interview was aimed at deepening the analysis of both performance and perception and providing a starting point for the patterns that emerged in the delivery.

To abstract usage problems from the bulk of data collected in the study and develop design recommendations, I followed the steps proposed by Seewald & Hassenzahl (2004). In the *experience* (or *data collection*) phase, I collected data concerning participants' performance in using SmarTerp as well as their perception. In the construction (or data analysis) phase, I analysed the data using the methods detailed in §6.7. The data sources, materials used to collect them, and the outcome of the analysis are summarised in the table below (Table 6.1).

In the *interpretation* phase, I identified recurring issues in participants' performance (e.g. a frequent pattern of error) or that were reported by participants (e.g. a frequent complaint). To identify recurring issues, I used the following processes. Starting from the task success rates, I looked for patterns in tasks with a particularly low score. Looking at the distribution of error categories, I searched for patterns where a given error category occurred with high frequency in a

Table 6.1: Overview of the study design: Data sources, aims and materials

| | Study design | | |
|---|---|---|---|
| | CAI-tool assisted SI test | Post-task questionnaire | Post-task interview |
| Type of data gathered | Quantitative and qualitative performance data | Quantitative perception data | Qualitative perception data |
| Data description | Participants' video-recorded interpretation with SmarTerp support: delivery, observations, and interview quotes | Questionnaire data | Interview data |
| Data collection materials | Source speech video with integrated SmarTerp interface (mock-up) | Digital questionnaire (7-point Likert scale items) | Semi-structured interview protocol |
| Outcome of analysis | a) Quantitative: success rates on tasks<br>b) Qualitative: error patterns | Participants' agreement with statements related to general satisfaction and pragmatic usability criteria | Users' explanation of usage problems and reflections about their experience using SmarTerp |

given task. I complemented this search with the observations I made during the test, participants' questionnaire responses, and recurring themes in the interview data. I considered whether data from the performance and perception sets converged into an explanation for the given issue (e.g. several participants reported to have struggled with a specific tool feature in a specific task where I could identify a recurring error pattern). This way, I developed an interpretation of recurring usage problems.

In the *prioritisation* phase, I ranked the usage problems by their frequency and impact. Impact ranks the issue by the severity of its consequences on successful task completion (i.e. complete and accurate interpretation). The levels of impact are:

1. *High*: prevents the user from completing the task and gave rise to critical errors in the test.
2. *Moderate*: causes user difficulty and gave rise to non-critical errors in the test.
3. *Low*: a minor problem that does not significantly affect task completion.

Frequency is given by the percentage of participants who experienced the issue:

1. *High*: 30% or more of the participants experienced the issue.
2. *Moderate*: 11–29% of participants experienced the issue.
3. *Low*: 10% or fewer of the participants experienced the issue.

In the *recommendation* phase, I proposed some design recommendations that could help solve the issues identified in the study. I finally discussed the proposed recommendations in a workshop with the SmarTerp HCI and development team to verify and finalise them.

## 6.4 Materials

### 6.4.1 Test speech design

Usability testing requires the observation of users' interaction with the product in tasks that are relevant to them and representative of the tasks they would normally perform with the product. In the context of a CAI tool, the tasks are defined by the problem triggers in the source speech that will prompt interpreters

to use the CAI tool. This implies that the usability test of a CAI tool requires the careful drafting of the test speech to create the necessary conditions for complete and meaningful observations.

I posited that the source speech design should satisfy five criteria. First, it should include *all problem trigger classes* for which CAI tool support is offered. Second, the problem triggers should be *sufficiently challenging* for all participants to increase the likelihood that they will consult the tool during SI. Third, the problem triggers in the source speech and their distribution should vary, creating tasks of *varying complexity* that make it possible to ascertain in which conditions and to what extent CAI tool support may be effective. I posited this criterion because my past research experience on the SI of numbers (Frittella 2017, 2019a) showed that the density of numerals in a speech unit and the complexity of the elements that accompany it contribute to determining the difficulty in interpreting the information. Because no previous study on the CAI tool-supported SI systematically analysed the impact of task complexity on the delivery, I identified this as a potentially significant variable to examine. Fourth, the *content* should not be so complex and unfamiliar to interfere with CAI tool consultation. Fifth, the *speech structure* should be clear in logic and alternate test tasks with discursive speech passages to prevent the difficulty in one task from impacting the following one and confounding the analysis. A similar criterion for the design of controlled test speeches had been articulated by Seeber & Kerzel (2012) as well as Prandi (2017). Sixth, further speech *characteristics* (e.g., audio quality, syntactic complexity, speed, etc.) should prevent factors other than the problem triggers to confound the analysis.

Starting from these principles, I designed the test speech to include all problem trigger classes for which SmarTerp provided support, i.e. acronyms (henceforth abbreviated as AC), named entities (NE), numbers (NU) and specialised terms (ST). Each item (e.g. the acronym, numeral, etc.) should be challenging enough to prompt the interpreter to look at the tool – hence, common acronyms (e.g. EU), well-known named entities (e.g. Barack Obama), one-digit numerals (e.g. 3), and specialised terms that are likely to be known to interpreters (e.g. artificial intelligence) were excluded. For each problem trigger class, more than one condition was chosen to create tasks of varying complexity. For instance, based on previous research on numbers (e.g. Frittella 2017, 2019a,b), I identified a 20-word sentence containing one numeral and no other factor of complexity to be a less problematic condition than a 60-word speech passage containing eight numerals accompanied by specialised terms. This way, I conceptually defined the tasks that should constitute my test speech.

In the tasks *isolated named entity* (NE), *isolated acronym* (abbreviated as AC), *isolated term* (TE) and *isolated numeral* (NU), the given problem trigger occurs in a "simple sentence, which is defined as a sentence of approximately 20–30 words, with simple syntax and logic and no further problem trigger. In the task *numeral and complex referent* (NR), one numeral is accompanied by a complex referent (i.e., another problem trigger such as an acronym/ a named entity/ specialised term/ numerical value) in a simple sentence (defined as above). The task *numerical information unit* (NIU) is a sentence of approximately 30 words constituted of a complex referent, a complex unit of measurement (i.e., an acronym/ a named entity/ specialised term/ numerical value) and five consecutive numerals, as in the structure (referent) increased/decreased by (X%) from Y in (time1) to Z in (time2). The task redundant number cluster (NCR) is a number-dense speech passage with the following characteristics: (1) constituted of three subsequent NIUs of approximately 10-15 words each; (2) the time and place references remain unvaried and are repeated in each NIU; (3) the unit of measurement and the referent remain unvaried; (4) the numeral changes in each NIU. The task non-redundant number cluster (NCN) is a number-dense speech passage with the following characteristics: (1) constituted of three subsequent NIUs of approximately 10–15 words each; (2) time, place, referent, unit of measurement and numeral change in each NIU; (3) either the referent or the unit of measurement is complex. I included two number-dense speech passages with different degrees of redundancy into the test speech to see whether the use of the tool would encourage participants to use a strategy, such as omitting repeated elements to reduce the requirements involved in processing the speech. The task *term list* (TL) includes three specialised terms occurring in the form of a list. The task *terms in a semantically complex sentence* (TS) is constituted of three specialised terms connected by complex logical links, e.g., implicit logical passages where comprehension requires relevant background knowledge and inference. The task *complex speech opening* (SO) is the address to participants and comprises three unknown named entities of people, their charges and two acronyms. Finally, in the task *conference programme* (CP), the programme for the rest of the conference day is presented and it includes several named entities and other problem triggers. The test speech may be consulted in the appendix. Table 6.2 summarises the task names and the corresponding abbreviations.

The test speech tasks may be grouped by their level of complexity as follows:

1. *Tasks of low complexity*: AC, NE, NU, and TE are the tasks of lowest complexity, consisting of a 20–30-word sentence characterised by simple syntax and logic and presenting only one problem trigger.

Table 6.2: Test speech tasks

| Code | Name |
|---|---|
| AC | Isolated acronym |
| NE | Isolated named entity |
| NU | Isolated numeral |
| NR | Numeral and complex referent |
| NIU | Numerical information unit |
| NCR | Redundant number cluster |
| NCN | Non-redundant number cluster |
| TE | Isolated term |
| TL | Term list |
| TS | Terms in a semantically complex sentence |
| SO | Complex speech opening |
| CP | Conference programme |

Table 6.3: Test speech task example

| Code | Name | Definition | Source speech segment |
|---|---|---|---|
| AC | Isolated acronym | One acronym in a *simple* sentence (20–30 words, with simple syntax and logic and no further problem trigger). | The signing of the AfCFTA [*the African Continental Free Trade Area*] by the African Member States significantly strengthens the movement towards this goal. |

2. *Tasks of medium complexity*: TL, TS and NR are slightly more complex given that several problem triggers co-occur within a sentence.

3. *Tasks of high complexity*: the remaining tasks may be regarded as highly complex given the co-occurrence of several problem triggers in the speech passage constituting the task.

After conceptually defining the tasks, I chose a topic and a communicative context that I predicted would be unknown to most interpreters to increase the likelihood that the items would be unknown to most of my study participants (i.e., a speech taking place at the African Union). After selecting the topic, I started creating speech passages matching my previously defined task criteria, as represented in Table 6.3.

Once I crafted all my tasks, I assembled them into a speech alternating tasks and discursive speech passages following a random order. The speech opened with a 45-word introduction free of problem triggers as a "warm-up time" for interpreters. It then presented the sections: greetings to participants, conference programme, body, and conclusion. In the body of the speech, tasks were followed by a "control sentence", providing a spill over region, and 0.3-second pause, when the speech was video-recorded. Each control sentence was approximately 30–50 words long and contained no problem trigger nor factor of difficulty. The whole speech is provided in the appendix of this volume.

> *Task*: The continent currently has a *gross domestic product of USD 3.42 trillion.*
> *Control Sentence*: This represents a remarkable achievement if we consider the fast pace of our economic growth over the past decades. [0.3 pause]

Once it was ready, the test speech was video recorded in high resolution. I video recorded the speech for the first test (the pilot study) reading it out in a native-like pronunciation. The speech was video recorded again by a native speaker of English and interpreter trainer for the second test (the main study). I repeated the same procedure to design the speech used for the practice section of the training module, which may be seen as comparable in structure and tasks.

### 6.4.2 Test video

The test speech was video recorded to generate a video. The speech video was then entered into a prototype of the SmarTerp CAI tool to generate the automatic aids. The output was video recorded to generate the final video showing

the speaker's video with the synchronised visual aids at constant latency of 0.2 seconds. I chose to use a recorded video rather than the live tool in the test to provide all participants with the same input. Controlling this variable was necessary for me to focus my analysis on UI design principles.

### 6.4.3 Post-task questionnaire

The usability of the tool is determined not just by users' performance but also by their mental state while using the product. In the usability test of SmarTerp, a post-test questionnaire aimed to gather quantifiable data capturing interpreters' satisfaction with the tool after having tried it in the interpreting test.

In developing the questionnaire, I considered utilising pre-existing validated tools, such as the User Experience Questionnaire. The advantage of using research-validated tools is that they offer a high degree of construct validity. However, at a preliminary assessment within the research team and the project consultant (three conference interpreters and researchers and one HCI expert), we found existing tools to be inadequate for our evaluation scenario. For instance, the UEQ requires users to express a judgment on properties of the tool through bipolar adjectives, such as "not understandable/understandable", "inferior/valuable", "not secure/secure", etc. We found some adjectives to be ambiguous in the context of a CAI tool, which made it necessary to either change the items or opt for a self-designed tool. However, the questionnaire could not be adjusted without compromising its validity. For this reason, I decided to design a questionnaire ad hoc for the test.

My questionnaire required users to express their agreement with statements related to some usability quality of the tool on a 7-point Likert scale. The questionnaire was hence scaled −3 to +3, where −3 represents the most negative response, 0 a neutral response and +3 the most positive. I chose a 7-point Likert scale considering the well-documented users' tendency to avoid extreme judgments – which is also the rationale behind the construction of the UEQ (Laugwitz et al. 2008). The first question asked participants to express their overall satisfaction with SmarTerp to capture their overall perception of the product. The following questions asked participants to express their judgment concerning pragmatic attributes of the product: its *effectiveness* (how effective SmarTerp was in supporting the SI task in participants' view), *ease of use* (how easy it was for interpreters to use SmarTerp during the test task), *ease of learning* (how easy it was for interpreters to learn to use SmarTerp), *timeliness* (whether interpreters found the support provided by SmarTerp well-timed), and *dependability* (whether interpreters felt that they could rely on SmarTerp during the test task). Finally, I

asked participants to express a judgement of the likelihood that they would include an ASR- and AI-based CAI tool into their SI workflow in the near future. I intentionally repeated this question, which I included in the enrolment questionnaire too, with the aim to explore whether the self-reported likelihood to use ASR-powered CAI tools changed after testing SmarTerp. The questionnaire is structured as follows. As shown below, I divided the questions into sections and gave them a heading for clarity, but these were not displayed in the original questionnaire.

*Part I: General Satisfaction*

On the whole, how satisfied are you with the CAI tool's support during the test? – Options: from 1 (very dissatisfied) to 7 (very satisfied)

*Part II: Satisfaction by Pragmatic Usability Criteria*

Do you agree with the following statements? – from 1 (strongly disagree) to 7 (strongly agree)

- The CAI tool was easy to use (perceived ease of use)
- The CAI tool helped me improve the accuracy of my delivery (perceived effectiveness)
- No training is required to use the CAI tool effectively (perceived ease of learning)
- The input provided by the CAI was timely (perceived timeliness)
- I felt that I could rely on the CAI tool's support (perceived dependability)

*Part III: Likelihood to Use the Tool*

How likely are you to use ASR-integrated CAI tools as a support during simultaneous interpreting in the near future? – from 1 (very unlikely) to 7 (very likely)

### 6.4.4 Semi-structured interview protocol

In the SmarTerp usability test, a semi-structured interview complemented the analysis of users' perceptions through the questionnaire. I decided to include an interview for three main reasons. The first was to expand the analysis of users' needs and requirements and update the CAI tool features based on such deeper understanding. The second was to shed light on the reasons for students' perception of the tool's usability expressed through the questionnaire. The third was to integrate participants' self-reported data into my interpretation of critical incidents and identification of usage problems, hence adding strength to the analysis. While in usability tests this is often accomplished by asking users to "think aloud" while performing test tasks, this is obviously not possible in the context of a simultaneous interpreting task.

I conducted the semi-structured interviews based on the protocol below. Interview questions in parts I and II of the protocol were aimed at shedding light on the reasons for participants' responses to the post-test questionnaire. Part III was aimed at gleaning insight into participants' perception of the tool's support in the rendition of individual problem triggers. Part IV asked participants to express their judgment on the functions and design features of the tool. It is important to note that although, as discussed earlier, users' opinions alone cannot be used to reliably derive design principles, within a mixed-method study, users' recommendations may be compared with researchers' observations to see whether they converge or not. Part V explored participants' perception of the tool in a real context of use. During the interview, I adapted the questions (by reformulating them, asking for specifications, asking to report a critical incident, making open-ended statements that invited participants to fill the gap, etc.) to probe into issues of interest without influencing participants' responses.

*Part I: General Satisfaction*

Your self-reported satisfaction with the tool was ... [user's evaluation in the post-task questionnaire] Could you please explain this choice?

In the enrolment questionnaire, your self-reported likelihood of using a CAI tool of this kind was ... In the post-test questionnaire, it was ... Why is that?

*Part II: Satisfaction by Pragmatic Usability Criterion*

Based on the enrolment questionnaire, your perceived ease of use / effectiveness / ease of learning / timeliness / dependability is (...). Could you motivate your choice? Can you recall an example?

*Part III: Satisfaction by Problem Trigger Class*

Based on the enrolment questionnaire, your self-reported difficulty with acronyms / named entities / numerals / specialised terms is (...). How do you normally deal with this problem trigger? Do you believe that the tool helped you in the test? How? Can you recall an example?

*Part IV: Design-related Recommendations*

What did you like the most about the tool and its interface? Why?

What did you like the least about the tool and its interface? Why?

*Part V: Perception of the Tool in Context*

In your view, how could a CAI tool of this kind make a difference to you as an interpreter?

How does the tool differ from a human boothmate?

Which one is more reliable?

If you were to choose whether to use it in an online or on-site assignment or both, which option would you choose?

- If on-site or both: would you still use the help of a human boothmate? What would his/her role be in that constellation?

## 6.5 Participants

### 6.5.1 Recruitment

The participants in the study were recruited through an open call for participants made circulate through the research team's professional network and SmarTerp's communication channels. Prospective participants who expressed their interest in participating in the study received an *informed consent* to sign. The document included key information about the study objective and scope, the deadlines and the data collection and treatment procedures. They were also asked to fill out a digital *enrolment questionnaire*. The questionnaire was aimed to profile and select participants based on our selection criteria: (a) 30–50 years of age, (b) more than 10 years of professional experience as an English-Italian conference interpreter in the simultaneous mode; (c) no less than 10 workdays as an English-Italian simultaneous interpreter each year, (d) no identifiable connection with the topic of the source speech (to create equal conditions for all participants). The first ten prospective participants who responded to our call and were found to fulfil our selection criteria were notified of their inclusion and provided access to the online training module.

*Part I: Participant's Personal Information*

What are your name and surname?

What is your chosen pseudonym? (You can choose any name different from your given name as the pseudonym which will identify your data to protect your privacy. If you do not choose any pseudonym, the research team will assign one to you.)

What is your country/ region?

What's your age group? Options: below 30 – excluded; 30–40; 40–50; above 50 – excluded.

*Part II: Qualifications and Professional Activity*

Do you hold a Master's degree in Conference Interpreting or equivalent academic qualification? Options: yes / no.

Are you a member of a professional association? If yes, which one(s)? How many years of professional experience as an English-Italian conference interpreter in the simultaneous mode do you have? Options: less than 10 – excluded; 10–19; 20–30; above 30.

On average, how many days do you normally work as an English-Italian simultaneous interpreter each year? Options: less than 10 – excluded; between 10 and 20; between 20 and 30; above 30.

Do you have any fields of specialisation? If yes, which one(s)? What is the country/region of your main clients? If Africa – excluded.

*Part III: Current Use of Technology*

How many days have you worked in the remote simultaneous interpreting mode over the past 12 months? Options: less than 10 – excluded; between 10 and 20; more than 20.

In how many of your last three on-site simultaneous interpreting assignments did you bring a laptop with you in the booth and use it to browse your glossary or search for unknown terms / information while you were interpreting? Options: none; 1 assignment out of three; 2 assignments out of three; 3 assignments out of three.

Have you ever used a computer-assisted interpreting (CAI) tool? If yes, which one and for what purpose?

*Part IV: Perception of Problem Triggers*

Do you find it difficult to interpret the following items? Please, rate their difficulty on a scale from 1 (very easy) to 7 (very difficult): acronyms, named entities, numbers, specialised terms.

*Part V: Prospective Use of Technology*

How likely are you to start using a technological tool with integrated automatic speech recognition during simultaneous interpreting in the near future? Choose the most suitable option on a scale from 1 (very unlikely) to 7 (very likely).

### 6.5.2 Training

After enrolment, participants completed an asynchronous (self-paced, approx. 1.5 hours) e-course which I had previously developed and made available via the LMS Moodle. The e-course comprised the following units: (1) Theoretical introduction to in-booth CAI tool support, explaining what CAI tools are and how in-booth support works; (2) Introduction to the CAI tool SmarTerp, presenting key functions of the tool (e.g., type of support provided and latency) and key interface features (e.g., location of elements in the interface, order of appearance, mode of display, etc.); (3) Practice session, consisting in a CAI-tool assisted SI exercise of a speech equivalent to the test speech in structure and complexity but different in topic and terminology. Participants were given over seven days to complete the e-course and their activity was monitored through the user data collected by the LMS. Quizzes were embedded in each theoretical section to test participants' knowledge. The practice session required them to upload their delivery to ensure that they did indeed complete the required training. Note that the word "training" in the context of this study refers to a combination of fundamental technical information with a practice session. No guidance was provided on how to effectively integrate the tool into the SI process (e.g., making participants aware of possible threats, suggesting the use of interpreting strategies, etc.).

### 6.5.3 First iteration: Pilot study

The participants in the first iteration of prototyping, testing and revision (i.e. the *pilot study*) were five Italian conference interpreters with English as their working language (either B or C). Because the main aim of conducting a pilot study was validating the research methodology and providing an initial orientation to the design work, the inclusion criteria were not strictly held in the recruitment phase. The only criteria for inclusion were holding an MA degree in conference interpreting and being a practising conference interpreter (ITA-A, ENG-B/C). All participants signed the informed consent and completed the training phase.

### 6.5.4 Second iteration: Main study

The participants in the main study were selected following the criteria for inclusion detailed earlier in this section. Accordingly, all main-study participants are conference interpreters with at least 10 years of professional experience, 20 RSI working days over the past twelve months, and 30 interpreting assignments as an English-Italian simultaneous interpreter each year.

## 6.6 Procedure

At the time of the test, the Covid-19 pandemic made it impossible to conduct presential sessions. I also judged that conducting the test remotely would not change the nature of the interpreting task, as the SmarTerp CAI tool was intended to be used on the interpreter's laptop and primarily in an RSI setting. I conducted the test remotely via the web conferencing platform Zoom and tested the CAI tool on each participant individually.

When the participant logged in at the agreed time, I welcomed him/her and reminded him/her of the purpose of the study. I then asked permission to video-record the session. Once recording started, the participants shared their screen and accessed the test video via a link I shared with them in the Zoom chat. The video started with a slide containing information about the communicative context of the speech (i.e. event name, speaker's name, title of the speech, time and place) which remained visible for one minute. During this time, the participant could reflect on the communicative context but could not search for more information. After one minute, an acoustic signal announced that the speech was starting. At that point, the participant started to interpret the speech simultaneously and could use the CAI tool SmarTerp. Through screen sharing, I could record the integrated view of participant's face and the CAI tool operating.

While the participant was interpreting, I noted significant patterns in an observation sheet that I had previously prepared. The observation sheet consisted of a table that listed the source speech tasks in their order of appearance in the test speech. The tasks were identified through a code. The empty column to the immediate left was dedicated to notes of errors or phenomena of interest that I observed while the participant was talking. Further to the left, an "interview" column provided a space to note comments on the critical incidents that participants spontaneously discussed during the interview. The structure of the observation sheet is depicted below (Table 6.4).

After the speech finished, I asked the participant to stop sharing his/her screen and to complete the post-test questionnaire that I sent via a link in the chat. Af-

Table 6.4: Structure of the interview sheet

| Task code | Item code | Source speech segment | Observations | Interview |
|---|---|---|---|---|
| task | subtask | | notes made during the interpreting test | notes made during the interview |

ter that, the participant took a 10-minute break. In the meantime, I read their answers and integrated them into the interview protocol to ask more personalised questions (for example, instead of asking "what do you think of the tool's efficiency?" I asked, "you rated the tool's efficiency as 5, why is that?"). After the break, I conducted the interview. I made notes on the protocol while interviewing participants but checked and completed the notes after each session, replaying the recorded interview. When, during the interview, participants spontaneously made comments on specific aspects of their delivery, or if I had a specific question for them on critical incidents, I noted the conversation in the "interview" column of the observation sheet.

