Fostering human-centered, augmented machine translation:

Analysing interactive post-editing

Vicent Briva-Iglesias, B.A., M.Sc.

Supervised by Prof. Sharon O'Brien (main) and Dr. Benjamin Cowan (secondary, UCD)

A thesis presented for the degree of Doctor of Philosophy

School of Applied Languages and Intercultural Studies

Dublin City University

March 2024

DECLARATION

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Vicent Briva-Iglesias

ID No.: 20213499

Date: March 05, 2024

ACKNOWLEDGEMENTS

My sincere gratitude goes to Prof. Sharon O'Brien. It has been a pleasure to work for 3 and a half years under your supervision, which has allowed me to learn from one of the most brilliant people I have ever met, as well as to improve every day, both academically and personally. Your dedication, help and diligence are to be remembered; a capital THANK YOU. I would also like to thank Dr. Benjamin Cowan, for his assistance and guidance during the statistical analyses.

Special mention should be made to the different participants of the thesis. Those participating in the pilot study, which was made possible thanks to funding from the European Association for Machine Translation (EAMT), and those participating in the main study, funded by the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real). These two bodies also deserve my gratitude for making this research thesis possible.

An essential part of the Dublin experience is the people who have made it possible, even 2,400 km away from Valencia, to call a new city and a new country home. They have also made the rainy days bearable, and the Irish paellas unforgettable. You know who you are, especially “el tridentito”.

Last, but not least, I would like to thank my family for their unconditional support. Being a first-gen academic can be difficult, hard, and complex. Everything I am and will be, I owe to you.

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION 1
CHAPTER 2. LITERATURE REVIEW 5
2.1. Translation technologies and computer-assisted translation tools 5
2.2. Machine translation architectures 7
2.2.1. Rule-based machine translation 7
2.2.2. Statistical machine translation 9
2.2.3. Neural machine translation 11
2.3. Post-editing 13
2.3.1. Traditional post-editing (TPE) 13
2.4. Translation quality 15
2.4.1. Translation quality evaluation 17
2.5. Translation productivity 21
CHAPTER 3. INTERACTIVE POST-EDITING 23
3.1. Introduction to interactive post-editing (IPE) 23
3.2. Historical context 26
3.2.1. TransType and TransType 2 26
3.2.2. CAITRA 34
3.2.3. CASMACAT 38
3.2.4. Lilt 49
3.2.5. Additional user evaluations of IMT systems for IPE tasks 53
3.3. Discussion on interactive post-editing (IPE) 67
CHAPTER 4. TRANSLATION AS A FORM OF HUMAN-COMPUTER INTERACTION 70
4.1. Human-computer interaction (HCI) 70
4.2. Usability 74
4.2.1. Usability engineering and testing 77
4.3. User experience (UX) 80
4.3.1. User experience from different angles 84
4.3.2. User experience evaluation 86
4.4. Studies of HCI factors in Translation Studies involving MT 89
4.4.1. Usability in Translation Studies involving MT 90
4.4.2. User Experience in Translation Studies involving MT 93
4.4.3. Ergonomics, situated interactions, and hedonomics in Translation Studies involving MT 98
4.5. Human-centered, augmented MT (HCAMT) 101

CHAPTER 5. RESEARCH RATIONALE AND RESEARCH QUESTIONS 104
5.1. Operationalising MTUX 105
5.2. Research questions and hypotheses 109
5.2.1. Factor one: Machine Translation User Experience (MTUX) 109
5.2.2. Factor two: translation productivity 110
5.2.3. Factor three: translation quality 110
5.2.4. Further exploration of Machine Translation User Experience 111
CHAPTER 6. METHODOLOGY 113
6.1. Pilot experiment 113
6.1.1. Participants 113
6.1.2. Content 114
6.1.3. IPE workbench 114
6.1.4. Design of the controlled user study 117
6.1.5. Results and lessons learned 118
6.2. Main longitudinal study 119
6.2.1. Participants 120
6.2.2. Design of the controlled, main longitudinal study 121
6.2.3. Texts 124
6.2.4. IPE workbench 126
6.2.5. Measures 127
6.3. The mixed-methods approach 129
CHAPTER 7. RESULTS OF THE MAIN LONGITUDINAL STUDY 132
7.1. RQ1. Is MTUX statistically significantly impacted by MTPE modality (TPE or IPE) and does this vary with increased experience? 132
7.1.1. Average MTUX scores 133
7.1.2. MTUX scores per factor 134
7.2. RQ2. Is translation productivity statistically significantly impacted by MTPE modality (TPE or IPE) and does this vary with increased experience? 143
7.3. RQ3. Is fluency statistically significantly impacted by MTPE modality (TPE or IPE) and does this vary with increased experience? 145
7.4. RQ4. Is adequacy statistically significantly impacted by MTPE modality (TPE or IPE) and does this vary with increased experience? 147
7.5. RQ5 to RQ7. Do pre-task perceptions of MTPE correlate with fluency, adequacy, or productivity? 149
Fluency scores 151
Adequacy scores 152
Productivity scores 153