## 6.7 Data analysis

### 6.7.1 Performance data

#### 6.7.1.1 Evaluation approach

In Chapter 4, I discussed the limitations of performance measures that have traditionally been used in empirical CAI research to evaluate tool effectiveness. Terminological accuracy and number accuracy may not be taken as a measure of overall interpreting "accuracy" or "quality" and hence tool effectiveness because a term or number that is correctly interpreted but wrongly contextualised still corresponds to an incorrect interpretation. Therefore, I decided to adopt a more comprehensive approach to the evaluation of interpreters' performance in the test, which I call, for lack of a better term, the *communicative approach*.

I developed this approach in my previous research on the SI of numbers (Frittella 2017, 2019a). In such work, I identified a limitation in previous studies on the SI of numbers that had analysed the interpreted numeral only. By doing so, they

failed to capture severe errors such as implausible numbers, sentences left unfinished, etc. In search of a more comprehensive analysis framework, I developed the *Processing Ladder Model for the SI of Numbers* (Frittella 2017, 2019a) inspired by Chernov (2004) *Probability Prediction Model*. The Processing Ladder Model analyses several levels of meaning of interpreted numerals to identify a broader range of phenomena and error patterns.

Given that empirical CAI research is still in its infancy and most previous studies focused on the rendition of terms and numerals in isolation, I decided to apply the framework I previously developed to the present analysis. In evaluating interpreters' performance during the test, I took into consideration the following "layers of meaning" in their delivery:

*Word:* was the item (i.e. the acronym, named entity, number or specialised term) accurately rendered?

*Sentence:* was the item accurately embedded in a well-formed and complete sentence?

*Text:* is the delivery internally consistent and congruous in meaning with the source speech?

*Context:* does the delivery make sense and is it plausible against the background knowledge of an informed audience?

*Function:* is the interpreter's delivery equivalent in function to the message intended by the speaker?

I consider a delivery segment as fully accurate only if all questions above may be answered positively. I applied this rationale both in defining criteria to calculate task success rates and in the analysis of error patterns.

#### 6.7.1.2 Preparation

To analyse participants' performance in the test, I started from the observation sheet (in Excel) and expanded it. The worksheet hence presented the relevant parts of the source speech (i.e. the "tasks") that were further segmented into smaller units (i.e. the "subtasks") corresponding to the individual items (i.e. the problem triggers) and the surrounding elements. I transcribed the relevant segment of participants' delivery to the right of the source speech segment. Then, I duplicated the spreadsheet and formatted one version for the quantitative analysis of task success rates and the other for the qualitative analysis of error categories.

#### 6.7.1.3 Task success rate

To calculate the success rate of participants' rendition of each task, I first coded the delivery segments corresponding to each subtask using the following notation:

- Accurate rendition, space is left blank.
- Error, marked as "e" and identifying severe semantic errors in which the meaning of the source speech passage is substantially changed in the delivery (i.e. considering the levels word to function of the communicative analysis framework described above).
- Omission, marked as "o" and identifying an omission of the item.
- Other issue, marked as "x" and identifying any other error of secondary importance (i.e. not corresponding to a substantial semantic error).

I then attributed a success rate to each subtask based on the following criteria:

- Correct rendition (100%): The delivery is correct and complete.
- Strategy (100%): The interpreter uses a strategy that does not change the meaning of the message and does not cause a loss of information, e.g., In 2021 ⟶ this year.
- Minor error or missing detail (95%, marked as "x"): The delivery is accurate and complete, apart from a minor imperfection or a missing detail, such as a missing adjective, speech disfluency, etc., e.g., Giovanie Biha ⟶ Giovanie "Beha".
- Partial rendition (proportional to task content, marked as "x"): Some element of the subtask is missing but the overall meaning is still comprehensible, e.g., Ministry of Trade and Industry ⟶ Ministry of Trade (−30%).
- Generalisation or summarisation (30%, marked as "o"): The interpreter omits the item and summarises the information, e.g., coal-bed methane has a huge potential ⟶ natural gas has a huge potential.
- Omission (0%, marked as "o"): The item is omitted without the use of a strategy, which causes the whole message and its informative content to go lost.

- Semantic error (0%, marked as "e"): The delivery is implausible, inconsistent or nonsensical, e.g., Africa has a GDP of USD 3.42 trillion ⟶ Africa has a GDP of USD 3.42 million.

Finally, I calculated the task success rate as the mean value of the success rates of its constitutive subtasks. The codes and corresponding success rates are summarised in Table 6.5 below.

Table 6.5: Task success rate – evaluation criteria

| Error / Phenomenon | Error Code | Success Rate |
|---|---|---|
| Correct rendition | | 100% |
| Strategy | | 100% |
| Minor error and missing detail | x | 95% |
| Partial rendition | x | Proportional to task content |
| Generalisation, summarisation | o | 30% |
| Omission | o | 0% |
| Semantic error | e | 0% |

#### 6.7.1.4 Error patterns

To identify qualitative error patterns, I started from the communicative error categories presented in Frittella (2017, 2019a). In this framework, each level of analysis of the *Processing Ladder Model* presented corresponding error categories. I chose this framework because it is in line with the general communicative approach I adopted in my analysis. The levels of analysis are *word, sentence, text,* and *context*. I added *strategy* as a separate category to ease the identification of patterns at this level. At each level of analysis, the specific codes are the following.

Analysis at the word level focuses on the delivery of problem triggers (acronyms and terms, numbers and named entities) in isolation, for which the CAI tool provides support. At the word level, the error categories are *omission* (if the problem trigger is not interpreted), *error* (if the problem trigger is incorrectly interpreted), *partial rendition* (if only part of the problem trigger is rendered), *pronunciation error* (if the problem trigger is incorrectly pronounced), *gender error* (if a person's gender is misinterpreted).

Analysis at the sentence level focuses on the accuracy of the sentence containing the problem trigger for which CAI tool support is provided. The first

error category is *misattribution*, if sentence components are incorrectly linked in the interpretation, for example, the address "Honourable Soraya Hakuziyaremye, Rwanda's MINICOM Minister; Ms. Giovanie Biha" is interpreted as "Honourable Soraya Hakuziyaremye; Ms. Giovanie Biha, Rwanda's MINICOM Minister". Sentence fragment occurs when the interpreted sentence is grammatically incomplete, for instance the referent of a numeral is missing as in the example "Namibia produced 2.52 million carats ['of diamonds' is omitted]". An error at the sentence level is registered if the problem trigger is correctly interpreted but an accompanying element of the sentence is misinterpreted, as in the sentence "Namibia imported [rather than produced] 2.52 million carats of diamonds".

Analysis at the text level focuses on the meaning and consistency of the speech passage containing the problem trigger. A first error category is *inconsistency*, when parts of the delivery are mutually contradictory, as in the delivery sample "by 2030, the African continent will have about 295 million people aged 15-to-64. In 2030, there will be 1 billion people aged 15-to-64 in Africa", where the second statement contradicts the first. A second error category is *distortion*, if the delivery, albeit internally consistent, differs substantially in meaning from the source speech, as in the following example:

> *Source*: … makes coal-bed methane reservoirs advantageous for commercial operations.
> *Delivery example*: coal-bed methane is hence important to boost the growth of the energy sector.

Analysis at the context level focuses on the external plausibility of the interpreted message. *Plausibility errors* correspond to implausible interpretation, for instance "by 2030, Africa will be home to 1 billion people". The error category *nonsense* is attributed when the delivery does not express a logical statement, as in the example below:

> *Source*: Though its contribution to the total energy mix is still modest, coal-bed methane has impressive potential.
> *Delivery example*: The energy mix is still in its first stage, but this bears great potential too.

Finally, strategy categories are omission of redundant item (if the interpreter omits an item that is repeated within the numerical task), lexical substitution (if the interpreter replaces the item with an equivalent lexical element), generalisation (if the interpreter produces a sentence with general meaning), and summarisation (if the interpreter summarises the meaning of the speech unit).

The table below summarises the categories of error and strategy used in the study and shows their correspondence to the quantitative analysis measures presented in the previous section.

Table 6.6: Categories of error and strategy in the deliveries

| Error Code | Success Rate | Name |
|---|---|---|
| o | 0% | Omission |
| e | 0% | Error (word level) |
| x | proportional to task content | Partial rendition |
| x | 95% | Pronunciation error |
| x | 95% | Gender error |
| e | 0% | Misattribution |
| e | 0% | Sentence fragment |
| e | 0% | Error (sentence level) |
| e | 0% | Inconsistency |
| e | 0% | Distortion |
| e | 0% | Plausibility error |
| e | 0% | Nonsense |
| | 100% | Omission of redundant item |
| | 100% | Lexical substitution |
| o | 30% | Generalisation |
| o | 30% | Summarisation |

### 6.7.2 Questionnaire data

To analyse the data gathered in the questionnaires, I entered them into an Excel spreadsheet and converted the 7-point Likert scale into a −3 to +3 scale. For each questionnaire item, I calculated the measures of central tendency: the mean (i.e. the average of the dataset), the median (i.e. the middle value in the dataset) and the mode (i.e. the value that occurs most often). While the mean is the most commonly used measure of central tendency in a numerical data set, it may not be a fair representation of the data because it is easily influenced by outliers, which is particularly problematic in small datasets. The median and mode are more robust, i.e. less sensitive to outliers. I, therefore, calculated all three measures of central tendency in the main study (10 participants) but only the mean in the pilot study (5 participants) because the sample size in the latter case was too small

to calculate the median and mode. To compensate for this and make it possible to identify the influence of outliers on the mean, I decided to report all individual values of the questionnaire responses. To interpret the questionnaire results, I adopted the standard interpretation of the scale means in the UEQ (Laugwitz et al. 2008):

- $< -0.8$: negative evaluation
- $[-0.8, 0.8]$: neutral evaluation
- $> 0.8$ positive evaluation

### 6.7.3 Interview data

To code the interview data, I turned the interview protocol into an Excel spreadsheet, in which each line corresponded to a question and each column to a participant. I transcribed participants' responses into the relevant cell of the spreadsheet, paying particular attention to quoting the exact words that participants used. I partly took notes during the interview and then replayed the whole interview afterwards to complete the transcription. When I found the answer ambiguous during the interview, I asked participants to reformulate their answer and then I asked for confirmation of my interpretation of their words. I then analysed the interview data as follows. For each question, I calculated how often a certain concept emerged in participants' answers. I divided the concepts into factors influencing participants' perception (e.g. of the CAI tool's features and usability) positively and negatively. Within each category, I counted how often the same concept had been expressed by study participants. I summarised the outcome in a table and translated representative quotes for the most important concepts into English.

# 7 First iteration: Pilot study

The purpose of the pilot study in the usability test of SmarTerp was to validate the methodology and provide initial orientation to the design. It hence represents both a tool for methodological validation and the first iteration of prototyping, testing and improvement. It was conducted with five participants, who were Italian (A) – English (B/C) conference interpreters (see the description of study participants' characteristics in Chapter 6). Because this test was conducted using a prototype, the tool was not always accurate. It omitted and misrepresented some source speech items, as described below. This first analysis hence shed light on some patterns of error that may occur when users are presented with inaccurate input by the CAI tool.

## 7.1 CAI tool inaccuracies

Due to imperfections in the CAI tool prototype, the test speech presented 15 issues in a total of 52 displayed items. These may be categorised as follows:

- *Errors* (code 'e'), no. 5: An item is displayed differently than it should have, e.g., "58,000" is displayed as "58.00.00".
- *Omissions* (code 'o'), no. 3: The CAI tool does not display an item that should have been displayed.
- 7 CAI tool *limitations* (code 'x'): The CAI tool does not display an element that participants consistently found significant, e.g., it does not display charges, like President, Minister, etc.

While errors and omissions are actual inaccuracies in the prototype performance, what I categorised as 'limitations' are elements that were not displayed by the tool because it was not programmed by default to provide such aids. I added this category to my analysis because, during the test, I observed that participants repeatedly stumbled on these items. I hence decided to dedicate particular attention to this recurring pattern which I interpreted as potentially significant.

## 7.2 Users' performance

### 7.2.1 Task success rates

Table 7.1 reports participants' success rates in speech tasks as well as the mean success rate on each task. As discussed in the method section, AC, NE, NU, and TE may be regarded as the tasks of lowest complexity, TL, TS, and NR as tasks of medium complexity, and NIU, NCR, NCN, SO, and CP may be expected to be of higher complexity. An asterisk identifies the speech tasks in which the test CAI tool prototype produced an error or omission.

Table 7.1: Mean task success rates (pilot study). "*": CAI tool error, omission or limitation is present in the task

| Code | Task | Mean |
|---|---|---|
| AC | Isolated acronym | 100% |
| NE | Isolated named entity | 96% |
| NU | Isolated numeral | 80% |
| NR | Numeral and referent | 100% |
| NIU | Numerical information unit* | 48% |
| NCR | Redundant number cluster* | 53% |
| NCN | Non-redundant number cluster* | 42% |
| TS | Terms in a semantically complex sentence | 45% |
| TL | List of three unknown terms* | 65% |
| TE | Isolated term | 78% |
| SO | Complex speech opening* | 78% |
| CP | Conference programme | 91% |

Two outliers may be found in the high success rates that participants obtained on low-complexity tasks, i.e. Diana's interpretation of NU and Sally's interpretation of TE. Based on the interviews, these were due to a distraction due to an interpretation error in the previous passage (Sally), and to the display of the order of magnitude "trillion" ($10^{12}$), which the tool displayed as "bilione" (Diana). While "bilione" is the correct Italian translation of "trillion", it is rarely used – common alternatives are: "trilione" ($10^{18}$), incorrect translation but used with increasing frequency as a loanword from English, and "mille miliardi", a correct and more native alternative but we expected it to be more difficult to process during SI.

### 7.2.2 Response to CAI tool inaccuracy

This section of the analysis shows what happened when participants received inaccurate aids by the CAI tool, or an expected aid was missing. Table 7.2 is a contingency table showing errors, omissions and other issues that occurred in participants' deliveries when the CAI tool provided a correct and an incorrect suggestion, omitted a suggestion that should have been provided or did not display a component of the information because of its intrinsic functional limitations.

Table 7.2: Contingency table: Accuracy of CAI tool and delivery

| | Delivery issues | | | |
|---|---|---|---|---|
| | Total | Errors | Omissions | Other |
| Correctly displayed items (185) | 56 (30%) | 15 (8%) | 16 (9%) | 25 (13%) |
| CAI tool issues (75) | 52 (69%) | 19 (25%) | 29 (39%) | 4 (5%) |
| CAI tool errors (25) | 14 (56%) | 12 (48%) | 1 (4%) | 1 (4%) |
| CAI tool omissions (15) | 14 (93%) | 0 (0%) | 14 (93%) | 0 (0%) |
| CAI tool limitations (35) | 24 (69%) | 7 (20%) | 14 (40%) | 3 (9%) |

The variation of errors, omissions, and other issues in the delivery with CAI tool accuracy is represented in Figure 7.1.

When the item displayed by the tool was correct, 70% of interpreted items were evaluated as correct; if we consider only delivery errors and omissions and exclude "other issues" from the count, 84% of interpreted items were rendered correctly. However, it should be taken into account that the tasks where no issue occurred were those of lowest complexity; therefore, the results cannot be seen as a direct impact of task complexity. When the CAI tool presented an issue, only 31% of interpreted items were evaluated as correct. If we exclude CAI tool limitations from the count and only consider the instances of errors and omissions, the proportion of correctly interpreted items equals 30%.

### 7.2.3 Error patterns

This section of the analysis reports the patterns of error that most frequently occurred both when the CAI tool provided participants with correct aids and when the aids were incorrect or missing.

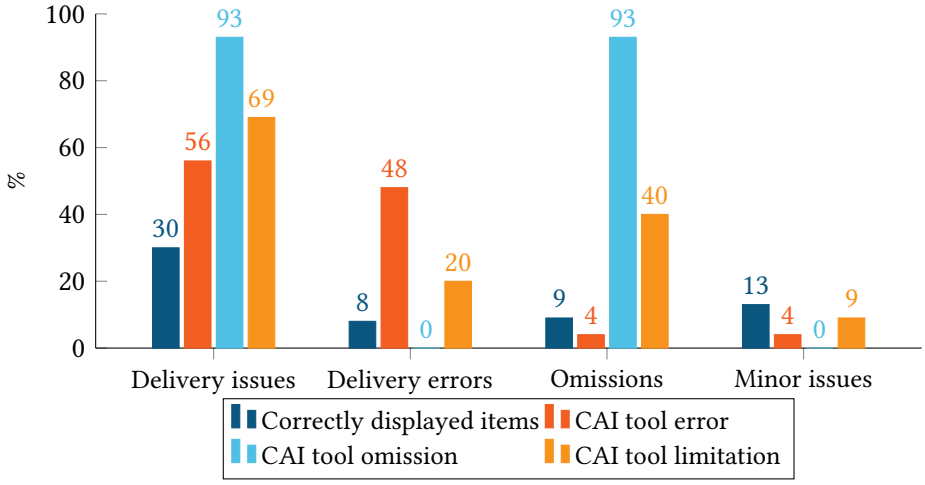


Figure 7.1: Column chart: Accuracy of CAI tool and delivery

#### 7.2.3.1 Correctly displayed items

When the item was correctly displayed by the CAI tool, nearly half (44.5%) of all delivery issues were categorised as "other issues" (cf. analysis criteria detailed in Chapter 6). One of these frequently occurring issues was *pronunciation errors* (no. 16), i.e., mispronunciation of a named entity (no. 13) and a specialised term (no. 3). The most mispronounced named entities were "Soraya Hakuziyaremye" (subtask code SO-2-1), mispronounced by 4/5 participants, and "Felix-AntoineTshisekedi Tshilombo" (subtask codes NE-2 and CP-1-2), mispronounced 7 times by 4/5 participants. The mispronounced terms were "praseodymium" (subtask code TL-1), mispronounced by 2/5 participants, and "terbium" (TL-2), mispronounced by 1 participant. Other recurring issues were *gender errors* (no 4), i.e., a person mentioned in the source speech is attributed the wrong gender by the interpreter, and a *partial rendition* of an acronym. Apart from "other issues", participants made a similar number of omissions (no 16) and more severe semantic errors (no 15) when the CAI tool displayed items correctly. No clear pattern of distribution of errors and omissions may be identified as both occur in the same tasks indiscriminately, i.e. in NCR, NCN, NIU and TS.

#### 7.2.3.2 CAI tool errors

Most CAI tool errors led to a *severe semantic error* in the delivery, as shown in the example below. In the delivery example provided, it is possible to see that

the CAI tool error disrupted the participant to the extent that she committed a plausibility error – considering that the global nickel production amounted to about 2.5 million tons in 2021, it is impossible that Madagascar alone produced 58 million. The subtask where the CAI tool error occurred had a mean success rate of 26%, with 2 plausibility errors, 2 misattributions and 1 summarisation strategy.

> *Source*: Madagascar alone produced approximately 58,000 [displayed as 58.00.00] metric tons of nickel in 2021.
> *Delivery example* (Minerva): Madagascar alone produced 58 million tons of nickel.

#### 7.2.3.3 CAI tool omissions

In nearly all cases, an item omitted by the CAI tool was omitted by study participants too (14/15, or 93% of cases). In the example below, all participants (5/5) omitted the item that was not provided by the tool and interpreted the ones provided:

> *Source*: Our soil is rich in high-value praseodymium [displayed], dysprosium [not displayed], and terbium [displayed].
> *Delivery example* (Logan): Our soil is rich in praseodymium and terbium.

#### 7.2.3.4 CAI tool limitations

Both errors and omissions were triggered by what I designated as "CAI tool limitations", which must be differentiated from CAI tool omissions because a given item is not left out because of a tool error but rather because of the way the tool functions. In the example below, for instance, the CAI tool did not display the referent "diamond production" because this was not recognised as a technical term. The mean success rate in this subtask was 46. 2 of 5 interpreters could not interpret the referent, which had been said by the speaker but not displayed by the tool, producing an incomplete sentence with a missing referent, as in the delivery example below. One other interpreter resorted to a strategy and summarised the overall meaning of the sentence without detailing the precise information: "Namibia's production increased too" (Carlo).

> *Source*: Namibia's diamond production [not displayed] amounted to 2.52 million carats.
> *Delivery example* (Sally): Namibia produced 2.52 million carats [missing referent].

Another case of tool limitation may be found when a person is introduced in the speech: his/her name and the agency s/he works for are both displayed but not their charge, as this is not recognised as a specialised term. While charges are lexical items for which professional conference interpreters should have a readily available equivalent in the target language, these were found to be often omitted or misinterpreted by study participants, as in the example below.

> *Source*: His Excellency Paul Kagame; Honourable Soraya Hakuziyaremye, Rwanda [not displayed] MINICOM Minister [not displayed], Ms Giovanie Biha UNECA Deputy Executive Secretary [not displayed].
> *Delivery example* (Minerva): His Excellency Paul Kagame, Honourable Soraya "Akugiaramie", Trade and Industry Minister, Mr Giovanie Biha Economic Commission for Africa's Secretary.

## 7.3 Users' perception

### 7.3.1 Post-task questionnaire

Table 7.3 reports the results of the post-task survey, which explored participants' satisfaction with SmarTerp and their perception of its usability. The table also shows participants' self-reported likelihood to use ASR-based CAI tools in the near future before and after the test with SmarTerp. Given the small number of survey responses, only the mean value was calculated. It is noticeable that the mean interest in ASR-powered CAI solutions (reflected by participants' self-reported likelihood to use such tools before and after testing SmarTerp) increased by a mean value of 0.8 points.

### 7.3.2 Interviews

#### 7.3.2.1 SmarTerp's UI and technical specifications

From the point of view of the tool's aesthetic qualities, participants generally referred to the "smart interface" (mentioned by four participants) as a factor increasing the appeal of SmarTerp also in comparison with other existing ASR-based CAI tools, although perceptions of aspects such as colour, font size and amount of information displayed vary among participants. For instance, two participants liked the colour choice (dark background with white typography), whereas two found it unpleasant.

The structural feature that received the most negative comments was the order of appearance of items. four participants said that they were confused by the

Table 7.3: Questionnaire results (pilot study)

| Usability aspect | Mean |
|---|---|
| Overall satisfaction | 1.8 |
| Ease of use | 2 |
| Effectiveness | 2.2 |
| Ease of learning | 1.4 |
| Timeliness | 1.8 |
| Dependability | 2 |
| Likelihood to use CAI (before) | 1.2 |
| Likelihood to use CAI (after) | 2 |

fact that items did not appear at the top of the interface and believed that the feeling of uncertainty and distraction caused by this UI feature contributed to their errors during the test. Furthermore, one participant found the division of the interface into three modules "overloading", and one thought that terms and acronyms should be displayed in the target language only.

Given that most participants immediately mentioned the tool as "easy to use", "intuitive", and "seamless", as well as suitable for everyone "even for the technology-damaged" (in Italian: "è a prova di tecnoleso", Minerva) right in opening the interview, it seems that they perceived the ease of use as a major strength of the tool, which is in line with the high scores in the post-task questionnaire. In participants' view, ease of use was given by the ASR system, which "is even better than a human boothmate: you don't even need to ask for help" (Sally). The technical specification that was most criticised by users is the tool's latency, which was defined as "too long" (three participants) or "a bit too long" (one participant).

#### 7.3.2.2 Usefulness

When asked what was, in their opinion, the main advantage in using the CAI tool, all participants (5/5) emphasised the support in dealing with interpreting problem triggers, or the "pain points of simultaneous interpreting" (Minerva). In the words of another participant: "the good thing is knowing that the most difficult part of the sentence will be there" (Sally). All participants reported that they found the tool most helpful for "unknown items", i.e. information that you cannot infer from the context and where you cannot apply a strategy: "you can round a number but you cannot approximate a named entity" (Minerva). The

tool was found to be especially needed because of the lack of preparation (Diana, Carlo, Minerva), for the following reason: "I usually do not need support in the rendition of terminology: if the terms are in my glossary, I know them. The four most important things are usually written on my notepad" (Minerva). Participants associated the availability of a support system for problem triggers with increased delivery accuracy (which justifies the high effectiveness scores in the post-task questionnaire) and reduced mental effort:

> The good thing is that you know that the most difficult part of the sentence is going to be provided by the tool. This way, you don't have to put any effort into energy-consuming tasks such as writing down numerals. (Sally)

Two participants also spoke of a greater feeling of security in using the tool and defined it as a "safety net" (Minerva) and a "further confirmation" (Sally, Diana).

When asked if they could recall particular speech passages where they found the tool most helpful, participants mentioned most often: (1) the *conference programme* (task code "CP"), given the high density of unknown named entities: "If I hadn't had the tool, I would have probably said either the persons' charge or their name but not both" (Sally); (2) *Number-dense passages* (task codes "NCR" and "NCN"): (3) *Unknown terms*, in particular in a list (task code "TL"). However, this is in contrast with the fact that at least two study participants reported that the tool was helpful in dense passages while they actually made severe errors.

A recurring theme in the interviews was the issue of trust. One participant commented in the interview that "trusting the tool is the prerequisite for using it. It's like having a GPS: you must trust it, otherwise what's the point of having one?" (Diana). In answering the question "what's the difference between the CAI tool and a human boothmate?", 3 participants declared that they would trust the tool more than a human interpreter: "The computer gives you the impression of being infallible. You can trust the human boothmate if you know her personally, but it does happen that you turn to your partner to request help and she gives you a blank stare back" (Sally).

Logan and Carlo were the participants who gave the lowest dependability scores to the tool. However, their low level of trust in the CAI tool had different grounds. Carlo showed an overall positive attitude towards technology, in general, and CAI tools, in particular. He reported in the interview to have tested and used several new technological solutions in the past and to be keen to integrate a tool similar to the one we tested into his workflow; he also showed understanding of how ASR and AI technology works and made predictions on the tool's performance based on his knowledge. Based on this knowledge, Carlo explained that he would need to "learn to trust the tool":

> I would trust the colleague to provide me with a hint on a specialised term I never saw before but not the tool because the latter cannot evaluate the adequacy of a solution in context. I would probably check the tool's suggestion after four consecutive assignments and if they consistently prove reliable, with time, I would learn to trust it. (Carlo)

Quite interestingly, Carlo was the only one of the participants who attempted to make selective use of the tool's prompts. When he did, he was observed looking away from the tool and using strategies such as abstraction and generalisation of the overall meaning of the passage whilst omitting the hint provided by the tool. When asked, in the interview, what motivated his behaviour he explained that he realised that he was having difficulty understanding the overall message and hence looked away, presumably to concentrate on his comprehension of the source speech beyond the individual problem triggers displayed by the CAI tool: "I did so because I knew that the aid was pointless if I couldn't understand the meaning in the first place ... The tool is helpful because I can ignore it" (Carlo).

Logan, instead, showed distrust and a low degree of familiarity with CAI tools and their functioning. Before agreeing to join the study, he asked whether such tools would one day replace human interpreters. In the interview, he mentioned the tool not providing the referent "diamond production" (whereas the item is not provided because it is not a technical term) as well as "Agenda 2063" appearing in the named entity column instead of the number column as two tool errors. He mentioned that, in several cases, he expected the tool to provide him with suggestions that did not come up and, in his view, impacted his delivery negatively: "my too-high expectations betrayed me" (Logan).

#### 7.3.2.3 Difficulty in using the tool

Despite the apparent simplicity and intuitiveness of CAI tools and the advantages reported by participants, most of them also pointed out difficulties and potential pitfalls in using CAI tools. Coming to the disadvantages reported by participants, three main themes emerged from our analysis: (1) ear-eye coordination, (2) CAI tool as prompt, (3) adjusting to CAI tool use. These three themes are explained in the discussion below.

##### 7.3.2.3.1 Ear-eye coordination

4/5 participants (all of them except for Carlo) reported that they had difficulty splitting their attention between listening to the source speech and using the

support of the CAI tool, which we will call *ear-eye coordination*. Participants explained this issue differently. Some of them reported a loss of grit and self-regulation: "It's just that I didn't put effort into listening anymore. The CAI tool made me lazy" (Logan). Another participant further elaborated on this same phenomenon: "because you know that the most difficult part of the sentence, the one that usually demands so much of your attentive control, is going to be provided by the CAI, you do not care of it anymore. For instance, in my case, I was not listening to numbers anymore, I just expected the CAI to do the work for me" (Sally). In other instances, participants provided an alternative explanation for the ear-eye coordination difficulty reported: "you must *learn* [italics added] to use it as a confirmation rather than your primary source of information ... you *must know how to* [italics added] distribute your attention between listening and looking at the tool" (Sally). Another participant reflected on the fact that the need to monitor one's interpreting process and output (e.g., in the case of plausibility errors) increases when a CAI tool is used: "the trade-off for the simplicity of use of the tool may lie in the need to monitor yourself more closely" (Diana). She went on to explain that "while without CAI an item that you don't know or hear is just gone, CAI puts you facing a choice."

#### 7.3.2.3.2 Tool as prompt

While, in theory, it is up to the interpreter to choose whether to use the CAI tool input or not, all participants except for Carlo reported that they felt prompted by the tool to say whatever they saw appear on the screen. Two participants defined the tool as "a temptation" (Diana, Sally) and added that they were tempted to use the tool's suggestion even in situations that would have better been dealt with otherwise, such as through the use of omission, generalisation, approximation and other similar strategies. The situations mentioned by participants were the following:

1. A high density of information (Sally), where a reduction in cognitive load was needed.
2. Redundant information, such as repeated numbers and examples (Sally), that participants felt could be left out strategically without compromising the overall message.
3. Loss of the overall meaning of the message, which makes the decontextualised suggestion unusable (Diana, Sally, Minerva, Carlo): "the tool was useful where I didn't need to understand the sentence, I just needed the

hint. In some cases, I had the hint but not the understanding and so I still needed to generalise" (Carlo).

4. Long décalage or fast delivery speed: "in the case of acronyms, although the extended form of the acronym is helpful to provide a complete and accurate rendition, it is a double-edged sword: it triggered me to say the whole acronym also when it was necessary, or I had better saved time" (Minerva).

Participants hence suggested that the tool should be used selectively, while failure to make goal-directed choices about how to use the CAI tool's suggestions may lead to mistakes.

#### 7.3.2.3.3 Getting to know the tool

Most participants expressed the need to "get to know the tool", or "to know what to expect from it". Participants explained the importance of knowing the tool as follows: "it would reshape your expectations: knowing how the tool works, knowing its limits, you also know how it can help you and you behave accordingly" (Minerva). With this expression, participants may have referred to several distinct concepts.

One meaning of "getting to know the tool" may be knowing where items appeared and how they would be displayed. One participant (Logan), for instance, explained that he expected the referent "diamonds" in the numerical task NCN to appear and was disappointed not to be provided with it by the CAI tool.

With this expression, one participant stressed the need for more hands-on experience allowing one to develop a practical feel and understanding of the tool's functioning: "one way is to read 'latency is two seconds' and another thing is to experience it while interpreting" (Carlo). Participants also expected that, through experience and repeated use, one would learn how to integrate the tool into the SI process more effectively, for instance by accommodating one's décalage to the tool's latency and using strategies to integrate the hint into one's rendition while waiting (Carlo, Diana, Logan).

As a further dimension of "getting to know the tool", participants engaged in reflection about the fact that the CAI tool did not help them process the speech semantically:

> Let's use the metaphor of a crutch: the CAI tool will be very helpful to walk, but it won't help you run! (Minerva)
> You are the one interpreting, which means understanding the link between

> pieces of information; the tool can only support you in the tasks that are challenging for the human brain. (Diana)

Participants reported a certain "disappointment" when they found out that the tool could not help them to the extent they had previously assumed:

> It was probably my high expectations that betrayed me. (Logan)
> If you're not understanding what the speaker is on about, and you expect the tool to be your "phone-a-friend lifeline", you will be disappointed! (Minerva)
> At some point in the speech, a series of technical terms came up [referring to the task "technical terms embedded in a semantically complex sentence", task code "TS"] and I was lost because I didn't know anything about the topic. I somewhat expected that the tool would help me out, but that obviously wasn't the case: CAI tools can provide you with terms but they do not stitch them together in a sentence for you! (Diana).

Possible errors and omissions of the tool were also mentioned as a negative factor. Participants were capable of recalling specific points of the test where they believed that the tool omitted an item or provided a wrong suggestion, which implies that CAI tool inaccuracies are strongly perceived by users. In regard to omissions, participants admitted that they got complacent after some exposure to the tool ("the CAI tool made me lazy", Carlo): "I expected problematic items to be displayed and so after a while, I stopped listening carefully to them; if they were not provided by the tool, I was dumbfounded" (Sally). In saying so, participants seemed to attribute the problem of CAI tool omissions leading to delivery omissions to their fault, at least partly. In contrast, errors seem to be much more negative for the tool's perceived dependability. Diana, who suggested that terms should be provided only in the target language to reduce the load of visual information, explained this point as follows:

> Trust is the fundamental prerequisite for using a CAI tool – it's like having a GPS: if you don't trust it, why use it? In the case of specialised terms in a real assignment, if a term is in my glossary, it is also in my head. If I consult the tool, it is because I need it: I need readily available, immediate support. I cannot waste time assessing the validity of the suggestion, if it's a term I've never heard before, I don't even have the knowledge to do so. I can only trust it. (Diana)

These difficulties are in contrast with participants' overall perception of the ease of learning of the tool, which was very high for most participants except for one. Participants who reported that little or no training is required to use the tool explained: "I don't think that there's much to be learnt: you literally don't have to do anything" (Logan). This may signal that novice CAI tool users may lack the awareness of their own difficulties and errors, which makes them unable to recognise their learning needs.

## 7.4 Usage problems and design recommendations

This first test iteration highlighted recurring error patterns that may be interpreted as usage problems, leading to recommendations for the improvement of SmarTerp. The first usage problem that was identified in this round of testing was that the order of appearance of items made it difficult for users to locate new information as it appeared on the screen. This usage problem had both high impact and frequency. 4/5 participants reported that the order of appearance of new items (ABCD, E⟶A, F⟶B etc.) appeared illogical and was confusing to them. The following is a representative quote from an interview:

> Why don't items appear in their logical chronological order? I would expect new items to appear at the top of the list with the others scrolling down so that I can check the previous item as well if I need to. The highlighting system is good, but items must appear in chronological order too. (Sally)

Participants explained that this feature decreased the tool's efficiency because they had to "look for bits of information on the screen" (Diana). A critical incident associated with the extra difficulty caused by the non-linear display system is the following.

> *Source*: His Excellency Paul Kagame; Honourable Soraya Hakuziyaremye, Rwanda's MINICOM Minister; Ms Giovanie Biha, UNECA Deputy Executive Secretary.
> *Delivery example* (Logan): His Excellency Paul Kagame; Soraya "Hakuziaremiei" [leans forward to read]; Giovanie Biha, Vice-Minister for Trade and Industry.

This feature seems to decrease both the perceived efficiency of using the tool and the actual effectiveness of the interaction. Given its high frequency and severe impact, we recommend that the order of appearance of new items on the screen

be changed into chronological order, with the new item appearing at the top of the list and the others already displayed scrolling down.

A further recurring error pattern, of medium impact but high frequency, was that users consistently interpreted people's names without the corresponding professional charge. This pattern could be interpreted both as a human error, given by the fact that interpreters could not distribute their attention effectively, and as a result of the fact that names were displayed by the tool but not charges, contrary to users' expectations. We decided that it would be appropriate to re-assess the problem after the second round of testing.

Finally, an issue of interest was the fact that both low accuracy and low recall impacted participants' delivery. Nearly half of the errors made by the CAI tool resulted in a severe semantic error in participants' delivery. At the same time, 93% of items (14/15) omitted by the CAI tool were omitted by study participants too. There are two ways to interpret these findings from the point of view of CAI tool development. On the one hand, one could notice that 93% is a much higher inaccuracy rate than nearly 50%. On the other hand, one could consider the impact of semantic errors as more severe than that of omissions. The severity of the issue may be further compounded by the lack of awareness of errors. A previous study observed that CAI tool users may "copy" tool errors without noticing them (Defrancq & Fantinuoli 2021). Also in this study, at least two participants reported that the CAI tool was "very useful" in the rendition of passages where they had committed severe errors. Finally, one may consider the crucial impact that wrong suggestions have on the perceived dependability of the tool: study participants tended to blame themselves for omissions made when the CAI tool omitted an item but blame the tool when it made an error. Considering all these aspects of the issue, it seems recommendable to continue following the principle "accuracy over recall" until disproved by future evidence.

## 7.5 Methodological validation

Along with conducting an initial formative evaluation of the tool, a purpose of the pilot study was to validate our research methodology and materials. The speech design was validated in that the tasks were challenging enough to prompt CAI tool consultation and the speech structure, alternating complex passages with passages not containing problem triggers, gave participants sufficient time to conclude the interpretation of a task before the next began, as expected. For the second iteration, we recorded another video with a more stable prototype and the updated CAI interface, i.e., after the design recommendations presented above

were implemented. An element of the study materials that we decided to revise before the second iteration was the questionnaire section asking participants to express their satisfaction with the support provided for individual problem triggers. In the interview, participants were not able to motivate the reason for this choice. For example, Logan and Carlo declared that they gave a low score to acronyms and specialised terms because "there were only one or two of them in the test speech". Since we did not find the outcome of these questions to produce reliable information, we decided to exclude them from the questionnaire in the main study and to dig deeper into users' perception of the usefulness of SmarTerp for the individual types of problem triggers through the post-test interview.

The interview questions allowed us to gain insight into participants' perspective, as we were expecting, and therefore the interview protocol remained unaltered.

Overall, the convergence of methods appears to be a strength of the study as it allows to mitigate the confusion that may arise from the interplay of several uncontrolled variables, related both to the machine side (e.g., CAI tool accuracy, functionalities, design features) and to the human side (idiosyncratic factors) of the interpreter-CAI tool interaction

# 8 Second iteration: Main study

The main study, conducted with 10 practicing conference interpreters, represents the second iteration of prototyping, testing and refinement of SmarTerp at a more mature stage of its development following the implementation of the design recommendations derived from the first test. After the pilot study, the order of appearance of items was changed into the scrolling list order with new items always appearing on top of the list. The prototype was also corrected to achieve ceiling performance (i.e. not to present the errors and omissions that characterised the first study). Only one omission involuntarily occurred in the task NIU. This chapter presents the results of the analysis of performance and perception data.

## 8.1 Users' performance

### 8.1.1 Task success rates

Table 8.1 provides the measures of central tendency in the dataset for the success rates for each task in the interpreting test. As expected, participants achieved a higher success rate on tasks of lowest complexity (AC, NE, NR, TL). NU represents an unexpected exception. The success rates are lower in tasks of higher complexity.

### 8.1.2 Error patterns

#### 8.1.2.1 Acronyms

Among the tasks of lowest complexity, AC (*isolated acronym*) is the one for which the highest success rate was registered. No specific error pattern was noticed in the rendition of this item, but the use of the visual aid varied across participants: five interpreted the extended version of the acronym only, four interpreted both the short and the extended version and one used the short version only. In the interview, participants explained that they found the support for acronyms advantageous because of the complexity of transcoding the sequence of digits from one language to the other. They also mentioned being provided with both the short and the extended version of the acronym as advantageous because this allows

Table 8.1: Success rates: central tendency measures (main study)

| | | % | | |
|---|---|---|---|---|
| Code | Task | Mean | Median | Mode |
| AC | Isolated acronym | 100 | 100 | 100 |
| NE | Isolated named entity | 90 | 98 | 98 |
| NU | Isolated numeral | 40 | 0 | 0 |
| NR | Numeral and complex referent | 90 | 95 | 95 |
| NIU | Numerical information unit | 79 | 92 | 92 |
| NCR | Redundant number cluster | 73 | 67 | 67 |
| NCN | Non-redundant number cluster | 55 | 64 | 67 |
| TS | Terms in a semantically complex sentence | 45 | 30 | 30 |
| TL | List of three unknown terms | 99 | 100 | 100 |
| TE | Isolated term | 90 | 100 | 100 |
| SO | Complex speech opening | 60 | 64 | 49 |
| CP | Conference programme | 80 | 79 | 92 |

the interpreter to select the most appropriate version based on the audience's background knowledge and their décalage. Only one participant was observed leaning forward and squinting at the screen to read the acronym more clearly.

> *Source*: The signing of the AfCFTA by the African Member States significantly strengthens the movement towards this goal.
> *Delivery example* (Laila): The signing of the [leans forward to read] AfCFTA, the African Continental Free Trade Area, is aimed at this goal.

#### 8.1.2.2 Named entities

While participants succeeded in reproducing *named entities*, both in isolation (task NE) and in combination with other problem triggers (tasks SO and CP), recurring errors were noticed, namely pronunciation errors and gender errors. Their frequency and context of occurrence are reported in Table 8.2.

The dataset contains 70 interpretations of named entities in total (7 per participant). Of these, 30 (43%) were mispronounced. The percentage of mispronunciations climbs to 70–90% for rare names. The frequency of pronunciation errors was high also for the rare element "praseodymium" (TL-1), mispronounced by 3 of 10 participants. Since the occurrence of mispronunciations does not depend

Table 8.2: Pronunciation errors

| | | Errors (/10) | |
|---|---|---|---|
| Code | Named entity | Pronunciation | Gender |
| SO-1 | Paul Kagame | 2 | 0 |
| SO-2-1 | Soraya Hakuziyaremye | 9 | 2 |
| SO-3-1 | Giovanie Biha | 2 | 4 |
| NE-1 | Felix-Antoine Tshisekedi Tshilombo | 8 | 0 |
| CP-1-2 | Tschisekedi Tschilombo | 7 | 0 |
| CP-2-2 | Kwesi Quartey | 2 | 0 |
| CP-3-1 | Victor Harison | 0 | 0 |

on the complexity of the source speech unit, the phenomenon may be ascribed to participants' difficulty in reading the item. Several participants were observed pausing, leaning forward, and squinting at the screen to read the named entity. In the interview, albeit acknowledging the usefulness of the transcription of named entities, participants reported difficulty in reading complex and long names, and two participants suggested that the mode of display should be adjusted to ease reading during SI. The reading difficulty may have diminished the overall effectiveness of the interpreter-CAI tool interaction and had a broader impact on the delivery than the mere rendition of the item. The delivery sample below exemplifies a recurring error pattern in the dataset, where the delay generated by having to pause and lean forward to read the named entity caused the interpreter to omit or misinterpret some following items.

> *Source*: Honourable Soraya Hakuziyaremye, Rwanda's MINICOM Minister.
> *Delivery example* (Jam): Soraya [leans forward to read] "Hazuziariame" [Minicom appears] Rwanda's ∅ Minister [male gender marker, gender error].

Coming to *gender errors*, these were moderate to high in two cases. Their occurrence may have several explanations. In some cases, it may be difficult to ascertain the gender by the name and because of the absence of gender markers in languages like English, whereas the target language requires the interpreter to make a choice. At the same time, errors of this kind were observed also when the gender was explicitly stated by the speaker, as in "Ms Giovanie Biha" (SO-3-1) interpreted as "Mr" by 4/10 participants.

#### 8.1.2.3 Isolated numeral

Another task of low complexity that registered a low success contrary to expectations is NU (isolated numeral). The mean success rate achieved by participants on the task NU was 40%, with 0% (in this case signifying a plausibility error) being the most frequent score. A delivery sample is provided below.

> *Source*: The continent currently has a gross domestic product of 3.42 trillion USD.
> *Delivery example* (Molly): The continent has a GDP of 3.42 billion dollars.

As explained in the discussion of the pilot study results, the order of magnitude trillion ($10^{12}$) was displayed as *bilione* in Italian, which is the correct translation but rarely used – common alternatives are: *trilione* ($10^{18}$), an incorrect translation but used with increasing frequency as a loanword from English, and *mille miliardi*, a correct and more native alternative but difficult to process during SI and to implement in the CAI tool. Six of 10 participants translated *bilione* (trillion) as *miliardo* and described the tool's suggestion as a mistake: "I don't understand how the tool got that wrong" (Molly). Seeing *bilione* on the screen, they probably associated it with the English order of magnitude *billion* and translated it into *miliardo* (the Italian order of magnitude for billion). Two of these participants used the tool's suggestion first and then wrongly corrected it into *miliardo*. Of the remaining four study participants, two opted for the Italian order of magnitude *trilioni* and only two accepted the tool's suggestion *bilioni*. Hence, albeit correct, the tool's suggestion seems to have been confusing and disruptive to interpreters, probably due to its low frequency of use. It must be noted that the statement "Africa's GDP is 3.42 billion dollars" corresponds to a plausibility error. Given that we can expect professional conference interpreters to be able to gauge, in normal conditions, the implausibility of this statement, the most likely explanation is that correcting the tool required too much of their attention for them to also check the plausibility of their delivery.

#### 8.1.2.4 Lists

The success rates on tasks of slightly higher complexity – when a series of terms are presented as a list (TL) or when a numeral is presented together with a complex referent (NR) – were high, with mean values between 90% and 100%. However, some recurring problems occurred in the neighbouring text, as in the example below, where while the information displayed by the tool was rendered accurately, the parenthetical information was misinterpreted. The parenthetical infor-

mation was rendered completely and accurately by only 1 participant, whereas 7 participants omitted it and the 2 remaining participants misinterpreted it.

> *Source*: This year, the market cap of AngloGold Ashanti – the largest mining company headquartered in Africa – was 12.13 billion USD.
> *Delivery example* (Lotta): This year, the market cap of AngloGold Ashanti – which represents the main headquarter – eh (.) nearly reached [numeral appears] 12.13 billion US dollars.

#### 8.1.2.5 Terms in semantically complex sentence

Problems were identified in the delivery when items displayed by the tool were connected by complex logical links. This is the case of the task TS, which presents three terms like TL but, differently from TL, the terms are not presented in the form of a bullet-point list but rather they are embedded in a more complex conceptual structure. The mean success rate dropped from 99% for TL to 45% for TS, with median and mode values dropping from 100% to 30%. Deliveries of TL break down into 2 correct renditions (100% success rate), 2 partial renditions (66% success rate), 4 generalisation strategies (30% success rate) and 2 semantic errors (0% success rate), where the interpreter's delivery was completely different in meaning from the source speech or nonsensical, as in the example below:

> *Source*: Furthermore, the porous high-rank coal matrix, with larger hydrocarbon-storage capacity, makes coal-bed methane reservoirs advantageous for commercial operations.
> *Delivery example* (Lotta): Furthermore, we must also consider an interesting coal matrix, hence, with a capacity of storage of hydrocarbon and coal-bed methane.