7.6. RQ8 to R10. Does MTUX correlate with fluency, adequacy, or productivity? 155
7.7. Discussion of the results 155
7.7.1. IPE produces a statistically significantly higher MTUX 156
7.7.2. IPE allows for working statistically significantly faster after some acclimatisation 158
7.7.3. IPE statistically significantly impacts fluency, but not adequacy 159
7.7.4. Perceptions of MT influence quality and productivity 160
CHAPTER 8. CONCLUSIONS, STRENGTHS, LIMITATIONS AND FUTURE WORK 163
8.1. Breaking the vicious circle: designing for pleasure rather than for absence of pain .164
8.2. Strengths 166
8.3. Limitations 168
8.4. Future work 168
REFERENCES 171
APPENDIX A. DCU ETHICS APPROVAL 194
APPENDIX B. RECRUITMENT JOB AD FOR THE MAIN STUDY – TRANSLATORS AND REVIEWERS 195
APPENDIX C. PRE-TASK QUESTIONNAIRE 198
APPENDIX D. MTUX QUESTIONNAIRE 199
APPENDIX E. FLUENCY DIFFERENCE BETWEEN TPE AND IPE 201

TABLE OF FIGURES

Figure 3.1. IPE process. 25
Figure 3.2. TransType (TT_1) interface. 28
Figure 3.3. TransType 2 (TT2) interface. 31
Figure 3.4. Screenshot of the interface of CAITRA. 35
Figure 3.5. Screenshot of CASMACAT's graphic user interface.. 39
Figure 3.6. The image shows the effect of week on Kdur per source text character. 46
Figure 3.7. Interface screenshot of PTM. 50
Figure 3.8. Screenshot of Lilt's graphic user interface 52
Figure 3.9. Forecat interface. 62
Figure 4.1. Building blocks for user-centered UX proposed by Roto (2006). 85
Figure 6.1. Graphic user interface of Lilt in the TPE modality 115
Figure 6.2. Graphic user interface of Lilt in the IPE modality 116
Figure 6.3 Design of the controlled, main longitudinal study 122
Figure 7.1. MTUX Score Comparison in the Evaluation Sessions (with SD bars) 133
Figure 7.2 MTUX score evolution during the learning sessions 134
Figure 7.3. Attractiveness MTUX Score Comparison in the Evaluation Sessions (with SD bars) 135
Figure 7.4. Perspicuity MTUX Score Comparison in the Evaluation Sessions (with SD bars) 136
Figure 7.5. Efficiency MTUX Score Comparison in the Evaluation Sessions Sessions (with SD bars) 138
Figure 7.6. Dependability MTUX Score Comparison in the Evaluation Sessions Sessions (with SD bars) 139
Figure 7.7. Stimulation MTUX Score Comparison in the Evaluation Sessions (with SD bars) 141
Figure 7.8. Novelty MTUX Score Comparison in the Evaluation Sessions Sessions (with SD bars) 142
Figure 7.9. Productivity Comparison in the Evaluation Sessions Sessions (with SD bars) 144
Figure 7.10. Productivity evolution during the learning sessions 145
Figure 7.11. Fluency Comparison in the Evaluation Sessions Sessions (with SD bars) 146
Figure 7.12. Fluency evolution during the learning sessions 147

Figure 7.13. Adequacy Comparison in the Evaluation Sessions 148
Figure 7.14. Adequacy evolution during the learning sessions 149
Figure 7.15. Correlation of translators' pre-task perceptions of MTPE with fluency scores.151
Figure 7.16. Correlation of translators' pre-task perceptions of MTPE with adequacy scores 152
Figure 7.17. Correlation of translators' pre-task perceptions of MTPE with productivity 154