#### 8.1.2.6 Highly complex tasks

The occurrence of errors in the interpretation of the speech seems to have been more frequent in the most complex tasks, i.e. those characterised by high information density and the co-occurrence of several problem triggers in the speech unit. Mean accuracy rates lay between 60% and 80% for tasks presenting by the named entity-acronym-charge sequence (especially SO and partly CP) and 79% and 55% for numerical tasks (NIU, NCR, NCN).

The main recurrent pattern that was noticed in these tasks is participants' tendency to interpret the tool's suggestions and omit or misinterpret other components of the message not provided by the tool. These components were not

provided because of the tool's functions: they do not classify as problem triggers and hence would not be extracted by the AI.

This error pattern had a significant impact on the rendition of the named entity-acronym-charge sequence in the task SO (complex speech opening), where the speaker greeted conference participants. Study participants tended to omit the person's charge, which was not displayed by the tool, as in the example reported below. With some variations, this is a pattern that was identified in the delivery of every participant, as testified by the fact that none of them scored close to 90–100% on this task. Two main explanations of this phenomenon were provided by study participants in the interview. The first is that they expected that charges would appear. The second is the difficulty in sharing attention between the acoustic and the visual input due to the excessive latency, mentioned by two participants: "the latency was too high and so I didn't hear Rwanda at all" (Stella).

> *Source*: His Excellency Paul Kagame; Honourable Soraya Hakuziyaremye, Rwanda MINICOM Minister, Ms Giovanie Biha UNECA Deputy Executive Secretary.
> *Delivery example* (Stella): His Excellency Paul Kagame; Soraya "Hakuziaramie" from the Ministry of Trade and Industry; Giovanie Biha from the Economic Commission for Africa.

This error pattern was identified also in number-dense tasks. In NCN (non-redundant number cluster), the referent 'diamond production' was not displayed by the tool. 6/10 participants misinterpreted the referent and, consequently, the whole task. As in the example below, the error pattern consists of an omission of the referent resulting in either a sentence fragment or a misattribution of the arithmetical value to the previous referent.

> *Source*: Madagascar alone produced approximately 58,000 metric tons of nickel in 2021. Namibia's diamond production amounted to 2.52 million carats in 2018.
> *Delivery example* (Mermaid): Madagascar alone has produced 58,000 tons of nickel Namibia 2.52 million of nickel.

During the interview, one participant (Oscar) explained that he found the suggestions confusing because the item 'nickel' remained highlighted. It could be that the persistence of irrelevant stimuli on the screen, combined with participants' difficulty in sharing attention between the acoustic and the visual input, could be a design-related factor increasing the likelihood of error hence requiring optimisation.

#### 8.1.2.7 CAI tool omission

The test speech presented only one case of CAI tool omission. The omitted numeral was omitted by 6/10 participants and misinterpreted by 1 participant, as in the example below.

> *Source*: Analysts forecast that African production of LNG will increase by 150% from 28 mtpy in 2018 to reach 84 mtpy by 2025.
> *Delivery example* (Jam): According to forecasts, African production of LNG will increase by 150% to reach 28 million tonnes per year (.) next year and 84 million tonnes in 2025.

## 8.2 Users' perception

### 8.2.1 Post-task questionnaire

Table 8.3 reports the measures of central tendency for the post-test questionnaire answers.

Table 8.3: Questionnaire results: central tendency measures (main study)

| | Mean | Median | Mode |
|---|---|---|---|
| Satisfaction | 1.8 | 2 | 3 |
| Ease of use | 2.2 | 3 | 3 |
| Effectiveness | 1.9 | 2 | 3 |
| Ease of learning | 0.8 | 1 | 3 |
| Timeliness | 0.7 | 0.5 | 0 |
| Dependability | 1.4 | 1.5 | 3 |
| Likelihood to use CAI (before) | 1.4 | 1.5 | 2 |
| Likelihood to use CAI (after) | 2.4 | 3 | 3 |

The questionnaire results provide a measurement of participants' self-reported satisfaction with the tool and their evaluation of usability attributes. These values are shown in Figure 8.1, which represents the standard interpretation of scale means based on the UEQ.

The usability qualities that were attributed the highest value are ease of use and effectiveness. Ease of learning, timeliness and dependability obtained significantly lower scores. The questionnaire results also show participants' stated

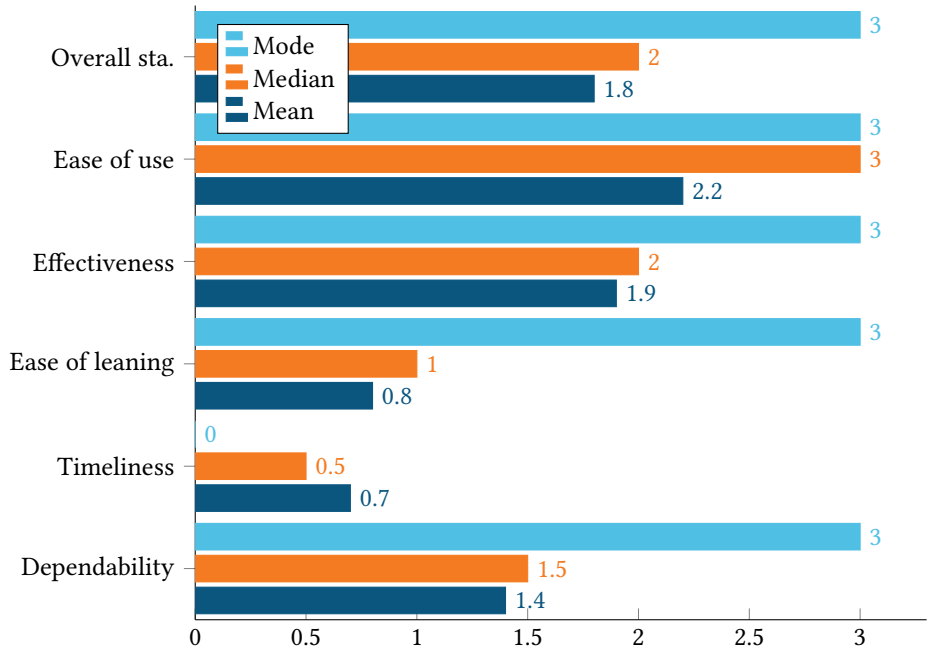


Figure 8.1: Questionnaire results: central tendency measures (main study)

likelihood that they would use ASR-based CAI tools in the booth in the near future. The self-reported likelihood after the test is compared with the judgment expressed by participants in the enrolment questionnaire, before testing the tool. It stands out that the self-reported likelihood of using an ASR-based CAI tool increased in 6 cases after testing SmarTerp and the mean value increased from 1.4 to 2.4.

### 8.2.2 Interviews

#### 8.2.2.1 SmarTerp's UI and technical specifications

In the interviews, several issues were raised by participants concerning SmarTerp's UI and technical specifications. Participants explained that locating items on the screen was a major difficulty in using the tool. One participant (Jam) reflected on the fact that this process should become automatic for the interpreter for the interaction to be as efficient as possible. Two UI design elements that were seen as facilitating this process were the chronological order of appearance of items (i.e. the scrolling list with new items placed on top, mentioned by

2 participants as a facilitating feature) as well as the highlight of new items or repeated items (mentioned by 2 participants). Other design features were seen as obstructive for the identification of relevant items on the screen. The most frequently reported issue (by 6 participants) is the need for repeated items to be placed at the top of the list. Participants explained that whenever they heard a problem trigger, they immediately looked at the top of the list; if they did not find the information there, "the eye had to wander" (Lotta), which in their view decreased the efficiency of the tool. This was the case when a repeated term was simply re-highlighted in its current position (e.g. third item from the top) rather than displayed on top of the list again. One participant reported that she felt disoriented by the fact that acronyms appeared in the terms module, in some cases, and in the named entities column, in other cases, based on whether the acronym abbreviated a term or a name. This forced her to "look for" the given item on the screen, which made her lose time and concentration. The fact that items that are not relevant anymore remain highlighted is another factor of complexity mentioned by some participants.

As far as the tool's technical specifications are concerned, most participants (8/10) regarded the latency as too high. 3 participants expressed the need to know where the terms come from, which influenced the perceived dependability of the tool. 2 participants suggested that adjusting the display of named entities and providing supporting information, in particular the charges accompanying named entities and the referents accompanying numerals, should help prevent some of the recurring problems identified in the deliveries and increase the effectiveness of the interaction. As far as the transcription of named entities is concerned, one participant suggested adopting a "syllabified coarse phonetic transcription" (Oscar) – syllabified to aid the identification of phonetic units and "coarse" because, differently from a scientific phonetic transcription, this should be easily readable for every interpreter. Another participant suggested adopting a "sound-like transcription based on the source language" (Laila), like the one used in publicly available ASRs systems (e.g. Google Cloud, Otter.ai etc.).

Participants' opinions diverged on some aspects of the interface, which may require more in-depth exploration or customisation. While 5 participants mentioned the division into columns as a strength of the tool, one participant saw this characteristic as an unnecessary complexity; she also predicted that this feature may become overloading in the context of speeches accompanied by PowerPoint presentations. Another controversial aspect was the usefulness of the source and target version of terms and acronyms, essential for one participant and unnecessarily complex for another.

#### 8.2.2.2 Perception of usability attributes

In the interviews, users were asked to justify their evaluation of the usability qualities of SmarTerp. Starting from the usability aspects that received the highest values in the post-test questionnaire, participants' evaluation of the tool's effectiveness and ease of use were highest. Based on the interviews, the potential of the tool to improve performance on problem triggers was regarded as the main factor making its use effective (mentioned by 6 participants):

> The tool can hugely increase accuracy: where in the past you would have used emergency strategies [e.g., approximation] because of our human limits (in my case, after 20 years in the booth I should interpret numbers with eyes closed, but that's absolutely not the case) the tool can help you achieve greater accuracy; if you combine your human ingenuity with the technical support, you can reach new heights. I believe that these tools will be compulsory in a few months. (Laila)

The automatic extraction and display of suggestions and the "intuitive" interface were mentioned as the main factors determining the tool's ease of use. However, some negative aspects were mentioned too as impairments to the tool's effectiveness and the ease of use. Amongst them, the most common are a sort of "distraction" caused by the tool and the need to "get used to it". Users' perception of these issues and their impact on the SI process is reflected in the following quote:

> The tool is very useful but also very complex to use: seeing a screen with moving items is not very natural. You must identify items on the screen and even a split second of delay can disrupt the interpretation, especially in highly dense passages. There is also a difficulty in dividing attention between the speech and the tool. Moreover, if you see the suggestions on the screen you feel compelled to say them but if you haven't understood the overall message, you can't do much with the suggestions. (Jam)

Similar positive and negative considerations were made by users on the ease of learning of SmarTerp. This is the usability attribute for which participants' opinions seem to diverge the most. Most users gave a +3 score to this usability attribute and explained that, in their view, one can start successfully using the tool immediately, without the support of a trainer, since it is "very self-evident: you get suggestions and what do you have to do? Absolutely nothing. It is foolproof" (Laila). Other participants referred again to the aforementioned drawbacks (getting used to the tool, the feeling of being "distracted" etc.) and added that the "intelligent" use (Jam) of the tool may require specific training.

The tool's timeliness was the attribute that scored lowest on all measures of central tendency, meaning that it was the attribute that received both the lowest mean value and the most negative evaluations. In fact, only two participants found the latency acceptable. The reason is summarised in the quote below:

> Too much time went by before the item appeared on the screen: waiting to see it, I lost concentration on what came next and something went lost [in the messages or the rest of the information]. You must remember that the usual speaking pace we are confronted with is very high. Even milliseconds can make a difference. (Stella)

Finally, participants were asked to evaluate the tool's dependability and explain what made them perceive the tool as reliable or unreliable. Participants who subscribed to the tool's reliability explained that they "expected" or "presumed" it to be accurate. Other users saw possible inaccuracies and omissions as major threats to the tool's reliability. The main factors that, according to users, should be evaluated are the adequacy of terminological suggestions (mentioned by 4 participants) and ASR performance (4) in the face of real-life challenges, such as different accents, non-native pronunciation, bad sound quality and a slow internet connection, in the case of on-site use of SmarTerp. It must be stressed that the tool's dependability seems to be a priority for users because, in the speed and complexity of SI, it might be too demanding to check the plausibility of the tool's suggestion:

> In some cases, I realised that the terminological suggestion provided by the tool might not have been the most adequate, but I did not have time to add my own version. Consider what happens when you use a CAI tool during SI: (1) you see the prompt, (2) you read it out, (3) while you read it, you assess its plausibility, (4) you don't have time to add an alternative solution because you have already committed to the CAI. (Molly)

> It was great to see all those long names and unknown terms on the screen. But I did not have time to examine them while interpreting. Since I had the suggestion, I wanted to use it, but I don't know if I would in a real-life assignment without knowing whether they are correct. What if they were wrong? (Laila)

Finally, 4 participants said that they expected items to appear which did not appear (as they were not meant to) and this negatively influenced their perception of the tool's dependability.

### 8.2.3 Drivers of users' satisfaction

The interview questions concerning participants' self-reported likelihood to use the tool as well as those concerning the major perceived advantages and disadvantages in the use of SmarTerp yield insights into interpreters' needs in the use of a CAI tool and make it possible to identify some possible factors that may drive their acceptance of such tools in the booth.

A first-time positive experience with the tool (which is commonly referred to as *user activation*) seems to be strongly influenced by the perception of substantial advantages in using the CAI tool. In other words, interpreters whose self-reported likelihood to use an ASR-based CAI tool increased after testing SmarTerp (6/10 study participants) declared to have changed their evaluation because the test made them realise that the tool has the potential to improve their performance beyond expectations, as reflected in the quotes below:

> My interest definitely increased: I had a chance to see what the tool can actually give you. (Mermaid)
> When I saw the opportunity to test the tool, I was sceptical at first. But after testing it, I am very impressed at what it can give you. (Oscar)
> When I filled out the enrolment questionnaire, I had just tried another technological tool that was meant to enhance consec [the smartpen] but that was a big disappointment and so I was a bit discouraged. But then I saw SmarTerp and it was love at first sight: it is evident that behind the tool there are people who know what the interpreting profession is about. (Lilla)

Other factors that contributed to users' activation and overall satisfaction are:

1. Improved *accuracy*, mentioned by 5/10 participants who claimed that the tool helped them reduce the number of errors and omissions.

2. Feeling of greater *security/self-confidence*, mentioned by 5/10 participants who defined the tool as "a lifeline" (Mermaid, Jam), "a parachute" (Lotta), "an umbrella when you're walking under the rain" (Oscar), and "a good boothmate" (Jam), who "is always there for you, when you're in need" (Stella). One participant explained: "I dread numbers, but this time I was 100 times more relaxed because I knew that I'd land on my feet" (Laila).

3. Reduced *effort*, mentioned by 3/10 participants who claimed that they felt less tired because they did not need to retain in memory or write down problem triggers during SI.

However, study participants (particularly but not only those whose self-reported likelihood to use an ASR-based CAI tool did not increase after testing SmarTerp) declared that they would need to assess several aspects of SmarTerp before they could decide to introduce it into their professional practice. Based on the interview data, we may expect that interpreters' continued use of a CAI tool (a factor known as *user retention*) may be conditional to the following needs:

1. Ascertaining the tool's absolute *dependability*, mentioned by 4/10 participants, who explained that the tool's dependability is the very prerequisite for using it: "you must trust that it is 100% reliable" (Lotta); "I wouldn't want a software telling me wrong things" (Stella). As explained earlier, dependability is a prerequisite both because of users' expectations and because it may be too demanding for the interpreter to check the plausibility of the CAI tool, as well as the source speech and the plausibility of their interpretation.

2. *Testing* the tool in the interpreter's *work domain*, a need mentioned by 3/10 participants, who suggested that the use of this tool may be more useful in specific interpreting settings or speech types rather than others; for instance, one participant explained: "in my field, a whole phrase made up of non-technical terms may represent a technical expression which must be rendered precisely into the target language; I doubt that the tool could help me in these cases" (Jam).

3. *Comparing* SmarTerp with other ASR-based solutions, mentioned by 3/10 participants.

4. Hearing other *users' opinions*, mentioned by 1 participant: "It looks like a promising development. However, before purchasing such a tool I'd need to consider a few factors (such as its cost) and wait for other interpreters' reviews and opinions of colleagues" (Toni).

5. Evaluating the *costs* of the tool, mentioned by 1 participant.

#### 8.2.3.1 Usefulness

Participants' reflections on the usefulness of SmarTerp to interpret problem triggers provide insights into users' expectations on a CAI tool. Users generally believe that the main advantages in using a CAI tool are increasing the accuracy and completeness of the rendition of problem triggers and reducing the mental effort in processing these difficult items:

> [Problem triggers] are the elements that most often slip away when we interpret, we do not understand or do not know how to translate. They are also the elements that take up the most space in your brain. (Lilla)

Through the support of a CAI tool, users expect to reduce errors and omissions in the interpretation of these items. In the case of acronyms, two participants added that being provided with the extended version of the acronym can increase the quality of the interpretation because the interpreter can provide additional information to the audience. At the same time, they expect that the CAI tool will alleviate some of the mental effort by offloading their memory or performing some processing steps for them. For instance, 4/10 participants declared in the interview that they commonly write down numerals when interpreting simultaneously but did not do so during the test because they expected that the tool would do so for them and reported that this way they felt more "relaxed" and "rested". Finally, some participants reflected on the fact that the tool can be helpful as a confirmation or "a litmus test" (Jam), i.e., in all those cases when they are not sure to have understood an item correctly.

These comments also offer insight into the problem situations in which a CAI tool is most likely to be perceived as useful by interpreters:

a. Interpreting items not found in preparation: although assignment preparation is a crucial aspect of interpreters' work, as one participant explained, "a specialised term you don't know can always come up" (Stella).

b. Dealing with highly technical assignments, which require a high degree of terminological precision by the interpreter.

c. Interpreting speech passages particularly dense in problem triggers because they are associated with mental effort and a high error rate; in this regard, three study participants mentioned in the interview that a human boothmate may provide wrong suggestions or fail to understand numbers and named entities and, hence, found the virtual boothmate more reliable for these items.

In general, we can expect that the CAI tool will be most useful in challenging situations. As one participant explained:

> The usefulness of the tool depends on the problem: a CAI tool is very useful for streams of numbers or numbers combined with a complex term. If the sentence to be interpreted is "population amounts to 150.000" you can make it by yourself. (Molly)

In the same way, participants reflected on the usefulness of the tool for particularly complex named entities, such as long names, unknown names, foreign names with a difficult pronunciation or names mispronounced by a foreigner.

In another section of the interview, study participants were invited to reflect on the differences between the virtual and the human boothmate. In general, they see consistency of performance accuracy (mentioned by 5 participants), readability (3) and availability (2) as a strength of the virtual boothmate compared to the human one. As explained by study participants, "human boothmates come and go, your CAI is always the same" (Toni); the virtual boothmate "gives you everything you need without complaining" (Oscar), "it is always there: it doesn't go on coffee break, it doesn't go to the toilet" (Lilla), "it doesn't get tired" (Jam), "it doesn't get distracted" (Laila). 4 participants stressed the objective user-tool relationship as an advantage: while interpersonal factors influence the relationship with the human boothmate, the use of the virtual boothmate is solely dependent on the interpreter's individual needs:

> It sometimes annoys me when colleagues are too helpful, but what can you do? It would be impolite to ask them to stop it or look away. The tool doesn't take offence if I ignore it. (Molly)

However, the tool is perceived as less reliable when it comes to terminology (2 particiapnts): the human boothmate is seen as a more reliable help because s/he can select the right term in context by virtue of his/her professional experience in the field. At the same time, the tool is seen as limited when it comes to providing help on the colleagues' overall understanding of the speech (2 participants).

When asked which context of use they would prefer for the tool (online, on-site or both), all participants replied "both". The general tendency is that of preferring online use because of the possible technical barriers to on-site use (unstable or unavailable interpret connection, cables required to connect the tool etc.). They also declared that in the context of combined use of human and virtual boothmate, they would prefer to use mostly the virtual boothmate and have the human boothmate serve as a "back up" helping in case of the tool's failure, checking the accuracy of the tool, monitoring the plausibility of the interpretation, and helping the interpreter when the overall meaning of the utterance is not clear to them.

#### 8.2.3.2 Difficulty in using the tool

During the test, study participants also perceived possible disadvantages in the use of a CAI tool, which negatively affected their overall satisfaction. The ones reported most often are:

a. *Failure to attend to the acoustic input*: participants explained that when they were looking at or waiting for the visual input, they lost their focus on the source speech. Because of this excessive attention on the visual input, participants explained that they failed to acoustically perceive elements in the immediate linguistic context, such as the charge associated with a person's name or the referent of a number, as shown earlier in the analysis of participants' performance.

b. *Loss of concentration on the overall meaning of the message*: one participant mentioned the fact that interpreters should always be able to "retell the story", i.e., to grasp the general meaning of the utterance rather than just transcode words; he then explained that a high degree of accuracy for specific items at the expense of holistic understanding may be a threat in the use of in-booth CAI support.

c. *Indiscriminate consultation*: a further risk may lie in making indiscriminate, non-strategic use of the tool, i.e., consulting it and relying on its input also when an alternative strategy may have been more effective. For example, Mermaid mentioned that she usually repeats the sound of unknown named entities; during the test, instead, she only looked at the tool and, in doing so, did not attend to the sound; since she found it difficult to correctly read out the long foreign names, she suggested that listening to and trying to reproduce the sound might have been a better strategy.

d. *Reliance-agency balance*: some participants mentioned that they expected the tool to help them, as for the participants who usually write down numerals but did not this time. This exposes them to problems if the tool is inaccurate or it presents limitations. For instance, one participant explained "sometimes I expected items to appear but they did not" (Oscar) and another commented, "if the tool helps you twice, you expect that it will help you a third time" (BCM). As discussed earlier in the report, this is reflected in the omissions of the item omitted by the CAI tool as well as the failure to interpret items not displayed by the tool because of its functions.

## 8.3 Usage problems and design recommendations

In this study, several usage problems were identified leading to recommendations for the improvement of the CAI tool SmarTerp. Below, they are divided into issues related to the CAI tool's general UI design features, the particular characteristics of how problem triggers are displayed, and the tool's technical specifications.

### 8.3.1 UI design

In the pilot study, it was noticed that the sequential order of appearance of items (ABCD, E ⟶ A, F ⟶ B, etc.) was confusing to users, who reported difficulty in locating relevant information on the screen. The SmarTerp developers hence changed the order of appearance to chronological (scrolling list with new items appearing on top). Repeated items were not repeated but only highlighted in their current position in the list. In the main study, two participants spontaneously mentioned the chronological order of appearance as "ideal". However, six users still found that the highlight of repeated items in their current position (e.g. third place from the top of the list) without moving them back to top represented an impediment to the detection of relevant input. In their view, relevant items should be always placed on top because that's the spot on the screen where they spontaneously directed their gaze when they heard or anticipated a problem trigger. The preparatory direction of the gaze onto the spot where items are supposed to appear might be one of the processes in CAI tool use that become automatic with repeated usage and make the interpreter-CAI tool interaction more efficient and effective. We, hence, recommend placing new items on top of the list to increase the tool's consistency favouring the development of automatic behaviour. However, the items should not be repeated to avoid unnecessary clutter. Two solutions are possible to avoid repetitions: (a) the repeated item moves from its current position to the top of the list and all the other items descend to fill the gap (which one study participant referred to as "Tetris mode"); (b) the repeated item swaps place with the item at the top of the list. The first option implies more movement on the screen, which may be either a distractor or a feature facilitating tool consultation thanks to the evident sign that a new item is being provided. While, at the current state of research, we cannot judge whether the scrolling of items would be advantageous or not, the first option allows us to keep the chronological order in the items on the screen: as the repeated item moves to the top, the second most recent item moves to the second position in the list. Since this mode of display is more consistent with the chronological order of appearance of new items, we recommend adopting this mode of display.

In order to facilitate the detection of items on the screen, it also seems recommendable to remove and/or de-highlight items when they are not relevant anymore. We currently do not know what the optimal length of permanence of items on the screen is. A solution could be to give users the option to decide whether they want items to disappear and customise the length of permanence (e.g. items disappear 30 seconds/1 minute/2 minutes etc. after they were pronounced). A safer option to implement may be to de-highlight items after a

certain time length (which may be customised too). In users' performance, some misattributions (i.e. interpretations in which a numeral is linked to the wrong referent) were interpreted as usage problems triggered by the fact that no longer relevant items remained highlighted. We noted that the referent constituting the previous numerical information unit persisted on the screen as a highlighted item, while the current referent had not been displayed. We hence hypothesised that the permanence of irrelevant highlights might be a factor confusing users.

By the same rationale, it seems recommendable to enable users to switch on and off the tool as well as individual modules to reduce clutter and satisfy interpreters' need for personalised help. Allowing users to customise the order of the modules within the CAI tool interface may also fulfil this purpose: users may benefit from the opportunity to place modules in the order that they find more logical or fields in the position that is most salient for them.

### 8.3.2 Display of problem triggers

The test confirmed some of the design team's hypotheses about the optimal display of individual problem triggers and confuted other ones, pointing to some optimisation potential. Our data suggest that not all interpreters may find the display of both source and target language of acronyms and specialised terms equally important – an idiosyncrasy which we observed both in the pilot and in the main study. Some interpreters feel that the source language adds reliability to the term. Others find the input excessive and superfluous since, as some argued, interpreters do not have enough time to compare source and target during SI. Hence, it seems recommendable to enable users to decide whether both source and target language or only one of both should be displayed. At the same time, users expressed the need to know the provenance of the terms displayed by the tool to gauge whether these are reliable or not. The UI may signal (through a colour code or an icon) whether the term/acronym comes from the interpreter's glossary or has been retrieved from other terminological sources, which was seen as a major factor affecting the tool's dependability.

When it comes to named entities, in the light of the high frequency of pronunciation errors, using an alternative graphic representation, such as a sound-like phonetic transcription (possibly syllabified) as suggested by study participants, should be explored in the future. The occurrence of gender errors also points to the fact that users may benefit from having access to additional information about the person, organisation etc. that is being mentioned. A possible option is introducing a pop-up window that opens upon mouse hovering displaying a

picture and some fundamental information about people that are mentioned, and possibly other named entities (such as places, names of organisations etc.) too.

Finally, coming to numbers, our observations concern the mode of display of numerals as well as the number of components of the numerical information unit which are provided. In designing the tool, we decided to display digits as Arabic numerals (with target-language punctuation) and provide the order of magnitude, if above 'thousand', as a target-language word. This choice was aimed at supporting interpreters in the last phase of numeral processing, which is recoding from Arabic graphic code into target-language phonological code. Compared to previous studies, which displayed the whole numeral in the Arabic code (e.g. Canali 2019) or a combination of Arabic digits and source-language orders of magnitude (e.g. Pisani & Fantinuoli 2021), in our study, no syntactic errors were identified for orders of magnitude 'million' and 'billion', which supports the effectiveness of our design strategy. However, a major problem was identified in participants' recoding of the order of magnitude 'trillion' (task NU, isolated numeral). Participants had difficulty gauging the reliability of the suggestion and either corrected it wrongly or chose an alternative translation. Note that simply replacing the order of magnitude with another terminological alternative would not represent a definite solution to the problem. The problem that may occur in all languages in the translation of rare orders of magnitude, especially where the target language does not present a univocal translation and one of the solutions may cause ambiguity under the influence of source-language interference. While it is difficult to propose a definite, one-size-fits-all solution, several options should be tested, and perhaps the alternative that is gauged as correct and unambiguous by most interpreters should be chosen. Furthermore, it seems necessary to explain to interpreters how orders of magnitude were translated. Coming to the amount of information displayed, the number was displayed together with the following element in the same item box, which typically is the referent or the unit of measurement. The initial hypothesis was that it might be ideal to display both referent and unit of measurement together with the numeral to provide the interpreters with the core of the numerical information unit. However, this is currently not possible because the syntactic position of NIU components may vary, and these are not always problem triggers recognised by the AI engine. After having observed that recurrent and severe errors occurred when interpreters were not provided with the referent, the research team wondered whether an alternative mode of display (e.g. a running transcript) might be better for numbers. Another option to avoid overloading the interface could be having a pop-up window with the transcription of the sentence containing the numeral open at the

click of the mouse or through mouse hovering, so that interpreters may be able to selectively look at the broader context in which the item occurred.

### 8.3.3 SmarTerp's technical specifications

Coming to SmarTerp's technical specifications, a first reflection pertains to the decision of whether to favour precision over recall, as recommended by Fantinuoli (2017). Fantinuoli hypothesised that "if a priority has to be set, precision has priority over recall, in order to avoid producing results that are not useful and may distract the interpreter". The fact that study participants expressed the need for the tool to be 100% reliable (dependability was defined as a prerequisite for the adoption of the CAI tool) may be considered as empirical evidence for this principle.

The excessive latency was pointed out by study participants as a major shortcoming of the CAI tool. In their view, error patterns such as the failure to perceive other components of the message were, at least partly, caused by the excessive latency. It could be that latency places a burden on interpreters' working memory (cf. Cowan 2010). If the interpreter waits to see the item appear on the screen to produce the target speech, his/her working memory may become overloaded by retaining understood but not-yet-interpreted items. This may either cause inaccuracy in the interpretation of the already-processed speech segment or make it impossible to listen and understand the subsequent speech segment. In the test, all items appeared at constant two-second latency, which, at the time of writing, is the lowest possible latency achieved by CAI tools. Even if technological advancement makes it possible to further reduce latency in the future, it may not be expected that this will consistently be lower than 2 seconds. Rather, it seems recommendable to train interpreters on how to effectively adjust their décalage to use the CAI tool as productively as possible.

In observing study participants interact with the CAI tool, I noticed two distinct approaches to adjusting décalage to the CAI tool latency (see example below). By the first approach, the interpreter interpreted the already-understood source speech segment without waiting for the item to appear. When the item did appear, s/he integrated it into her delivery. By doing this, décalage was close to the source speech rather than dictated by the tool's latency. By the second approach, the interpreter followed the CAI tool and waited for items to be displayed to start producing the source speech segment. décalage was hence dictated by the tool's latency. Interpreters who successfully implemented the first approach (*following the speaker, not the tool*), were the best performers in complex and dense speech passages (e.g. SO, NCR, NCN). Interpreters who adopted

the second approach (*following the tool, not the speaker*) were generally less successful at coping with complex speech passages. Their delivery was often characterised by errors in the second or third sentence, possibly because of memory overload. If this hypothesis gets confirmed by future evidence, it could mean that interpreters' décalage strategy may offset a possible disadvantage derived by CAI tool latency.

> *Source*: In 2019, Africa produced nearly 8.41 mbd [million barrels per day] of oil. Madagascar alone produced approximately 58,000 metric tons of nickel in 2021. Namibia's diamond production amounted to 2.52 million carats in 2018.
>
> *Delivery example 1* (interpreter following the speaker): In 2019 Africa [2019 appears] produced about 8.41 million barrels per day of oil. Madagascar alone produced 58.000 tons of nickel in 2021. Diamond production in Namibia amounted to 2.52 million carats in 2018.
>
> *Delivery example 2* (interpreter follows the tool): Africa produced, as far as oil is concerned, [Madagascar appears] 8.41 million barrels per day, as far as Madagascar is concerned also 58,000 tons [Namibia appears] in Madagascar. Instead, as far as Namibia is concerned, we talked about 2.56 million increase of nickel production.

# 9 Discussion

The analysis of usability test data in Chapters 7 and 8 yielded practical recommendations for the improvement of SmarTerp's UI. Taking SmarTerp as a case study of interpreter-CAI tool interaction during SI and interpreting the findings against the backdrop of previous research, this chapter discusses the scientific contribution of this work. First, the chapter summarises the study's contribution to the field's current knowledge of interpreter-CAI interaction, both from the point of view of understanding of users' performance and of users' perception of CAI tools. Afterwards, the chapter discusses possible evidence about the need to train novice users of CAI tools. In a next step, based on the insights gained from this study, the chapter suggests (albeit only tentatively) general heuristics of CAI tool UI design and summarises open UI design questions. The chapter then discusses the limitations of the work, traces future research trajectories and, finally, provides methodological recommendations for future studies wishing to use the methods developed in the present work.

## 9.1 General principles of interpreter-CAI tool interaction

### 9.1.1 Users' performance

#### 9.1.1.1 Mediating variables affecting users' performance

This study found that several mediating variables external to the CAI tool's UI design can influence the outcome of interpreter-CAI interaction. The first is *CAI tool accuracy*. The data gathered in this study suggests that an inaccurately displayed item or an omission of the CAI tool is likely to trigger either an error or an omission in interpreters' delivery. For instance, CAI tool omissions corresponded to delivery omissions in 93% of cases (i.e. across 3 cases of omission interpreted by 5 participants) and in 60% of cases in the main study (in one instance of CAI tool omission interpreted by 10 participants). The second mediating factor is *task complexity* (a variable that had not been accounted for in the design of previous studies, cf. Chapter 4). In the present study, task complexity was defined by the density of problems triggers in the speech passage to be interpreted

and the complexity of the semantic relationships connecting them, whereas other potential factors of complexity, such as complex syntax, were controlled for (cf. Chapter 6). High task complexity was associated with much lower success rates and with the occurrence of severe semantic errors. This means that, while one isolated problem trigger in a simple sentence is likely to be accurately rendered with the support of the CAI tool, speech passages dense in problem triggers or where problem triggers are connected by complex logical relations (requiring greater analysis effort from the interpreter) are more likely to be misinterpreted. A further mediating variable is *CAI tool latency*. The large majority of study participants perceived the tool's latency as excessive. They ascribed errors in their interpretation to "having to wait for the tool", which, in their view, caused them to forget or fail to process other elements of the unfolding messages. Given the intense use of working memory (WM) during SI (e.g. Cowan 2000, Mizuno 2005), a possible explanation for the supposed impact of latency on interpretation lies in WM overload. It is possible that if the interpreter waits too long for the tool's input to start interpreting the items held in working memory, WM saturation is reached so that already-processed items disappear from memory, or it is not possible to process additional information. Future research should confirm the impact of these mediating variables and possibly identify other significant ones exerting an influence on CAI.

#### 9.1.1.2 Level of performance affected and error analysis

In the present analysis, errors were detected extending the analysis beyond the level of the isolated term/numeral and including larger units of analysis (the sentence, the coherence and cohesion of the speech passage, the plausibility of the message etc.), i.e. adopting a "communicative approach" to the analysis of deliveries (cf. Chapter 6). This is a major difference of this study compared to previous empirical CAI research, which focused on the accuracy of interpreted numbers and terms only, without focusing on the overall meaning of the interpreted message (cf. Chapter 4). A significant proportion of errors in participants' deliveries was detected analysing the delivery beyond the mere problem trigger, which means that CAI tool use may lead to overall incorrect delivery even when individual items were correctly rendered. Examples of errors detected at a communicative analysis were problem triggers that were accurately rendered but wrongly contextualised in the delivery, omission or misinterpretation of information accompanying the problem trigger or implausible delivery.

#### 9.1.1.3 Effectiveness and efficiency of the interaction

In usability studies, efficiency and effectiveness are typically evaluated with different metrics (such as success rates for the former and time on task for the latter) and regarded as two distinct concepts: while an interactive system may not be very efficient (i.e. sub-optimal in terms of the time and effort investments required of users to complete the task) it may still be evaluated as effective (i.e. it allows users to complete the task successfully). In the case of in-booth CAI tools, the distinction of efficiency and effectiveness as two separate concepts seems not to be as neat. Given that SI is a complex cognitive activity, the slightest interference may disrupt task execution. Hence, a decrease in efficiency, producing a delay in the delivery and causing the cognitive load to increase, impacts the tool's effectiveness. For instance, study participants claimed that excessive CAI tool latency delayed their interpretation and contributed to their failure to process other elements of the unfolding speech.

### 9.1.2 Users' perception

#### 9.1.2.1 Perceived usefulness

The perception of the CAI tool's effectiveness (described by study participants as its contribution to a complete and accurate rendition of problem triggers, reduced effort in the interpretation of these elements and a greater feeling of security) seems to have been a major driver of user activation, i.e. a positive first-time experience with the CAI tool. In other words, in order to perceive the tool as satisfactory, users need to feel that the tool helps them achieve better outcomes than they would by themselves and with less effort. To achieve this goal, the tool must be effective not just in easy problem situations, for instance, an isolated numeral in a simple sentence, which we may expect professional conference interpreters to be capable of solving themselves, but rather in the most complex ones, such as the interpretation of long and complex named entities, high number density and the co-occurrence of several problem triggers. Study participants explained that complex tasks are the ones where they were most in the need of CAI tool support.

#### 9.1.2.2 Perceived dependability

Issues in interpreter-CAI tool interaction may have interfered with the other sub-processes such as monitoring the plausibility of the own delivery and checking the plausibility of the tool's suggestions. Participants stressed that during

SI "even a split second makes a difference" and reported that, in this context, they found it difficult to monitor both themselves and the tool. This perception is reflected in the high rate of delivery errors corresponding to CAI tool errors in the pilot study (48%, across 5 cases of CAI tool error interpreted by 5 participants). Interpreters perceive the CAI tool as a source of immediate and reliable help when in need. Hence, they regard the tool's dependability as a fundamental prerequisite for them to adopt the tool in real life.

#### 9.1.2.3 Perception of human vs artificial boothmate

The ambition of ASR-integrated CAI tools is to represent an ideal virtual boothmate. To fulfil this role, the virtual boothmate should possess some characteristics of the human boothmate perceived by interpreters as "ideal" – although the help provided by the virtual and the human boothmate currently are and are likely to remain different in nature, as acknowledged by our study participants too. The study participants described a "good" boothmate as one who is available, reliable and knows what type of help the individual interpreter needs. Participants spoke of trusted colleagues who know their interpreting style, their preferences and needs well and provide help accordingly. The personalisation of help is regarded as crucial because excessive or unnecessary input may be disruptive during SI. Therefore, it seems recommendable for a CAI tool to provide sufficient customisation options for the individual interpreter to adjust the amount and type of support received to meet their individual needs. However, because users most commonly do not use the customisation options available in the tools they use, it seems recommendable to instruct CAI tool users on how to adjust the options to best suit their needs.

## 9.2 Training needs of CAI tool users

While ASR- and AI-powered CAI tools address existing needs of interpreters, which makes them potentially very useful, error patterns recurrently occurring in interpreters' delivery warn us from taking the success of CAI tool use for granted. Despite the apparent simplicity and intuitiveness of CAI tools, achieving an effective integration of CAI tools into SI appears to be a complex task. Such complexity risks to offset the potential gains of utilising these tools, if users are not instructed to use them appropriately. At present, both the content and methods of CAI tool training (i.e. *what should we teach?* and *how should we teach it?*) remain to be defined.

All study participants in this study had completed an e-course introducing them to the UI features and technical specifications of SmarTerp before taking the test. They had had a chance to practise on the CAI tool in an interpreting exercise of equal length, complexity and structure. However, several errors and problems occurred anyway in their delivery. This suggests that the content of in-booth CAI tool training cannot be reduced to mere information about the tools and unguided practice.

While further research is needed to precisely define the content of CAI tool training, some issues that emerged in the study may point to potential learning needs. It must be stressed that not all study participants were aware of their learning needs. Some participants contradictorily claimed that "no training is needed to use a CAI tool effectively" but, at the same time, they reported several difficulties in the use of the tool, such as "getting used to it" and developing specific strategies to integrate it into SI. Some users also claimed that they did not think the tool had any negative impact on their delivery also where they made considerable errors. Developing awareness for potential problems in CAI tool use and analysing one's performance may be a first learning need to be addressed in training.

Drawbacks of using the CAI tool that were reported by study participants, and may point to learning needs, include:

- Difficulty in "ear-eye coordination", i.e. attending to both the CAI tool and the speaker simultaneously.
- Loss of concentration on the overall meaning of the message, in favour of an excessive concentration on the problem trigger.
- Being prompted by the tool, i.e. interpreting an item that participants saw appear on the screen as an impulsive, immediate reaction to the visual input, although participants were aware of not having understood how to contextualise the item.
- Indiscriminate consultation, i.e. consulting the tool also when an alternative strategy may have been more effective.
- Knowing the tool enough to formulate realistic expectations about what items will be displayed.
- Reliance-agency balance, i.e. striking a balance between using the aids provided by the tool and remaining vigilant.

## 9.3 CAI tool UI design

### 9.3.1 Tentative heuristics

The results of the usability tests conducted in this study point to some heuristic principles that may guide the UI design of CAI tools. Although these must be corroborated by further evidence, they may provide hypotheses for future usability-focussed studies on CAI.

#### 9.3.1.1 Display numerals as a mix of Arabic digits and target-language orders of magnitude, if larger than thousand, but watch out for rare orders of magnitude

Previous studies on the CAI of numbers displayed numerals entirely in the Arabic code or using a mix of Arabic code for digits and *source* language phonological code for orders of magnitude above 'thousand' (Canali 2019, Pisani & Fantinuoli 2021). Presented with this graphic representation of numerals, users still made some transcoding errors in the rendition of orders of magnitude (e.g. 'billion' ⟶ 'million'). In this study, numerals were presented using a mix of Arabic code for digits and *target* language phonological code for orders of magnitude above 'thousand'. The fact that no transcoding errors for orders of magnitude until 'billion' were found in the dataset validates this design principle. However, the order of magnitude 'trillion' represents an exception. This order of magnitude is rare and its translation into Italian may be ambiguous. Therefore, the solution we chose confused users and caused severe errors in the delivery. The TL translation of rare orders of magnitude in in-booth CAI tools should hence be carefully chosen and its clarity should be verified with users.

#### 9.3.1.2 Make the UI interface as consistent as possible

Actions on the UI interface (e.g. where new items appear on the screen) should be as consistent and predictable as possible. An issue recurrently mentioned by study participants during the interviews is a difficulty in "finding information on the screen". The activity of locating relevant information amongst all other irrelevant stimuli on the screen is a cognitive process known in psychology as visual search (Davis & Palmer 2004). Our study participants reported that the additional effort and delay caused by non-automatic visual search caused them to focus excessively on the visual input and fail to attend to the acoustic input. A user interface that is maximally consistent should ease the development of automatic search behaviour.

#### 9.3.1.3 De-highlight irrelevant items

In order to facilitate users' identification of relevant items on the UI, it seems recommendable to de-highlight items that are no longer relevant.

#### 9.3.1.4 Favour precision over recall

Because CAI tool dependability seems to be a fundamental need of users and some study participants even claimed that they were unable to check the accuracy of the aids during SI, it seems recommendable to favour the precision of displayed aids over recall, as suggested by Fantinuoli (2017).

#### 9.3.1.5 Signal the origin of specialised terms and acronyms

Users were mistrustful of the terms suggested by the CAI tool. In the case of CAI tools that search for terminology in external sources (e.g. electronic dictionaries, databanks etc.), it seems necessary to signal the origin of the displayed term (e.g. via colour-codes or icons) so that users may decide whether to use the information at a quick glance.

### 9.3.2 Open design questions

In this study, we could not find any empirical evidence for some principles that guided the design of SmarTerp. It is still to be demonstrated empirically whether these UI features ensure that CAI tool interface is usable. Each of these critical features could be the object of a dedicated study.

The first open question relates to the separation of problem triggers into distinct interface sections ("modules"). Study participants' opinions of this design feature diverged. Some users found that the division into modules makes the UI better organised and more consistent. Other users found that it made the interface excessively and unnecessarily cluttered. The impact of this UI feature could be explored comparing different interface options in comparable and controlled tasks (A/B testing).

The second open question is about the display of terms and acronyms both in the source and target language. Users' behaviour did not point to any use of the source-language version of the displayed item. This could depend on the artificial nature of the study: interpreters may have taken the accuracy of items for guaranteed and not performed the accuracy check that they would normally perform in a real assignment. However, some users commented that they would

have never had the time and concentration to check whether the tool suggestion was accurate.

Open questions also relate to the display of named entities. Participants often mispronounced them or misinterpreted their gender (in the case of people). Furthermore, the excessive concentration required to read out the named entity from the screen recurrently led to errors or a loss of other fundamental information components in the following sentence segment. Pronunciation errors are more likely to occur in the interpretation of complex named entities of a foreign language that the interpreter does not master. Gender errors may be more likely when languages that are gender-neutral in spoken speech (such as Mandarin Chinese) are interpreted simultaneously into languages that are gender-sensitive. To cope with pronunciation errors, some study participants suggested using not the official written form of named entities but a sound-like transcription similar to that produced by publicly available ASR systems (like Otter.ai or Google Translate). One study participant (Carlo) suggested adding a "Netflix-style pop-up" displaying additional information about the people mentioned, such as a person's picture, gender, age etc.

A further question related to interpreters' difficulty in interpreting not just numerals but the whole numerical information unit correctly. A question is whether displaying the transcript of the sentence in which the numeral occurs in response to mouse hovering over the numeral might represent a more effective support for interpreters rather than the isolated number.

Both in the display of named entities and numerals questions related to the amount of information provided to users and whether more information should be made accessible when users request it through an interaction with the system (e.g. hovering over or clicking on the item). There seems to be a parallel with intelligibility features in translation memories, that allow users to find out more about the aid provided to them.

## 9.4 Limitations

This study has several limitations. From a general scientific perspective, the study design was adequate to account for some possible mediating variables, that were incorporated into the design of the test speeches, but not all possible variables. Further variables such as the impact syntactic complexity, delivery pace, language combination, remain currently unexplored. Furthermore, given the artificial nature of the test, an evaluation of CAI tool use in a real-life assignment may reveal further insights and, possibly, yield a better understanding of actual users' needs.

From a usability perspective, the main limitations of the work derive from the small sample size and the broad exploratory character of the inquiry. A larger sample size is required to give more robustness to the design recommendations that were developed within this work. Furthermore, to develop more robust UI design heuristics, focused studies are required to zoom in on specific interface principles, for instance through usability testing.

Given these limitations, the UI design heuristics and the principles of interpreter-CAI tool interaction that were identified based on the interpretation of the study findings should be regarded as initial hypotheses requiring further exploration

## 9.5 Future work

Empirical CAI research is still in its infancy and several research questions remain to be addressed by future studies. Future work may have three major orientations: scientific, pedagogical, and usability. The *scientific orientation* consists in exploring the CAI-tool supported SI to contribute to the field's scientific knowledge. The use of a CAI tool adds a further element of complexity to the already complex cognitive task of SI. Therefore, this may represent a vehicle to increase understanding of SI from a cognitive perspective – and, possibly, be of interest to cognitive psychologists. More in general, scientific orientation may be seen as the basic research providing the fundamental knowledge applied in the usability and pedagogical orientations. A major and necessary contribution to basic scientific knowledge may be obtained pursuing the following objectives: (1) Developing a cognitive model of interpreter-CAI interaction that may explain which structures are activated during CAI-tool assisted SI and why errors occur; (2) Defining the impact of CAI tools on users with varying interpreting expertise (e.g., students vs professionals); (3) Exploring the impact of specific classes of problem trigger and identifying moderating variables; (4) Ascertaining the psychophysiological impact of CAI tool availability, for instance, exploring the hypothesis that it may reduce stress.