LIST OF ABBREVIATIONS

AI	Artificial intelligence
CAT	Computer-assisted translation
HCAI	Human-centered artificial intelligence
HCAMT	Human-centered, augmented machine translation
HCI	Human-computer interaction
MT	Machine translation
MTUX	Machine translation user experience
MTPE	Machine translation post-editing
UX	User experience

PUBLICATIONS AND PRESENTATIONS FROM THIS RESEARCH

Publications

Briva-Iglesias, Vicent, and Sharon O'Brien. 2024. 'Pre-Task Perceptions of MT Influence Quality and Productivity: The Importance of Better Translator-Computer Interactions and Implications for Training'. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation.*

Briva-Iglesias, Vicent. 2023. 'Translation Technologies Advancements: From Inception to the Automation Age'. In *La Família Humana: Perspectives Multidisciplinàries de La Investigació En Ciències Humanes i Socials*, Lucía Bellés-Calvera; María Pallarés-Renau, 137–52. Emergents 3. Publicacions de la Universitat Jaume I. Servei de Comunicació i Publicacions.

Briva-Iglesias, Vicent, and Sharon O'Brien. 2022. 'The Language Engineer: A Transversal, Emerging Role for the Automation Age'. *Quaderns de Filologia - Estudis Lingüístics 27* (0): 17–48. https://doi.org/10.7203/qf.0.24622.

Briva-Iglesias, Vicent, and Sharon O'Brien. 2023. 'Measuring Machine Translation User Experience: A Comparison between AttrakDiff and User Experience Questionnaire'. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 335–44.

Briva-Iglesias, Vicent, Sharon O'Brien, and Benjamin R. Cowan. 2023. 'The Impact of Traditional and Interactive Post-Editing on Machine Translation User Experience, Quality, and Productivity': *Translation, Cognition & Behavior 6 (*1). https://doi.org/10.1075/tcb.00077.bri.

Presentations

- *Human-centred machine translation via interaction design*, Translating and the Computer 45. Luxembourg, November 2023.
- *The impact of traditional and interactive post-editing on machine translation user experience, quality, and productivity*, 4th International Conference in Translation, Interpreting and Cognition (ICTIC4). Chile, September 2023.
- *Machine translation user experience (MTUX): Towards more human-centric engagement forms*, In "The evolving role of the post-editor" Tutorial at The 24th Annual Conference of the European Association for Machine Translation. Finland, June 2023.
- *Measuring machine translation user experience (MTUX): A comparison between AttrakDiff and User Experience Questionnaire*, The 24th Annual Conference of the European Association for Machine Translation. Finland, June 2023.

- *What is machine translation user experience and why should we start looking at it? An overview in multilingual communication processes*, 3rd Annual Conference Language in the Human-Machine Era (LITHME). Leeuwarden, The Netherlands, May 2023.
- *Human-centred translation technology development: What can we learn from human-computer interaction?* 3rd Annual Conference of the Translation Studies Network for Ireland (TSNI3). Ireland, April 2023.
- *The language engineer: An emerging, transversal role for the automation age*, GALA Global 2023. Ireland, March 2023.
- *Pre-task perceptions and their impact on final translation quality: Implications for training*, Translating and the Computer 44. Luxembourg, November 2022.
- *How to measure machine translation experience: Attrakdiff vs User Experience Questionnaire*, New Trends in Translation and Technology Conference. Rhodes, Greece, July 2022.
- *Els avenços de les tecnologies de la traducció: Des dels inicis fins a l'era de l'automatització*, Jornades de Foment en Investigació en Ciències Humanes i Sociales. Castelló de la Plana, May 2022.

ABSTRACT

Fostering human-centered, augmented machine translation: Analysing interactive post-editing

Vicent Briva-Iglesias

Recent language technology developments have disrupted the translation and interpreting professions. However, the focus has been on using more computational power and training larger language models, often neglecting the users of such technology (do Carmo and Moorkens 2022).