The *pedagogical orientation* comprises all research conducted to define the content and methods of instructional interventions on in-booth CAI tool use in a scientific and systematic way. This orientation is of interest both for the development of CPD solutions for professionals and the training of new generations of interpreters. To advance towards the development of research-based solutions, the following research gaps should be filled: (1) Modelling the skills and knowledge structures underlying effective interpreter-CAI interaction; (2) Defining effective instructional strategies to train those skills.

Finally, the *usability orientation* aims at developing recommendations for the further development of CAI tools, identifying the optimal UI features and technical specifications. Some research gaps that should be filled to develop this orientation are: (1) Developing research-validated tools and measurements: for instance, usability studies rely on a number of validated questionnaires; using these tools across studies increases their construct validity and allows for comparability and replicability; (2) Developing industry benchmarks allowing to put the evaluation of a CAI tool in perspective; (3) Exploring the impact of particular interchangeable UI elements, which are currently selected based on the personal intuition of interpreters/designers with little scientific justification. These are, for instance, the use of a running transcript vs isolated problem triggers, unitary field vs division of suggestion into modules etc.

## 9.6 Methodological recommendations

The present study represents an example of usability-oriented empirical CAI research. Future studies may build on the methods used in the present work to evaluate the usability of CAI tools. I recommend that these studies consider the following methodological recommendations:

1. *Participant selection*: to develop valid recommendations, participants should be representative of the target users for which the product was designed. If non-representative users are selected for convenience or to achieve other scientific aims, this should be specified in the limitations of the study; in this case, usage problems and possible solutions should serve as hypotheses rather than definite design recommendations.

2. *Sample size*: given the time-intensive nature of usability tests, it is common to have small samples, and it is believed that 5 participants are sufficient to detect major usage problems (Barnum 2020, Nielsen & Landauer 1993). However, the limitations of a small sample should be acknowledged. If a quantitative research question and statistical validity are pursued with a small sample, an experimental design collecting a large number of data points on the impact of one specific variable may be preferable to a usability test. Alternatively, other HCI testing methods should be considered.

3. *Test speech design*: because usability tests should involve tasks representative of real-life challenges that users would overcome through the use of the product, a sufficient degree of experimental control is required on

all actions that users perform with the product during the test session to evaluate the usability of the tool on those tasks. In the evaluation of a CAI tool, the design of the test speech is crucial because it is the source of the tasks that users accomplish with the support of the tool. In the limitations of the study, it should be specified that this design strategy represents a limitation to the ecological validity of findings. We recommend specifying the precise characteristics of the test tasks and, if possible, disclosing test materials to encourage scientific scrutiny and replicability.

4. *Approach to the analysis of deliveries*: the identification of critical incidents and usage problems through a process of abstraction requires the analysis of interpreters' performance. A micro-analysis focusing solely on the interpreted problem trigger may be useful to respond to specific research questions. However, a broader analysis of the interpreting product is required to infer the impact of the CAI tool on the interpreting process and derive considerations on possible tool-related problems. It is recommendable to adopt an analysis framework that allows for such an in-depth nuanced analysis – which we obtained adopting a communicative approach and an adaptation of the Redundancy Ladder Model (Frittella 2019a). If a micro-unit of analysis is used in the study (e.g., only interpreted numerals or specialised terms are analysed without) it should be specified that the results of the analysis are not reflective of broadly-conceived "delivery accuracy/quality" and are not measurements of "CAI tool effectiveness".

5. *Inferring the cause of critical incidents*: human factors always influence the outcome of users' interaction with a product. This may be particularly true of SI – a cognitively taxing task of bilingual communication performed by human interpreters in real-time and under time pressure. The influence of contingent and idiosyncratic factors represents a major potential threat to the usability testing of a CAI tool as a critical incident may be caused by multiple factors. In my study, I found it particularly helpful to integrate observations and performance metrics with interview passages in which participants explained why, in their view, particular critical incidents occurred. I considered their explanation as a form of participant triangulation strengthening the reliability of my interpretation.

6. *Interpreting users' recommendations*: while study participants' recommendations may be useful to surface their perceived problems and needs, they should be evaluated in the light of a whole range of possible biases: users

may be unable to locate or adequately explain a problem, they may misattribute a problem to the wrong cause, report that they want something which actually does not work in practice etc. In sum, while it is important to listen to users' opinions, one should not "ask users to be designers" (Barnum 2020). Their recommendations may be taken as a starting point for further exploration or a source of participant triangulation but not as an infallible source of design recommendations.

7. *Validity of findings*: recommendations for the design and development of a product through usability testing are a powerful tool to improve a product. Nonetheless, it should be made clear that these are based on small-scale studies which may have yielded only a partial understanding or biased view of the issue. Like all scientific research, the results should be held as valid only until disproved by further evidence and scientific scrutiny should be encouraged through a transparent and well-documented process of (a) formulating research questions on UI design elements and technical specifications of the tool, (b) collecting adequate data to explore these questions, (c) providing a detailed, objective description of findings, (d) and explaining researchers' interpretation of underlying usage problems.

8. *Generalisability of findings*: the results of small-scale inquiries are, by their own nature, not generalisable in a statistical sense. However, we suggest that the concept of analytical generalisation (Yin 2013) may apply to the design principles identified in usability tests. Analytical generalisation involves making projections about the results of a study not to a population but to a theory. Evidence for the applicability of a principle gathered through multiple studies strengthens the generalisability of the given principle to other similar cases. Usability testing was used in the present work in line with the aim to derive recommendations on the tool's UI design and technical specifications. Future work may focus on applying other user research methods in line with different aims and research questions, such as contextual inquiry and focus groups to deepen understanding of users' needs, A/B testing to compare two different designs etc.

# 10 Conclusion

The interpreting profession is facing a phase of change due to technological advancement. CAI tools are one such technology that has the potential to become deeply entrenched in the way SI is performed. Thanks to the recent introduction of ASR- and AI-technology into CAI solutions, it is now possible for conference interpreters to utilise these tools during the complex cognitive task of SI to receive automated support for the rendition of particularly problematic items, known as "problem triggers". The occurrence of acronyms, named entities, numbers, and specialised terms in the source speech systematically corresponds to higher-than-normal rates of omissions, generalisations, approximations as well as severe errors and is thought to increase the processing requirements of the interpreting task. ASR-integrated CAI tools, representing an artificial boothmate for interpreters, have the potential to increase accuracy and alleviate some of the mental effort in processing these problematic items. However, CAI tools are used during a task, simultaneous interpreting, that is already extremely complex from the point of view of the numerous sub-processes taking place concurrently in the mind of the interpreter. Details matter in the design of any user interface, where seemingly small features can significantly impact users' performance. This is all the truer in the case of a tool designed to be used in a task that is as cognitively complex as SI. Therefore, CAI tool UI design is of utmost importance to ensure that these potentially useful tools are supportive, not disruptive, for the interpreter.

Despite the importance of CAI tool usability, the design of CAI tools has been intuitive rather than systematic and evidence based. While empirical research, for how scarce, has been conducted on the use of CAI tools in the booth, no previous study systematically evaluated the CAI tool interface with the aim to provide empirical evidence for design principles that could make the UI more usable. Differently from CAT research, which has been enriched by cross-fertilisation with the field of HCI, empirical CAI research is still in its infancy and has been mostly initiated by scientific aims. Empirical CAI research has predominantly addressed exploratory research questions (e.g. *how do untrained students use CAI tools?*) or experimental hypotheses (e.g. *do CAI tools lead to an improvement in interpreters' performance?*). Study designs and measures have been consistent with

the experimental research tradition in interpreting studies, but no study has explicitly drawn on a methodological and theoretical HCI framework to evaluate the usability of existing interfaces and inform their future development. Previous studies have created our current knowledge base on CAI and some methodological contributions paved the way for possible new lines of research (e.g. Prandi 2022, which offers the starting point for cognitive CAI research). However, they have not specifically addressed the tool's usability and insights concerning these aspects were obtained incidentally rather than as a major and intended output – with some recent exceptions (Montecchio 2021, EABM 2021b), which, however, must be considered in the light of their limitations. Hence, it may be more appropriate to say that empirical CAI research so far has mostly provided evidence for the tool's *utility* rather than usability.

This book presented a case study of interpreter-centred design and development of an ASR- and AI-powered CAI tool, SmarTerp, and detailed the application of usability testing methods to the empirical evaluation of this solution. After a literature review of usability research methods in usability engineering, translation technology (CAT tools), and interpreting technology (CAI tools), the empirical part of the work shed light on the rationale for the development of methods and materials. The work then presented the results of the two usability tests (i.e. the *Pilot Study* and the *Main Study*) that were conducted with two groups of conference interpreters (no 5 and 10, respectively) to develop design recommendations and improve the solution. The study presented a convergent mixed-method design in which quantitative and qualitative performance and perception data were gathered through a CAI-tool supported SI test, a post-task questionnaire and a semi-structured interview. Other than fulfilling the practical aim of improving the UI of SmarTerp's CAI tool, the study contributed to the field's scientific understanding of interpreter-CAI interaction and moved some steps forward towards the development of data-driven usability heuristics for CAI tool design, as argued in the discussion. Nevertheless, given its novelty and interdisciplinary nature, the study represents a methodological contribution to empirical CAI research. The work may, hopefully, pave the way for a strand of usability-focused empirical CAI research. To this aim, the transcript of the test speeches are provided in the appendix and the study limitations, as well as possible future trajectories and methodological recommendations for future work, are addressed in the discussion.

Usability testing, along with other user research methods, is a fundamental tool to ensure that technological solutions meet conference interpreters' needs and are suitable for the complex cognitive task they intend to support. A usability-focused line of empirical CAI research contributes to ensuring the development

of interpreter-centred solutions, that may help professionals leverage technological affordances, achieve better service quality and keep up with changing market requirements.

Future work should focus not only on the "machine side" of interpreter-CAI interaction but also on the human side. Previous research highlighted problems that are not caused by usability issues but rather by interpreters' improper use of CAI tools, such as their "overreliance" (Defrancq & Fantinuoli 2021) on the visual aids. The present work confirmed that interpreters encounter several difficulties in the use of CAI tools during SI. Despite the apparent simplicity of these tools, given the automaticity offered by ASR and AI technology, the effective integration of visual aids into the SI process appears to be far from simple. Exploring the complexity in interpreter-CAI interaction and identifying the root causes of the problems experienced by interpreters to inform the development of training resources seems highly desirable. Supporting interpreters in the use of CAI tools through training does not only offer practical benefits, i.e. increasing the effectiveness of interpreter-CAI interaction, but also has an ethical dimension. As concerns multiply with new technologies quietly nudging aside humans, assuming greater roles in the T&I industry, helping humans leverage technological innovations to enhance their service represents a contribution to a more thoughtful, ethical system with human interpreters as the pivot of progress.

Research on CAI tool usability and research informing CAI tool training are hence complementary in supporting the profession at a time of unprecedented change. In this crucial time of a "technological turn", research on interpreting technology (be its focus usability or training) is needed to inform change in the field and ensure that technological development is fair and sustainable. This need calls for a redesign not just of tools but of the very role of research and the researcher. To live up to its social responsibility, research will need to work as the link between different stakeholders, ensuring that their concerns are central to technological development.

# Appendix A: Test speech

## A.1 Briefing (communicative context)

Welcome address by Dr. Rene N'Guettia Kouassi, Director of the African Union, at the African Union Conference, on 21 January 2021, on-site and remote.

## A.2 Reading instructions

1. Read at a speed of about 110 words per minute.
2. Make a 0.3-second pause when you see the following in the speech: [0.3 pause].
3. Please, make a shorter, natural pause at the end of an idea or paragraph.
4. The [*extended name of acronyms*], in square brackets and italics, is provided in the transcript only for reference and should not be read out.
5. The speech section headings (Greetings, Programme, Africa's assets, GDP, Population, Natural resources, Conclusion) are provided in the transcript only for reference and should not be read out

## A.3 Transcript

Good morning ladies and gentlemen,
Thank you so much for joining us here today. As you know organising this conference this year has taken extra organisational efforts given the Covid-19 pandemic. Therefore, we are particularly glad and grateful that the conference could take place and that some of us could even be here in person in spite of the adverse circumstances. If everyone has taken their seats and the technicians confirm that the live streaming is functioning properly, I shall start with my address. [0.3 pause]

- H.E. Paul Kagame,
- Hon. Soraya Hakuziyaremye, Rwanda's MINICOM [Ministry of Trade and Industry] Minister,
- Ms Giovanie Biha, UNECA [United Nations Economic Commission for Africa] Deputy Executive Secretary,
- Ladies and gentlemen. [0.3 pause]

It is my utmost pleasure to see so many participants in this historic room that we chose to host our conference. And it is also remarkable that many participants are following this conference from countries across the continent. On behalf of H.E. President Felix-Antoine Tshisekedi Tshilombo, it is my honour and pleasure to welcome you to this conference. It is our hope that today's conference will promote a fruitful dialogue and stimulate collaborations among African countries, which is one of the key objectives of the African Union. [0.3 pause]

Before I go on talking about the African Union and the road ahead for Africa, I would like to remind you of today's programme:

1. After my introductory remarks and the address by our Chairman, H.E. Tschikedi Tschilombo, we will hear the keynote speech 'Made in Africa: towards realizing Africa's Structural Transformation for the achievement of Agenda 2063' by H.E. Kwesi Quartey, AUC's [African Union Commission] Deputy Chairman.
2. We will then hear an address by H.E. Professor Victor Harison, our new Commissioner for the DEA [Department of Economic Affairs], titled 'Rethinking continental initiatives and regional cooperation'.
3. After a 30' break,
4. we will have a 90' panel discussion
5. and finish off with a Q&A session. [0.3 pause]

Ladies and Gentlemen,
As you know this annual gathering reflects the mission of the African Union: To achieve greater unity, cohesion and solidarity between the African countries and African nations and to accelerate the political and social-economic integration of the continent. There are several signs that we are on the right track.

The continent currently has a gross domestic product of USD 3.42 trillion. This represents a remarkable achievement if we consider the fast pace of our economic growth over the past decades. [0.3 pause]

The economic growth will be accompanied, in the years to come, by a massive demographic growth. Our growing population represents an invaluable asset for Africa. This is why the African Union has been promoting the development of an African single market, which would have the largest consumer base in the world. The signing of the AfCFTA [the African Continental Free Trade Area] by the African Member States significantly strengthens the movement towards this goal. [0.3 pause]

Furthermore, while other countries are facing the challenge of an ageing population, Africa's working-age population is growing rapidly and is projected to surpass that of any other continent by 2030:

- By 2030, the African continent would add about 295 million new people aged 15-to-64.
- The growth would push the number of 15-to-64-year-old Africans up by 40% by 2030.
- By 2030, Africa would hence be home to nearly 1 billion people of 15 to 64 years of age. [0.3 pause]

There is no doubt that Africa's population is a precious resource for the further growth of the continent. And this is certainly not our only asset. Let us not forget that Africa has a wealth of natural resources:

- In 2019, Africa produced nearly 8.41 mbd [million barrels per day] of oil.
- Madagascar alone produced approximately 58,000 metric tons of nickel in 2021.
- Namibia's diamond production amounted to 2.52 million carats in 2018. [0.3 pause]

Africa's abundance of natural resources represents a huge potential for the continent's economic growth and could generate shared prosperity for all. Our soil is rich in high-value praseodymium, dysprosium and terbium. These are very rare elements, for which global demand is constantly growing. [0.3 pause]

Such growing demand is reflected in the market value of our mining companies. This year, the market cap of AngloGold Ashanti – the largest mining company headquartered in Africa – was USD 12.13 billion. [0.3 pause]

Africa is also rich in natural gases – much cleaner and cheaper than fossil fuels, which explains the constantly growing global demand. Analysts forecast that African production of LNG [liquefied natural gas] will increase by 150% from 28 mtpy [million tonnes per year] in 2018 to reach 84 mtpy by 2025. [0.3 pause]

We not only expect natural gas demand to continue rising over the years but also to support Africa's transition to green growth. Though its contribution to the total energy mix is still modest, coal-bed methane has impressive potential. The development and utilization of coal-bed methane are of great social and economic benefit. It is a clean-burning fuel for domestic and industrial uses. [0.3 pause]

Furthermore, the porous high-rank coal matrix, with larger hydrocarbon-storage capacity, makes coal-bed methane reservoirs advantageous for commercial operations. [0.3 pause]

Ladies and gentlemen,
We are advancing towards the AU's vision of an integrated, prosperous and peaceful Africa, driven by its own citizens and representing a dynamic force in the global arena. This, however, can only be realized with the full participation of all stakeholders. Which is why events like this conference are so important. Thank you for your attention and your participation. I wish you fruitful discussions.

Table A.1: Tasks in test speech.

| Task code | Name | Definition | Source speech segment |
|---|---|---|---|
| AC | Isolated acronym | One acronym in a simple sentence (20-30 words, with simple syntax and logic and no further problem trigger). | The signing of the AfCFTA [the African Continental Free Trade Area] by the African Member States significantly strengthens the movement towards this goal. |
| NE | Isolated named entity | One named entity in a simple sentence. | On behalf of H.E. President Felix-Antoine Tshisekedi Tshilombo it is my honour and pleasure to welcome you to this conference. |
| NU | Isolated numeral | One complex numeral (3 digits, order of magnitude = 'trillion') in a simple sentence. | The continent currently has a gross domestic product of USD 3.42 trillion. |
| NR | Numeral + complex referent | One complex numeral (4 digits, order of magnitude = 'billion') and a complex referent (an acronym/named entity/specialised term/a numerical value) in a simple sentence. | This year, the market cap of AngloGold Ashanti – the largest mining company headquartered in Africa – was USD 12.13 billion. |
| NIU | Numerical information unit | A complex numerical information unit (NIU), approx.. 30 words, constituted of<br>(1) a complex referent,<br>(2) a complex unit of measurement (an acronym/named entity/specialised term/a numerical value),<br>(3) five consecutive numerals, as in the following structure: (referent) increased/decreased by (X%) from Y in (time1) to Z in (time2). | Analysts forecast that African production of LNG [liquefied natural gas] will increase by 150% from 28 mtpy [million tonnes per year] in 2018 to reach 84 mtpy by 2025. |

| Task code | Name | Definition | Source speech segment |
|---|---|---|---|
| NCR | Redundant number cluster | A number-dense speech passage with the following characteristics:<br>(1) constituted of three subsequent NIUs, approx.10-15 words each;<br>(2) the time and place references remain unvaried and are repeated in each NIU;<br>(3) the unit of measurement and the referent remain unvaried, but the referent is expressed with a different synonym in each NIU;<br>(4) the numeral changes in each NIU. | • By 2030, the African continent would add about 295 million new people aged 15-to-64.<br>• The growth would push the number of 5-to-64-year-old Africans up by 40% by 2030.<br>• By 2030, Africa would hence be home to nearly 1 billion people of 15 to 64 years of age. |
| NCN | Non-redundant number cluster | A number-dense speech passage with the following characteristics:<br>(1) constituted of three subsequent NIUs, approx. 10-15 words each;<br>(2) time, place, referent, unit of measurement and numeral change in each NIU;<br>(3) either the referent or the unit of measurement is complex. | • In 2019, Africa produced nearly 8.41 mbd [million barrels per day] of oil.<br>• Madagascar alone produced approximately 58,000 metric tons of nickel in 2021.<br>• Namibia's diamond production amounted to 2.52 million carats in 2018. |
| TE | Specialised term in a simple sentence | A specialised term in a simple sentence. | Though its contribution to the total energy mix is still modest, coal-bed methane has impressive potential. |

| Task code | Name | Definition | Source speech segment |
|---|---|---|---|
| TL | List of three specialised terms in a simple sentence | A list of three specialised terms in a simple sentence. | Our soil is rich in high-value praseodymium, dysprosium and terbium. |
| TS | List of three specialised terms in a semantically complex sentence | A series of three terms in a semantically complex sentence (i.e., characterised by implicit logical passages where comprehension requires relevant background knowledge and inference). | Furthermore, the porous high-rank coal matrix, with larger hydrocarbon-storage capacity, makes coal-bed methane reservoirs advantageous for commercial operations. |

# Appendix B: Training speech

## B.1 Briefing (communicative context)

Speech by Yoshihide Suga, Prime Minister of Japan. Welcome address at the High-Level Meeting on Japan's Economic Recovery in the Covid-19 aftermath, February 3, 2021 (via Webcast).

## B.2 Reading instructions

1. Read at a speed of about 110 words per minute.
2. Make a 0.3-second pause when you see the following in the speech: [0.3 pause].
3. Please, make a shorter, natural pause at the end of an idea or paragraph.
4. The [*extended name of acronyms*], in square brackets and italics, is provided in the transcript only for reference and should not be read out.
5. The speech section headings (Greetings, Programme, Africa's assets, GDP, Population, Natural resources, Conclusion) are provided in the transcript only for reference and should not be read out

## B.3 Transcript

Good morning ladies and gentlemen,
Our session is just about to start in few minutes, so please make sure to check your camera and microphone settings. While we wait for all participants to connect, let me say that I am very pleased that you could all join this forum. We intended to create an arena for thoughtful discussion and debate, which is all the most important in these challenging times of an enduring global pandemic. I see that all participants have connected now, so I will start with my welcome address. [0.3 pause]

- H.E. Tarō Asō, Deputy Prime Minister,
- Honourable Hiroshige Sekō, METI [Ministry of Economy, Trade and Industry] Minister,
- Distinguished Fumio Hayashi, Professor at GRIPS [National Graduate Institute for Policy Studies],
- Ladies and Gentlemen, distinguished colleagues. [0.3 pause]

Good morning.
It is my pleasure to have the opportunity today to exchange views with leaders in administrative, financial, and economic areas in Kanagawa-ken [Kanagawa Prefecture], which is taking place online due to the continuing impact of the novel coronavirus (COVID-19). [0.3 pause]

At the January MPM [Monetary Policy Meeting], we heard the outlook for Japan's economic activity through the coming fiscal year. While the pandemic had an impact on our economy and society, the overall outlook appears to be positive. [0.3 pause]

Despite the adverse circumstances, our country remains a global economic force. Japan has a GDP of USD 5.38 trillion. This means that, although our economy had the biggest decline on record in the aftermath of the pandemic, it is now recovering. [0.3 pause]

In my address today, I would like to remind us of the strengths of the Japanese economy and open the discussion on how to make it even more resilient and competitive in the years to come. Before I continue with my address, let me remind you of the programme for today's conference:

1. My welcome address will be followed by a keynote speech by Honourable Hiroshige Sekō, Minister of Economy, Trade and Industry.
2. Kuroda Haruhiko, Governor of the Bank of Japan, will then honour us with a presentation on the topic 'Japan's Economic and Monetary Policy in the aftermath of the Covid-19 pandemic.'
3. After a 45' break,
4. we will finish off with our roundtable on economic recovery, digital policy and reinvention. [0.3 pause]

Distinguished colleagues, While COVID-19 has had a significant impact on social and economic activities, this also could be an opportunity to reform the economic structure that will strengthen the growth potential.
Let us think for instance of the various efforts to use information and communication technology in such areas as telework and online medical care. This increased use led to new innovative solutions and boosted the growth of our communication network, which is considered the most advanced in the world. The importance of telecommunications in our country is reflected in the market value of SFTBF [SoftBank], which has reached USD 159.60 billion. This is one of the examples of how our society and economy can make the most of the harsh experience of COVID-19. [0.3 pause]

Exports have traditionally been a core pillar of the Japanese economy. Before the pandemic, Japan was and remains a global exporter of metals. Japan's export of hot-rolled iron increased by 3.51% from 12.5 million tons in YTD [year to date] 2018 to 16.85 million tons in YTD 2019. Although exports decreased when the pandemic hit, they have been witnessing a steady increase over the current fiscal year. [0.3 pause]

Our machinery and equipment are still considered a synonym of quality around the world. Japanese machinery, like our high-performance spinning extruders are leaders in the plastics manufacturing industry. [0.3 pause]

Recent trends in Japan's exports offer further evidence of an ongoing trade recovery built around tech. Our top exports include core pieces of equipment in microchip manufacturing, such as microassemblies, semiconductor lithography equipment, and wire enamel. These are key components and devices supporting the information-oriented society and we expect that their export volumes will continue to increase in the next years. [0.3 pause]

Looking at the road ahead, we are committed to economic growth that is sustainable not only from a financial and social but also from an environmental point of view. Japan is progressing towards its green goals of meeting a larger share of its electricity demand through biomass-derived power:

- Japan's domestic production of woody biomass increased to 10.2 million BDt [bone-dry metric tons] in 2020.
- By 2022, the new biomass plant in Aichi-ken will add 75 megawatts to Japan's production capacities.

- The prefecture of Kyushu alone imported 0.71 million metric tons of PKS [palm kernel shells] in 2018.

These figures are promising because biomass-derived power sets Japan on course to become a carbon-neutral society. [0.3 pause]

In our debates today as well as our policymaking for the years to come, we cannot fail to take into account one of the main challenges that our country will have to face – that of an ageing population due to high longevity and low birth rates:

- In Japan, the percentage of persons aged 65 years old and over exceeded 28.7% of the total population as of 2020.
- The elderly population of Japan increased by 300,000 in 2020 from a year earlier.
- A total of 36.17 million people above 65 years of age lived in Japan in 2020. [0.3 pause]

The Japanese Government and business community seek to offset the effects of an ageing population on economic growth and government budget resources. The ageing population shapes demand and opportunities in various segments. Innovation in the field of medicine, such as our Peptide Discovery Platform System, also provides a great contribution to improving the living standards of our elderly population. The synthetically created macrocyclic peptides, with their high degree of specific binding, offer the opportunity to target protein surfaces traditionally considered "undruggable". [0.3 pause]

Ladies and Gentlemen,
There is concern that the COVID-19 shock we are facing will remain for a long time as scarring effects. It is important to avoid such "scarring effects" on the economy as much as possible in order to create a situation where Japan's economy will return to a sustainable growth path without any sudden deceleration once the impact of COVID-19 subsides. Therefore, the top priority is to overcome the crisis as swiftly as possible. At the same time, it is also important to strengthen the growth potential while making the most of the lessons learned from this crisis. I am confident that this event will represent a major contribution to defining the strategy for the future of our country.
Thank you.

Table B.1: Tasks in training speech.

| Task code | Name | Definition | Source speech segment |
|---|---|---|---|
| AC | Isolated acronym | One acronym in a simple sentence (20-30 words, with simple syntax and logic and no further problem trigger). | At the January MPM [Monetary Policy Meeting], we heard the outlook for Japan's economic activity through the coming fiscal year. |
| NE | Isolated named entity | One named entity in a simple sentence. | It is my pleasure to have the opportunity today to exchange views with leaders in administrative, financial, and economic areas in Kanagawa-ken [Kanagawa Prefecture] |
| NU | Isolated numeral | One complex numeral (3 digits, order of magnitude = 'trillion') in a simple sentence. | Japan has a GDP of USD 5.38 trillion. |
| NR | Numeral + complex referent | One complex numeral (4 digits, order of magnitude = 'billion') and a complex referent (an acronym/named entity/specialised term/a numerical value) in a simple sentence. | The importance of telecommunications in our country is reflected in the market value of SFTBF [SoftBank], which has reached USD 159.6 billion. |
| NIU | Numerical information unit | A complex numerical information unit (NIU), approx.. 30 words, constituted of<br>(1) a complex referent,<br>(2) a complex unit of measurement (an acronym/named entity/specialised term/a numerical value),<br>(3) five consecutive numerals, as in the following structure: (referent) increased/decreased by (X%) from Y in (time1) to Z in (time2). | • In Japan, the percentage of persons aged 65 years old and over exceed 28.75% of the total population as of 2020,<br>• The elderly population of Japan increased by 300,000 in 2020 from a year earlier.<br>• A total of 36.17 million people above 65 years of age lived in Japan in 2020. |

| Task code | Name | Definition | Source speech segment |
|---|---|---|---|
| NCR | Redundant number cluster | A number-dense speech passage with the following characteristics:<br>(1) constituted of three subsequent NIUs, approx.10-15 words each;<br>(2) the time and place references remain unvaried and are repeated in each NIU;<br>(3) the unit of measurement and the referent remain unvaried, but the referent is expressed with a different synonym in each NIU;<br>(4) the numeral changes in each NIU. | • Japan's domestic production of woody biomass increased to 10.2 million BDt [bone-dry metric tons] in 2020.<br>• By 2020, the new biomass plant in Aichi-ken will add 75 MW to Japan's production capacities.<br>• The prefecture of Kyushu alone imported 0.71 million metric tons of PKS [palm kernel shells] in 2018 |
| NCN | Non-redundant number cluster | A number-dense speech passage with the following characteristics:<br>(1) constituted of three subsequent NIUs, approx. 10-15 words each;<br>(2) time, place, referent, unit of measurement and numeral change in each NIU;<br>(3) either the referent or the unit of measurement is complex. | • Japan's domestic production of woody biomass increased to 10.2 million BDt [bone-dry metric tons] in 2020.<br>• By 2020, the new biomass plant in Aichi-ken will add 75 MW to Japan's production capacities.<br>• The prefecture of Kyushu alone imported 0.71 million metric tons of PKS [palm kernel shells] in 2018 |
| TE | Specialised term in a simple sentence | A specialised term in a simple sentence. | Japan machines, like our high-performance spinning extruders are leaders in the plastics manufacturing industry |

| Task code | Name | Definition | Source speech segment |
|---|---|---|---|
| TL | List of three specialised terms in a simple sentence | A list of three specialised terms in a simple sentence. | Our top exports include core pieces of equipment in microchip manufacturing, such as microassemblies, semiconductor lithography equipment, and wire enamel. |
| TS | List of three specialised terms in a semantically complex sentence | A series of three terms in a semantically complex sentence (i.e., characterised by implicit logical passages where comprehension requires relevant background knowledge and inference). | Innovation in the field of medicine, such as our Peptide Discovery Platform System, also provides a great contribution to improving the living standards of our elderly population. The synthetically created macrocyclic peptides, with their high degree of specific binding, offer the opportunity to target protein surfaces traditionally considered “undruggable”. |
| SO | Complex speech opening | Address to the participants with three unknown given names and surnames and two unknown acronyms. | • H.E. Tarō Asō, Deputy Prime Minister,<br>• Honourable Hiroshige Sekō, METI [Ministry of Economy, Trade and Industry] Minister,<br>• Distinguished Fumio Hayashi, Professor at GRIPS [National Graduate Institute for Policy Studies] |

| Task code | Name | Definition | Source speech segment |
|---|---|---|---|
| CP | Conference programme | Announcement of the conference programme. | • My welcome address will be followed by a keynote speech by Honourable Hiroshige Seko, Minister of Economy, Trade and Industry<br>• Kuroda Haruhiko, Governor of the Bank of Japan, will then honor us with a presentation on the topic 'Japan's Economic and Monetary Policy in the aftermath of the Covid-19 pandemic'.<br>• After a 45' break, we will finish off with our roundtable on economic recovery, digital policy and reinvention. |

# References


Alcina, Amparo. 2008. Translation technologies: Scope, tools and resources. *Target* 20(1). 79–102.

Alessandrini, Maria Serena. 1990. Translating numbers in consecutive interpretation: An experimental study. *The Interpreters' Newsletter* 3. 77–80. http://hdl.handle.net/10077/2159.

Andres, Dörte. 2015. Consecutive interpreting. In Franz Pöchhacker (ed.), *Routledge encyclopedia of interpreting studies*, 84–87. Abingdon: Routledge.

Arthern, Peter J. 1978. Machine translation and computerised terminology systems: A translator's viewpoint. In Barbara M. Snell (ed.), *Translating and the Computer*. London: Aslib. https://aclanthology.org/1978.tc-1.5.

Asare, Edmund K. 2011. *An ethnographic study of the use of translation tools in a translation agency: Implications for translation tool design*. Kent, OH: Kent State University. (Doctoral dissertation).

Automatic Language Processing Advisory Committee (ALPAC). 1966. *Languages and machines: Computers in translation and linguistics*. Tech. rep. 1416. Washington, D.C.: Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council.

Baigorri-Jalón, Jesús. 2014. *From Paris to Nuremberg: The birth of conference interpreting*. Trans. by Holly Mikkelson & Barry Slaughter Olsen (Benjamins Translation Library 111). Amsterdam: John Benjamins.

Baker, Mona & Gabriela Saldanha (eds.). 2019. *Routledge encyclopedia of translation studies*. London: Routledge. DOI: 10.4324/9781315678627.

Barnum, Carol M. 2020. *Usability testing essentials: Ready, set... test!* 2nd edn. Burlington, MA: Morgan Kaufmann.

Berber-Irabien, Diana. 2010. *Information and communication technologies in conference interpreting*. Tarragona: Universitat Rovira i Virgili. (Doctoral dissertation).

Bevan, Nigel, James Carter & Susan Harker. 2015. ISO 9241-11 revised: What have we learnt about usability since 1998? In Masaaki Kurosu (ed.), *Human-computer interaction: Design and evaluation, 17th international conference on Human-Computer Interaction (HCI 2015)*, 143–151. Cham: Springer.

Bevan, Nigel, Jim Carter, Jonathan Earthy, Thomas Geis & Susan Harker. 2016. New ISO standards for usability, usability reports and usability measures. In Masaaki Kurosu (ed.), *Human-computer interaction: Theory, design, development and practice, 18th international conference on Human-Computer Interaction (HCI 2016)*, 268–278. Cham: Springer.

Bevan, Nigel, Jurek Kirakowskib & Jonathan Maissela. 1991. What is usability. In *Proceedings of the 4th international conference on HCI*.

Biagini, Giulio. 2015. *Glossario cartaceo e glossario elettronico durante l'interpretazione simultanea: Uno studio comparativo*. Trieste: Università di Trieste. (MA thesis).

Bilgen, Baris. 2009. *Investigating terminology management for conference interpreters*. Ottawa: University of Ottawa. (Doctoral dissertation).

Bowker, Lynne. 2002. *Computer-aided translation technology: A practical introduction*. Ottawa: University of Ottawa Press.

Bowker, Lynne & Desmond Fisher. 2010. Computer-aided translation. In Yves Gambier & Luc van Doorslaer (eds.), *Handbook of translation studies*, vol. 1, 60–65. Amsterdam: John Benjamins.

Braun, Sabine. 2019. Technology and interpreting. In Minako O'Hagan (ed.), *The Routledge handbook of translation and technology*, 271–288. New York: Routledge. DOI: 10.4324/9781315311258-16.

Braun, Susanne & Andrea Clarici. 1996. Inacuracy for numerals in simultaneous interpretation: Neorolinguistic and neuropsychlogical perspectives. *The Interpreters' Newsletter* 7. 85–102. http://hdl.handle.net/10077/8993.

Brooke, John. 1996. Sus: A "quick and dirty" usability scale. In Patrick W. Jordan, B. Thomas, Ian Lyall McClelland & Bernard Weerdmeester (eds.), *Usability evaluation in industry*, 189–194. London: CRC Press.

Brown, John Seely, Allan Collins & Paul Duguid. 1989. Situated cognition and the culture of learning. *Educational Researcher* 18(1). 32–42.

Brown, Tim. 2009. *Change by design: How design thinking transforms organizations and inspires innovation*. New York: Harper.

Bundgaard, Kristine, Tina Paulsen Christensen & Anne Schjoldager. 2016. Translator-computer interaction in action: An observational process study of computer-aided translation. *Journal of Specialised Translation* (25). 106–130.

Burmester, Michael, Marc Hassenzahl & Franz Koller. 2002. Usability ist nicht alles: Wege zu attraktiven Produkten (Beyond usability: Appeal of interactive products). *i-com* 1(1). 32–40. DOI: 10.1524/icom.2002.1.1.032.

Buxton, William & Richard Sniderman. 1980. Iteration in the design of the human-computer interface. In *Proc. of the 13th annual meeting, Human Factors Association of Canada*, 72–81. Downsview: Human Factors Association of Canada.

Cadwell, Patrick, Sheila Castilho, Sharon O'Brien & Linda Mitchell. 2016. Human factors in machine translation and post-editing among institutional translators. *Translation Spaces* 5(2). 222–243.
Canali, Sara. 2019. *Technologie und Zahlen beim Simultandolmetschen: Utilizzo del riconoscimento vocale come supporto durante l'interpretazione simultanea dei numeri*. Rome: Università degli studi internazionali di Roma. (MA thesis).
Carl, Michael & Sabine Braun. 2017. Translation, interpreting and new technologies. In Kirsten Malmkjær (ed.), *The Routledge handbook of translation studies and linguistics*, 374–390. London: Routledge.
Chan, Sin-Wai. 2007. Taking a technological turn: The making of a dictionary of translation technology. *Journal of Translation Studies* 10(1). 113–130.
Chan, Sin-Wai. 2015. The development of translation technology: 1967–2013. In Sin-Wai Chan (ed.), *The Routledge encyclopedia of translation technology*, 3–31. New York: Routledge.
Cheng, Ran. 2021. SDL Trados and Tmxmall: A comparative study of computer-aided translation tools. *Journal of Networking and Telecommunications* 2(4). 76–79. DOI: 10.18282/jnt.v2i4.1379.
Chernov, Ghelly V. 2004. *Inference and anticipation in simultaneous interpreting: A probability-prediction model* (Benjamins Translation Library 57). Amsterdam: John Benjamins.
Cheung, Martha. 2011. Reconceptualizing translation: Some Chinese endeavours. *Meta: journal des traducteurs/Meta: Translators' Journal* 56(1). 1–19.
Chmiel, Agnieszka. 2008. Boothmates forever? On teamwork in a simultaneous interpreting booth. *Across Languages and Cultures* 9(2). 261–276.
Christensen, Tina Paulsen, Marian Flanagan & Anne Schjoldager. 2017. Mapping translation technology research in translation studies: An introduction to the thematic section. *Hermes-Journal of Language and Communication in Business* 56. 7–20.
Colominas, Carme. 2008. Towards chunk-based translation memories. *Babel* 54(4). 343–354.
Cooper, Alan. 2004. *Why high-tech products drive us crazy and how to restore the sanity*. Carmel, IN: Sams Publishing.
Coppers, Sven, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna Van der Lek-Ciudin, Tom Vanallemeersch & Vincent Vandeghinste. 2018. Intellingo: An intelligible translation environment. In Regan Mandryk & Mark Hancock (eds.), *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. New York: Association for Computing Machinery. DOI: 10.1145/3173574.3174098.

Corpas Pastor, Gloria & Lily May Fern. 2016. *A survey of interpreters' needs and practices related to language technology*. Tech. rep. FFI2012-38881-MINECO/TI-DT-2016-1. Universidad de Málaga.

Costa, Hernani, Gloria Corpas Pastor & Isabel Durán Muñoz. 2014. A comparative user evaluation of terminology management tools for interpreters. In Patrick Drouin, Natalia Grabar, Thierry Hamon & Kyo Kageura (eds.), *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, 68–76. Dublin: Association for Computational Linguistics & Dublin City University.

Cowan, Nelson. 2000. Processing limits of selective attention and working memory: Potential implications for interpreting. *Interpreting* 5(2). 117–146.

Cowan, Nelson. 2010. The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science* 19(1). 51–57.

Creswell, John W. & J. David Creswell. 2017. *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: Sage.

Cronin, Michael. 2010. The translation crowd. *Revista Tradumàtica* 8. 1–7. DOI: 10.5565/rev/tradumatica.100.

Cronin, Michael. 2012. *Translation in the digital age*. London: Routledge.

Davis, Elizabeth T. & John Palmer. 2004. Visual search and attention: An overview. *Spatial Vision* 17(4–5). 249–255. DOI: 10.1163/1568568041920168.

De Merulis, Gianpiero. 2013. *L'uso di InterpretBank per la preparazione di una conferenza sul trattamento delle acque reflue: Glossario terminologico e contributo sperimentale*. Bologna: Università di Bologna. (MA thesis).

Defrancq, Bart & Claudio Fantinuoli. 2021. Automatic speech recognition in the booth: Assessment of system performance, interpreters' performances and interactions in the context of numbers. *Target* 33(1). 73–102.

Desmet, Bart, Mieke Vandierendonck & Bart Defrancq. 2018. Simultaneous interpretation of numbers and the impact of technological support. In Claudio Fantinuoli (ed.), *Interpreting and technology* (Translation and Multilingual Natural Language Processing 11), 13–27. Berlin: Language Science Press. DOI: 10.5281/zenodo.1493291.

Doherty, Stephen. 2019. Translation technology evaluation research. In Minako O'Hagan (ed.), *The Routledge handbook of translation and technology*, 339–353. New York: Routledge. DOI: 10.4324/9781315311258-20.

Donovan, Clare. 2006. Where is interpreting heading and how can training courses keep up. In *Future of conference interpreting: Training, technology and research*. London: University of Westminster.

Duflou, Veerle. 2016. *Be(com)ing a conference interpreter: An ethnography of EU interpreters as a professional community* (Benjamins Translation Library 124). Amsterdam: John Benjamins.

Dumas, Joe. 2007. The great leap forward: The birth of the usability profession (1988–1993). *Journal of Usability Studies* 2(2). 54–60.

Ehrensberger-Dow, Maureen. 2019. Ergonomics and the translation process. *Slovo.ru: Baltic Accent* 10(1). 37–51.

Ehrensberger-Dow, Maureen & Gary Massey. 2014a. Cognitive ergonomic issues in professional translation. In John W. Schwieter & Aline Ferreira (eds.), *The development of translation competence: Theories and methodologies from psycholinguistics and cognitive science*, 58–86. Newcastle upon Tyne: Cambridge Scholars Publishing.

Ehrensberger-Dow, Maureen & Gary Massey. 2014b. Translators and machines: Working together. *Man vs. Machine* 1. 199–207.

Ehrensberger-Dow, Maureen & Sharon O'Brien. 2015. Ergonomics of the translation workplace: Potential for cognitive friction. *Translation Spaces* 4(1). 98–118.

Ergonomics for the Artificial Booth Mate (EABM). 2021a. *Cai*. https://www.eabm.ugent.be/cai/.

Ergonomics for the Artificial Booth Mate (EABM). 2021b. *Survey*. https://www.eabm.ugent.be/survey/.

European Commission, Directorate-General for Interpretation. 2018. *Interpreting and translating for Europe*. Luxembourg: Publications Office of the European Union. DOI: 10.2862/183834.

Falbo, Caterina. 2016. Going back to Ancient Egypt: Were the Princes of Elephantine really "overseers of dragomans"? *The Interpreters' Newsletter* 21. 109–114. DOI: 10.13137/1591-4127/13734.

Fantinuoli, Claudio. 2006. Specialized corpora from the web and term extraction for simultaneous interpreters. In Marco Baroni & Silvia Bernardini (eds.), *Wacky! Working papers on the web as corpus*, 173–190. Bologna: CEDIB.

Fantinuoli, Claudio. 2016. InterpretBank: Redefining computer-assisted interpreting tools. In João Esteves-Ferreira, Juliet Macan, Ruslan Mitkov & Olaf-Michael Stefanov (eds.), *Proceedings of Translating and the Computer 38*, 42–52. London: AsLing. https://aclanthology.org/2016.tc-1.5.

Fantinuoli, Claudio. 2017. Speech recognition in the interpreter workstation. In João Esteves-Ferreira, Juliet Macan, Ruslan Mitkov & Olaf-Michael Stefanov (eds.), *Proceedings of the Translating and the Computer 39*, 25–34. London: AsLing.

Fantinuoli, Claudio. 2018a. Computer-assisted interpreting: Challenges and future perspectives. In Gloria Corpas Pastor & Isabel Durán-Muñoz (eds.), *Trends in e-tools and resources for translators and interpreters* (Translation and Multilingual Natural Language Processing 11), 153–174. Leiden: Brill Rodopi.

Fantinuoli, Claudio. 2018b. Interpreting and technology: The upcoming technological turn. In Claudio Fantinuoli (ed.), *Interpreting and technology*, 1–12. Berlin: Language Science Press. DOI: 10.5281/zenodo.1493289.

Fantinuoli, Claudio, Giulia Marchesini, David Landan & Lukas Horak. 2022a. *KUDO interpreter Assist: Automated real-time support for remote interpretation.* arXiv. DOI: 10.48550/arXiv.2201.01800.

Fantinuoli, Claudio, Giulia Marchesini, David Landan & Lukas Horak. 2022b. Kudo interpreter assist: automated real-time support for remote interpretation. *arXiv preprint arXiv:2201.01800.*

Fantinuoli, Claudio & Maddalena Montecchio. 2022. *Defining maximum acceptable latency of AI-enhanced CAI tools.* arXiv. DOI: 10.48550/arXiv.2201.02792.

Flanagan, Marian & Tina Paulsen Christensen. 2014. Testing post-editing guidelines: How translation trainees interpret them and how to tailor them for translator training purposes. *The Interpreter and Translator Trainer* 8(2). 257–275.

Fossati, Giona. 2021. *SmarTerp: Applying the user-centered design process in a computer-assisted interpreting (CAI) tool.* Madrid: E.T.S. de Ingenieros Informáticos de la Universidad Politécnica de Madrid. (MA thesis).

Frittella, Francesca Maria. 2017. *Numeri in interpretazione simultanea: Difficoltà oggettive e soggettive.* New York: Europa Edizioni.

Frittella, Francesca Maria. 2019a. "70.6 billion world citizens": Investigating the difficulty of interpreting numbers. *Translation & Interpreting* 11(1). 79–99. DOI: 10.12807/ti.111201.2019.a05.

Frittella, Francesca Maria. 2019b. Numbers: From stumbling block to training tool. *The Interpreters' Newsletter* (24). 35–56. DOI: 10.13137/2421-714X/29524.

Gacek, Michael. 2015. *Softwarelösungen für DolmetscherInnen.* Vienna: Universität Wien. (MA thesis). DOI: 10.25365/thesis.35667.

Gaido, Marco, Susana Rodríguez, Matteo Negri, Luisa Bentivogli & Marco Turchi. 2021. *Is "Moby Dick" a whale or a bird? Named entities and terminology in speech translation.* arXiv. DOI: 10.48550/arXiv.2109.07439.