To date, the goal of technology development has been the creation of an intelligent agent that emulates human behaviour to increase automation. As a response, a novel technology design framework has gained a foothold recently: human-centered artificial intelligence, where instead of human replacement, the aim is to produce a powerful tool that augments human capabilities, enhances performance, and empowers users, who are at all instances in supervisory control of such systems (Shneiderman 2022). If applied to machine translation (MT), we can talk about human-centered, augmented MT (HCAMT). This shift, moving from emulation to empowerment, places humans at the centre of AI/language technology. This PhD thesis presents the concept of Machine Translation User Experience (MTUX) as a way to foster HCAMT. Consequently, we conduct a longitudinal user study with 11 professional translators in the English-Spanish language combination that analyses the effects of traditional post-editing (TPE) and interactive post-editing (IPE) on MTUX, translation quality and productivity. MTUX results suggest that translators prefer IPE to TPE because they are in control of the interaction in this new form of translator-computer interaction and feel more empowered in their interaction with MT. Productivity results also suggest that translators working with IPE report a statistically significantly higher productivity than when working with TPE. Quality results also indicate that translators offer more fluent translations in IPE, and equally adequate translations in both post-editing modalities. All these results allow for reflection on the potential adoption of IPE as a more HCAMT post-editing modality, which empowers the users, who have been increasingly reluctant to interact with machine translation post-editing in industry workflows (Cadwell, O'Brien, and Teixeira 2018).

This PhD thesis establishes the methodology for fostering HCAMT tools, systems and workflows through the study of MTUX. The successful implementation of HCAMT in translation and interpreting may lead to sustainable, diverse, and ethically sound development in MT systems and other technological tools through a wide variety of users and use-cases.

CHAPTER 1. INTRODUCTION

In an era where the boundaries of language and culture are increasingly blurred by global communication and digitalisation, the significance of machine translation (MT) in bridging linguistic gaps has never been more pronounced (Vieira, O'Hagan, and O'Sullivan 2021). This PhD thesis situates itself within this critical juncture, where the convergence of MT and human-computer interaction (HCI) bears new possibilities and challenges in global, digital communication.

Recent developments in language technologies have been rapid, mainly due to the availability of more data online, as well as improved computing algorithms and the use of more computational power (Brown et al. 2020). Currently, the research focus is on using larger amounts of data to develop larger and larger language models (LLMs) through resource-intensive and resource-extensive training processes, both economic and natural (Chuan Li 2020; Zhong et al. 2023).

In the literature review (Chapter 2 to Chapter 4), we can see that the experience and needs of the MT user have been overlooked in the development and adoption of new language technologies. It is important to emphasise that the user is one of the most important elements of the user-MT interaction (if not the most important) and, currently, the main focus in MT development is to deliver higher quality (Moorkens et al. 2018). This is when we see that most of the current technological development processes focus on creating an autonomous system that automates human tasks, which is what artificial intelligence (AI) is all about (Shneiderman 2022a). According to the recently passed European Union Artificial Intelligence Act (EU AI Act) (European Union 2024, 39), an AI system is a "machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments". As per "automation", the Cambridge Dictionary defines this concept as "the use of machines and computers that can operate without needing human control" (Cambridge Dictionary 2024).

In the language services industry, we are starting to see some sectors where MT is used directly as an AI system without any human intervention if assimilation is being pursued

(Schmidtke and Groves 2019). It is important to stress that using unsupervised MT must be carefully considered depending on the risks, value and goals of such use (Way 2013). Thus, in a context of quality saturation in various language combinations, where high-quality MT is already achieved in some major language combinations, we ask whether this path of technological development is the appropriate way forward: more data and more computational power to train systems that are a drain on economic and natural resources (Zhong et al. 2023), spending hundreds of thousands (or millions) of euros to achieve limited quality improvements (Chuan Li 2020). In addition, many people also view this path of technological development with concern about the possible loss of jobs or the decision-making about people's lives based on recommendations from these technologies (Eloundou et al. 2023; Oviedo-Trespalacios et al. 2023). For instance, the integration of MT in translation production workflows has often been met with resistance, primarily due to concerns over quality, loss of control, and the dehumanising nature of MT-assisted translation (e.g. Fırat 2021; Moorkens 2020; Cadwell, O'Brien, and Teixeira 2018). Is there a more sustainable and ethically sound alternative in the development and use of language technologies that would allow efforts and investments to improve these technologies to be channelled to all parties involved?

It is in this context that a new technology design framework is gaining momentum: human-centered technology or human-centered artificial intelligence (HCAI) (Shneiderman 2022b). The goal of HCAI is to pursue "intelligence amplification (IA)" as opposed to AI. The goal of HCAI is to build a tool that enhances human capabilities, improves human performance and empowers users, who should always possess supervisory control of the technology in an enjoyable interaction with the technology. This is in addition to the concept of "augmentation" (Raisamo et al. 2019; O'Brien 2023). If we apply all these concepts to MT, we can talk about human-centered, augmented machine translation (HCAMT). Thus, in order to develop a methodology to inform the development of HCAMT tools, systems and workflows for a wide range of MT users and use cases, we used the language services industry and the translator-MT interaction as a use case through a two-week longitudinal study.

In analysing the application of MT in the language services industry through post-editing (Chapter 2), we saw that there might be ways of interacting with MT that follow a more HCAMT approach than traditional workflows, such as interactive post-editing (IPE) (Chapter

3). The hypothesis of this thesis is based on the fact that, in IPE tasks, it is the machine that adapts to the human in real time, with the person being in supervisory control of the interaction. In contrast, in traditional post-editing (TPE) tasks, it is the human who adapts to static MT proposals. Consequently, this PhD thesis analyses and explores the user experience (UX) of professional translators in IPE and TPE tasks in one of the early HCI- and HCAI-based studies for fostering HCAMT.

Reviewing the field of HCI (Chapter 4), we note that UX should be a key element in technology development and adoption (Albert and Tullis 2022), but that it has received minuscule and negligible attention in the Translation Studies and the MT technology development communities (Briva-Iglesias and O'Brien 2023). Using Roto's (2016) UX framework as a basis, where the user-MT interaction is a situated activity where the user, the context where the interaction takes place and the system need to be considered as a whole, together with the HCAI framework, we conducted a pilot study with 15 professional translators, who performed TPE and IPE tasks over two consecutive days. In this pilot study, we designed and tested a methodology to measure the user-MT interaction and coined the concept of machine translation user experience (MTUX) as a person's perceptions and responses resulting from the use and/or anticipated use of MT (Briva-Iglesias and O'Brien 2023; Briva-Iglesias, O'Brien, and Cowan 2023). When analysing MTUX, both pre-task perceptions and post-task perceptions of MT users are highly relevant. Although the pilot results were promising, only two interactions were not enough to obtain clear, statistically significant results. Thus, having tested the methodology and validated the research design, we conducted one of the few longitudinal studies in Translation Studies and MT, and analysed the MTUX, quality and productivity of 11 professional translators over two weeks of interaction with TPE and IPE. This longitudinal study was also designed to account for the variable experience levels of the translators with the different MT post-editing (MTPE) modalities, which permitted us to make a fair comparison of different measures for the TPE and IPE modalities. Every translator had substantial experience with TPE tasks, but none of them had experience with IPE. Thus, this PhD thesis aimed to respond to the following overarching research question:

- RQ. Is IPE a better alternative to TPE in terms of machine translation user experience (MTUX), translation productivity, and translation quality?

Delving into the importance of UX in user-MT interactions, this study posits that a system, tool or workflow that produces a better MTUX can statistically significantly impact the users' satisfaction, but also their performance (in terms of quality and productivity) and overall engagement with MT. A HCAMT approach not only facilitates smoother interactions but also empowers MT users by providing them with tools and interfaces that are intuitive, responsive, and adaptable to their specific needs. This focus on personalisation and MTUX is paramount in addressing the diverse challenges faced by professional translators or other MT users. Furthermore, the inclusion of productivity and quality within the context of MTUX is not merely a matter of operational efficiency but a critical consideration for the sustainability of the translation profession. As the demand for rapid, high-quality translations continues to grow (Moorkens 2017), the ability of translators to leverage MT effectively becomes a key competitive advantage. Therefore, enhancements in MTUX can lead to significant gains in productivity without sacrificing the quality of translation, thereby striking a balance that benefits all stakeholders involved in the translation production process.

At the core of this research, we advocate for HCAMT tools, workflows, and systems, and defend a shift in perspective from viewing MT as a standalone solution to seeing it as a tool that, when thoughtfully integrated into the translation production process, can augment human capabilities rather than replace them. Through a comprehensive examination of MTUX, this thesis seeks to illuminate the pathways through which HCAMT can be realized, emphasizing the critical role of UX in bridging the gap between human limitations and machine efficiency, always situating the human at the centre of the interaction. By weaving together insights from HCI, Translation Studies, MT, and UX design, this thesis presents a holistic view of the translator-machine interaction landscape. This integrated approach not only enriches our understanding of the dynamics at play but also charts a course for the future of research on MT, where technology and human expertise converge to create a more connected and comprehensible world. Results and methodologies should not be applicable only to professional translators, but to the whole spectrum of MT users.