Garcia, Ignacio. 2015. Computer-aided translation: Systems. In Sin-Wai Chan (ed.), *Routledge encyclopedia of translation technology*, 68–87. New York: Routledge.

Gile, Daniel. 1987. La terminotique en interprétation de conférence: Un potentiel à exploiter. *Meta: journal des traducteurs/Meta: Translators' Journal* 32(2). 164–169.

Gile, Daniel. 2009. *Basic concepts and models for interpreter and translator training* (Benjamins Translation Library 8). Amsterdam: John Benjamins.

Goodman, Elizabeth, Mike Kuniavsky & Andrea Moed. 2012. *Observing the user experience: A practitioner's guide to user research*. Waltham, MA: Elsevier.

Gould, John D. 1988. How to design usable systems. In Martin G. Helander (ed.), *Handbook of human-computer interaction*, 757–789. Amsterdam: North-Holland.

Gould, John D. & Clayton Lewis. 1985. Designing for usability: Key principles and what designers think. *Communications of the ACM* 28(3). 300–311.

Green, Spence, Jason Chuang, Jeffrey Heer & Christopher D. Manning. 2014. Predictive translation memory: A mixed-initiative system for human language translation. In Hrvoje Benko (ed.), *Proceedings of the 27th Annual ACM Symposium on user interface software and technology*, 177–187. New York: The Association for Computing Machinery. DOI: 10.1145/2642918.2647408.

Hansen-Schirra, Silvia. 2012. Nutzbarkeit von Sprachtechnologien für die Translation. *Trans-kom* 5(2). 211–226.

Hart, Sandra G. & Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Peter A. Hancock & Najmedin Meshkati (eds.), *Human mental workload*, vol. 52, 139–183. Amsterdam: North-Holland.

Hassenzahl, Marc. 2003. The thing and I: Understanding the relationship between user and product. In Mark A. Blythe, Kees Overbeeke, Andrew F. Monk & Peter C. Wright (eds.), *Funology*, 31–42. Dordrecht: Springer. DOI: 10.1007/1-4020-2967-5_4.

Holmes, James S. 1975. *The name and nature of translation studies*. Amsterdam: Translation Studies Section, Department of General Literary Studies, University of Amsterdam.

Holzinger, Andreas. 2005. Usability engineering methods for software developers. *Communications of the ACM* 48(1). 71–74.

House, Juliane (ed.). 2014. *Translation: A multidisciplinary approach*. London: Palgrave Macmillan.

Hutchins, W. John. 1998. The origins of the translator's workstation. *Machine Translation* 13(4). 287–307.

Hutchins, W. John. 2015. Machine translation: History of research and applications. In Sin-Wai Chan (ed.), *Routledge encyclopedia of translation technology*, 120–136. London: Routledge.

Hwang, Wonil & Gavriel Salvendy. 2010. Number of people required for usability evaluation: The 10±2 rule. *Communications of the ACM* 53(5). 130–133.

ISO. 1998. *Ergonomic requirements for office work with visual display terminals (VDTs).* Standard ISO 9241-11:1998. Geneva.

ISO. 2010. *Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems.* Standard ISO 9241-210:2010. Geneva.

ISO. 2011. *Systems and software engineering: Systems and software quality requirements and evaluation (SQuaRE): System and software quality models.* Standard ISO/IEC 25010:2011. Geneva.

ISO. 2016. *Systems and software engineering: Systems and software quality requirements and evaluation (SQuaRE): Measurement of quality in use.* Standard ISO/IEC 25022:2016. Geneva.

ISO. 2018. *Ergonomics of human-system interaction – Part 11: usability: definitions and concepts.* Standard ISO 9241-11:2018. Geneva.

Jakobsen, Arnt Lykke. 2017. Translation process research. In John W. Schwieter & Aline Ferreira (eds.), *The handbook of translation and cognition*, 19–49. New York: Wiley-Blackwell.

Jensen, John B. 2006. *The strategic partnership in the conference interpreting booth.* Paper presented at the Annual Meeting of the American Translators Association.

Jiménez-Crespo, Miguel A. 2020. The "technological turn" in translation studies: Are we there yet? A transversal cross-disciplinary approach. *Translation Spaces* 9(2). 314–341.

Jones, Roderick. 2014. *Conference interpreting explained.* London & New York: Routledge.

Kalina, Sylvia. 2000. Interpreting competences as a basis and a goal for teaching. *The Interpreters' Newsletter* (10). 3–32. http://hdl.handle.net/10077/2440.

Kalina, Sylvia. 2007. "Microphone off": Application of the process model of interpreting to the classroom. *Kalbotyra* 57(3). 111–121. DOI: 10.15388/klbt.2007.7564.

Kalina, Sylvia. 2010. New technologies in conference interpreting. In Hannelore Lee-Jahnke & Erich Prunc (eds.), *Am Schnittpunkt von Philologie und Translationswissenschaft: Festschrift zu Ehren von Martin Forstner*, 79–96. Bern: Peter Lang.

Kalina, Sylvia. 2015. Preparation. In Franz Pöchhacker (ed.), *Routledge encyclopedia of interpreting studies*, 318–319. Abingdon: Routledge.

Kappus, Martin & Maureen Ehrensberger-Dow. 2020. The ergonomics of translation tools: Understanding when less is actually more. *The Interpreter and Translator Trainer* 14(4). 386–404.

Kay, Martin. 1980. *The proper place of men and machines in language translation.* Research report CSL-80-11. Palo Alto, CA: Xerox Palo Alto Research Center.

Koponen, Maarit, Wilker Aziz, Luciana Ramos & Lucia Specia. 2012. Post-editing time as a measure of cognitive effort. In Sharon O'Brien, Michel Simard & Lucia Specia (eds.), *Workshop on Post-Editing Technology and Practice*. San Diego, CA: Association for Machine Translation in the Americas. https://aclanthology.org/2012.amta-wptp.2.

Koskinen, Kaisa & Minna Ruokonen. 2017. Love letters or hate mail? Translators' technology acceptance in the light of their emotional narratives. In Dorothy Kenny (ed.), *Human issues in translation technology*, 8–24. Abingdon: Routledge.

Krüger, Ralph. 2016a. Contextualising computer-assisted translation tools and modelling their usability. *trans-kom* 9(1). 114–148.

Krüger, Ralph. 2016b. Fachübersetzen aus kognitionstranslatologischer Perspektive. *trans-kom* 8(2). 273–313.

Krüger, Ralph. 2019. A model for measuring the usability of computer-assisted translation tools. In Heike Elisabeth Jüngst, Lisa Link, Klaus Schubert & Christiane Zehrer (eds.), *Challenging boundaries: New approaches to specialized communication*, 93–117. Berlin: Frank & Timme.

Kurz, Ingrid. 1985. The rock tombs of the princes of Elephantine: Earliest references to interpretation in Pharaonic Egypt. *Babel* 31(4). 213–218.

Lagoudaki, Elina. 2006. Translation memories survey 2006: User's perceptions around TM usage. In *Proceedings of Translating and the Computer 28*. London: Aslib. https://aclanthology.org/2006.tc-1.2.

Lagoudaki, Pelagia Maria. 2008. *Expanding the possibilities of translation memory systems: From the translator's wishlist to the developer's design*. London: Imperial College London. (Doctoral dissertation).

Läubli, Samuel. 2020. *Machine translation for professional translators*. Zürich: University of Zürich. (Doctoral dissertation). DOI: 10.5167/uzh-193466.

Läubli, Samuel & Spence Green. 2019. Translation technology research and human-computer interaction (HCI). In Minako O'Hagan (ed.), *The Routledge handbook of translation and technology*, 370–383. New York: Routledge. DOI: 10.4324/9781315311258-22.

Läubli, Samuel & David Orrego-Carmona. 2017. When Google Translate is better than some human colleagues, those people are no longer colleagues. In João Esteves-Ferreira, Juliet Macan, Ruslan Mitkov & Olaf-Michael Stefanov (eds.), *Proceedings of Translating and the Computer 39*, 59–69. London: AsLing.

Läubli, Samuel, Patrick Simianer, Joern Wuebker, Geza Kovacs, Rico Sennrich & Spence Green. 2021. The impact of text presentation on translator performance. *Target* 34(2). 309–342. DOI: 10.1075/target.20006.lau.

Laugwitz, Bettina, Theo Held & Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In Andreas Holzinger (ed.), *HCI and usability for education and work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society (USAB 2008)*, 63–76. Berlin: Springer. DOI: 10.1007/978-3-540-89350-9_6.

Lazar, Jonathan, Jinjuan Heidi Feng & Harry Hochheiser. 2017. *Research methods in human-computer interaction*. Cambridge, MA: Morgan Kaufmann.

LeBlanc, Matthieu. 2013. Translators on translation memory (TM): Results of an ethnographic study in three translation services and agencies. *Translation & Interpreting* 5(2). 1–13.

Lewis, James. 2012. Usability testing. In Gavriel Salvendy (ed.), *Handbook of human factors and ergonomics*, 4th edn., 1267–1312. Hoboken, NJ: John Wiley.

Mayhew, Deborah J. 2007. Requirements specifications within the usability engineering lifecycle. In Andrew Sears & Julie A. Jacko (eds.), *The human-computer interaction handbook*, 2nd edn., 913–921. Boca Raton: CRC Press.

Mazza, Cristina. 2001. Numbers in simultaneous interpretation. *The Interpreters' Newsletter* (11). 87–104. http://hdl.handle.net/10077/2450.

Melby, Alan K. 1978. Design and implementation of a machine-assisted translation system. In *Proceedings of the 7th international conference on Computational Linguistics*. Bergen.

Mellinger, Christopher D. & Gregory M. Shreve. 2016. Match evaluation and over-editing in a translation memory environment. In Ricardo Muñoz Martín (ed.), *Reembedding translation process research*, 131–148. Amsterdam: John Benjamins.

Merlini, Raffaella. 2015. Dialogue interpreting. In Franz Pöchhacker (ed.), *Routledge encyclopedia of interpreting studies*, 102–107. Abingdon: Routledge.

Mizuno, Akira. 2005. Process model for simultaneous interpreting and working memory. *Meta* 50(2). 739–752.

Montecchio, Maddalena. 2021. *Defining maximum acceptable latency of ASR-enhanced CAI tools: Quantitative and qualitative assessment of the impact of ASR latency on interpreters' performance*. Mainz: Johannes Gutenberg-Universität Mainz. (MA thesis).

Moorkens, Joss & Sharon O'Brien. 2013. User attitudes to the post-editing interface. In Sharon O'Brien, Michel Simard & Lucia Specia (eds.), *Proceedings of the 2nd Workshop on Post-editing Technology and Practice*, 19–25. Nice: Association for Machine Translation in the Americas. https://aclanthology.org/2013.mtsummit-wptp.3.

Moorkens, Joss & Sharon O'Brien. 2017. Assessing user interface needs of post-editors of machine translation. In Dorothy Kenny (ed.), *Human issues in translation technology*, 109–130. Abingdon: Routledge.

Moser-Mercer, Barbara. 1992. Terminology documentation in conference interpretation. *Terminologie et traduction* 2(3). 285–303.

Nakamura, Jeanne & Mihaly Csikszentmihalyi. 2002. The concept of flow. In C. R. Snyder & Shane J. Lopez (eds.), *Handbook of positive psychology*, 89–105. New York: Oxford University Press.

Nielsen, Jakob. 1992. The usability engineering life cycle. *Computer* 25(3). 12–22.

Nielsen, Jakob. 1993. *Usability engineering*. San Francisco, CA: Morgan Kaufmann.

Nielsen, Jakob. 1994. Usability inspection methods. In Catherine Plaisant (ed.), *Conference companion on Human factors in computing systems (CHI'94)*, 413–414. New York: Association for Computing Machinery. DOI: 10.1145/259963.260531.

Nielsen, Jakob. 2010. What is usability? In Chauncey Wilson (ed.), *User experience re-mastered: Your guide to getting the right design*, 3–22. Burlington, MA: Morgan Kaufmann.

Nielsen, Jakob. 2012. *Usability 101: Introduction to usability*. Nielsen Norman Group. https://www.nngroup.com/articles/usability-101-introduction-to-usability/.

Nielsen, Jakob & Thomas K. Landauer. 1993. A mathematical model of the finding of usability problems. In Bert Arnold, Gerrit van der Veer & Ted White (eds.), *Proceedings of the INTERACT'93 and CHI'93 Conference on Human factors in Computing Systems*, 206–213. New York: Association for Computing Machinery. DOI: 10.1145/169059.169166.

O'Brien, Sharon. 2005. Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine Translation* 19(1). 37–58.

O'Brien, Sharon. 2012. Translation as human–computer interaction. *Translation Spaces* 1(1). 101–122.

O'Brien, Sharon, Maureen Ehrensberger-Dow, Marcel Hasler & Megan Connolly. 2017. Irritating CAT tool features that matter to translators. *Hermes: Journal of Language and Communication in Business* 56. 145–162.

O'Brien, Sharon, Minako O'Hagan & Marian Flanagan. 2010. Keeping an eye on the UI design of translation memory: How do translators use the "concordance" feature? In Willem-Paul Brinkman & Mark Neerincx (eds.), *Proceedings of the 28th Annual European Conference on Cognitive Ergonomics*, 187–190. New York: Association for Computing Machinery. DOI: 10.1145/1962300.1962338.

O'Hagan, Minako. 2013. The impact of new technologies on translation studies: A technological turn? In Carmen Millán & Francesca Bartrina (eds.), *The Routledge handbook of translation studies*, 521–536. Abingdon: Routledge.

Olohan, Maeve. 2019. Technology, translation. In Mona Baker & Gabriela Saldanha (eds.), *Routledge encyclopedia of translation studies*, 574–578. London: Routledge.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In Pierre Isabelle (ed.), *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318. Stroudsburg, PA: Association for Computational Linguistics. DOI: 10.3115/1073083.1073135.

Pisani, Elisabetta & Claudio Fantinuoli. 2021. Measuring the impact of automatic speech recognition on number rendition in simultaneous interpreting. In Caiwen Wang & Binghan Zheng (eds.), *Empirical studies of translation and interpreting: The post-structuralist approach*, 181–197. New York: Routledge. DOI: 10.4324/9781003017400-14.

Pöchhacker, Franz. 2004. *Introducing interpreting studies*. London: Routledge.

Pöchhacker, Franz. 2015a. Interpreting studies. In Franz Pöchhacker (ed.), *Routledge encyclopedia of interpreting studies*, 201–206. Abingdon: Routledge.

Pöchhacker, Franz. 2015b. Modes. In Franz Pöchhacker (ed.), *Routledge encyclopedia of interpreting studies*, 268–269. Abingdon: Routledge.

Prandi, Bianca. 2015. Use of CAI tools in interpreters' training: A pilot study. In João Esteves-Ferreira, Juliet Macan, Ruslan Mitkov & Olaf-Michael Stefanov (eds.), *Proceedings of Translating and the Computer 37*. London: AsLing. https://aclanthology.org/2015.tc-1.8.

Prandi, Bianca. 2017. Designing a multimethod study on the use of CAI tools during simultaneous interpreting. In João Esteves-Ferreira, Juliet Macan, Ruslan Mitkov & Olaf-Michael Stefanov (eds.), *Proceedings of Translating and the Computer 39*, 76–113. London: AsLing.

Prandi, Bianca. 2018. An exploratory study on CAI tools in simultaneous interpreting: Theoretical framework and stimulus validation. In Claudio Fantinuoli (ed.), *Interpreting and technology* (Translation and Multilingual Natural Language Processing 11), 29–59. Berlin: Language Science Press. DOI: 10.5281/zenodo.1493293.

Prandi, Bianca. 2022. *A cognitive inquiry into the product and process of computer-assisted simultaneous interpreting: An experimental study on terminology*. Germersheim: Johannes Gutenberg-Universität Mainz. (Doctoral dissertation). Published as *Computer-assisted simultaneous interpreting: A cognitive-experimental*

*study on terminology* (Translation and Multilingual Natural Language Processing 22). Berlin: Language Science Press, 2023. DOI: 10.5281/zenodo.7143056.

Prandi, Bianca. 2023. *Computer-assisted simultaneous interpreting: A cognitive-experimental study on terminology* (Translation and Multilingual Natural Language Processing 22). Berlin: Language Science Press. DOI: 10.5281/zenodo.7143056.

Pyla, Pardha S., Manuel A. Pérez-Quiñones, James D. Arthur & H. Rex Hartson. 2005. Ripple: An event driven design representation framework for integrating usability and software engineering life cycles. In Ahmed Seffah, Jan Gulliksen & Michel C. Desmarais (eds.), *Human-centered software engineering: Integrating usability in the software development lifecycle*, 245–265. Dordrecht: Springer. DOI: 10.1007/1-4020-4113-6_13.

Pym, Anthony. 2011. What technology does to translating. *Translation & Interpreting* 3(1). 1–9.

Risku, Hanna. 2002. Situatedness in translation studies. *Cognitive Systems Research* 3(3). 523–533.

Risku, Hanna. 2004. *Translationsmanagement: Interkulturelle Fachkommunikation im Informationszeitalter* (Translationswissenschaft 1). Tübingen: Gunter Narr.

Rubin, Jeffrey & Dana Chisnell. 2008. *Handbook of usability testing: How to plan, design and conduct effective tests*. Indianapolis, IN: John Wiley.

Rütten, Anja. 2007. *Informations- und Wissensmanagement im Konferenzdolmetschen*. Frankfurt am Main: Lang.

Rütten, Anja. 2016. Professional precariat or digital elite? Workshop on interpreters' workflows and fees in the digital era. In João Esteves-Ferreira, Juliet Macan, Ruslan Mitkov & Olaf-Michael Stefanov (eds.), *Proceedings of Translating and the Computer 38*. London: AsLing. https://aclanthology.org/2016.tc-1.13.

Rütten, Anja. 2018. Can interpreters' booth notes tell us what really matters in terms of information and terminology management? In João Esteves-Ferreira, Juliet Macan, Ruslan Mitkov & Olaf-Michael Stefanov (eds.), *Proceedings of Translating and the Computer 40*, 132–144. London: AsLing.

Sand, Peter. 2015. *The new Interplex: Glossaries made fast and easy*. AIIC Webzine. https://aiic.org/document/813/AIICWebzine_Winter2011_5_SAND_The_new_Interplex_glossaries_made_fast_and_easy_EN.pdf.

Schmettow, Martin, Wendy Vos & Jan Maarten Schraagen. 2013. With how many users should you test a medical infusion pump? Sampling strategies for usability tests on high-risk systems. *Journal of Biomedical Informatics* 46(4). 626–641.

Schneider, Dominik, Marcos Zampieri & Josef van Genabith. 2018. Translation memories and the translator: A report on a user survey. *Babel* 64(5-6). 734–762.

Scriven, Michael. 1967. The methodology of evaluation. In Ralph W. Tyler, Robert M. Gagné & Michael Scriven (eds.), *Perspectives of curriculum evaluation* (AERA Monograph Series on Curriculum Evaluation 1), 39–83. Chicago: Rand McNally.

Seeber, Kilian G. & Dirk Kerzel. 2012. Cognitive load in simultaneous interpreting: Model meets data. *International Journal of Bilingualism* 16(2). 228–242.

Seewald, Frauke & Marc Hassenzahl. 2004. Vom kritischen Ereignis zum Nutzungsproblem: Die qualitative Analyse in diagnostischen Usability Tests. In Marc Hassenzahl & Matthias Peissner (eds.), *Tagungsband Usability Professionals 2004*, 142–148. Stuttgart: Fraunhofer Verlag.

Shneiderman, Ben. 1980. *Software psychology: Human factors in computer and information systems*. Boston, MA: Little, Brown & Company.

Snell-Hornby, Mary. 2006. *The turns of translation studies: New paradigms or shifting viewpoints?* Amsterdam: John Benjamins.

Still, Brian & Kate Crane. 2017. *Fundamentals of user-centered design: A practical approach*. Boca Raton, FL: CRC press.

Stoll, Christoph. 2009. *Jenseits simultanfähiger Terminologiesysteme: Methoden der Vorverlagerung und Fixierung von Kognition im Arbeitsablauf professioneller Konferenzdolmetscher*. Trier: Wiss. Verlag Trier.

Teixeira, Carlos S. C., Joss Moorkens, Daniel Turner, Joris Vreeke & Andy Way. 2019. Creating a multimodal translation tool and testing machine translation integration using touch and voice. *Informatics* 6(1). Article 13. DOI: 10.3390/informatics6010013.

Teixeira, Carlos S. C. & Sharon O'Brien. 2017. Investigating the cognitive ergonomic aspects of translation tools in a workplace setting. *Translation Spaces* 6(1). 79–103.

Tripepi Winteringham, Sarah. 2010. The usefulness of ICTs in interpreting practice. *The Interpreters' Newsletter* (15). 87–99. http://hdl.handle.net/10077/4751.

Van Cauwenberghe, Goran. 2020. *La reconnaissance automatique de la parole en interprétation simultanée*. Gent: Universiteit Gent. (MA thesis). https://lib.ugent.be/catalog/rug01:002862551.

van Kuijk, Jasper, Heimrich Kanis, Henri Christiaans & Daan van Eijk. 2017. Barriers to and enablers of usability in electronic consumer product development: a multiple case study. *Human-Computer Interaction* 32(1). 1–71.

Vargas-Sierra, Chelo. 2019. Usability evaluation of a translation memory system. *Quaderns de Filologia: Estudis Lingüístics* XXIV. 119–146. DOI: 10.7203/QF.24.16302.

Walker, Callum. 2021. *An eye-tracking study of equivalent effect in translation: The reader experience of literary style*. Cham: Palgrave Macmillan. DOI: 10.1007/978-3-030-55769-0.

Whiteside, John, John Bennett & Karen Holtzblatt. 1988. Usability engineering: Our experience and evolution. In Martin G. Helander (ed.), *Handbook of human-computer interaction*, 791–817. Amsterdam: North-Holland.

Whyman, Edward K. & Harold L. Somers. 1999. Evaluation metrics for a translation memory system. *Software: Practice and Experience* 29(14). 1265–1284.

Will, Martin. 2009. *Dolmetschorientierte Terminologiearbeit: Modell und Methode*. Tübingen: Gunter Narr.

Will, Martin. 2015. Zur Eignung simultanfähiger Terminologiesysteme für das Konferenzdolmetschen. *Zeitschrift für Translationswissenschaft und Fachkommunikation* 8(1). 179–201.

Yin, Robert K. 2013. *Case study research: Design and methods*. Thousand Oaks, CA: Sage.

Zaretskaya, Anna. 2017. *Translators' requirements for translation technologies: User study on translation tools*. Málaga: Universidad de Málaga. (Doctoral dissertation). https://hdl.handle.net/10630/16246.

# Name index

# Usability research for interpreter-centred technology

Technological advancement is reshaping the ways in which language interpreters operate. This poses existential challenges to the profession but also offers new opportunities. This books moves from the design of the computer-assisted interpreting tool SmarTerp as a case study of impact-driven interpreting technology research. It demonstrates how usability testing was used to achieve an interpreter-centred design. By contextualising the case study within the past and current developments of translation and interpreting technology research, this book seeks to inform and inspire a redefinition of the conceptualisation of interpreting technology research – not just as a tool to understand change but to drive it into a sustainable and human direction.