CHAPTER 2. LITERATURE REVIEW

This chapter presents a literature review on the most important concepts of translation technologies. Section 2.1 presents an overview of translation technologies and computer-assisted translation (CAT) tools, followed by a brief description of the main machine translation (MT) architectures (Section 2.2). Then, Section 2.3 explores how MT has been traditionally introduced into the language services industry through post-editing. Finally, key concepts such as translation quality (Section 2.4) and translation productivity (Section 2.5) are also presented.

2.1. Translation technologies and computer-assisted translation tools

Although the term "translation technology" seems to be related to today's mobile devices, the cloud, or the popular concept of AI, the technologies applied to the professional world of translation are not that *new*. To speak of them, we need to go back to the late 1950s. Bar-Hillel was commissioned by the US government to produce a report on the state of MT at the time, which was one of the most important areas of research in the field of computer science. After touring the country's leading research centres, Bar-Hillel (1959) issued a rather unfavourable report in which he agreed that the goals of the research teams were unrealistic and that "fully-automatic high-quality translation" was far from being a reality. He therefore recommended that research should focus on the development of less ambitious but more practical technologies for translators. Some years later, the report of the Automatic Language Processing Advisory Committee (ALPAC 1966), also commissioned by the US government, was even more negative. This report indicated that the studies on the nascent MT systems of the second half of the 20th century failed as they did not meet quality expectations and recommended that interest in MT should be transferred to machine-assisted translation, which was aimed at "improving human translation, with an appropriate use of machine aids" (Ibid.: 25). These two reports were the seed that germinated into what we know today as translation technologies (Chan 2014).

At that time, the idea of automatically retrieving previously translated text segments was born in order to facilitate the translation process (Melby 1978; Arthern 1979). Subsequently, several companies began to take an interest in the development of software related to translation processes. Sumita and Tsutsumi (1988) released Easy to Consult, which was an

electronic dictionary that could not only look for single words but also for sentences or phrases of more than two words. Soon after, Trados developed TED and Multiterm, a plug-in for a text processor and a multilingual terminology management tool, respectively. These two tools were later on merged and formed the first Translator's Workbench editor (Garcia and Stevenson 2005). The introduction of these tools, created with the intention of enabling translation of the same content in less time and increasing productivity, started to alter professional translation and documentation processes. According to Chan (2014), translation technologies experienced a first period of rapid growth from 1993 to 2003, and a second period of global development from 2004 to 2013. In the first period, online information sources (dictionaries, encyclopaedias, linguistic corpora in various languages, etc.) appeared and replaced the paper resources previously used. The first terminology management tools also appeared, with the intention of being able to show the terminological equivalences of various languages automatically and faster. In the second period, these series of new tools and features merged into what we know as computer-assisted translation (CAT) tools. CAT tools are a type of computer software that assists translators and facilitates certain aspects of the translation process (Bowker and Fisher 2010). Garcia (2014, 4) defines them as:

> At its core, every CAT system divides a text into "segments" (normally sentences, as defined by punctuation marks) and searches a bilingual memory for identical (exact match) or similar (fuzzy match) source and translation segments. Search and recognition of terminology in analogous bilingual glossaries are also standard. The corresponding search results are then offered to the human translator as prompts for adaptation and reuse.

With technological advances and the adoption of CAT tools in the language services industry (ELIS Research 2023), these tools evolved from being simple applications used to retrieve previously translated fragments through translation memories, manage specific terminology with glossaries or automate quality management processes, to substantially more complex environments (Briva-Iglesias 2023). In addition, CAT tools, which started out as desktop applications, moved later on to the web or the cloud, such as Matecat or Phrase TMS (Rothwell et al. 2023). Currently, CAT tools also have MT components that can be used individually to obtain translation proposals with the aim of assisting the translator or in conjunction with translation memories via a technology called *fuzzy match repair* (Bulté and Tezcan 2019). In this latter case, fuzzy matches with a high coincidence with the text to be

translated are retrieved and leveraged to improve the results of an MT system, mixing both the advantages of translation memories and MT. Thus, CAT tools have evolved from being simple tools with few functionalities, normally supported by office tools such as Microsoft Word, to become an all-in-one solution with many features to ease the work of translators (Rothwell et al. 2023).

2.2. Machine translation architectures

Despite the fact that research on MT was deprioritised following the ALPAC report, research continued, albeit to a lesser extent. Yet, to talk about MT, we must first know what is meant when the term is used. Ginestí and Forcada (2009, 43) define MT as:

> the process of translating, by means of a computer system (consisting of computers and programs), computerized texts written in the source language to computerized texts written in the target language. A computerized text is a computer file that contains text in a known format. [translation by the author]

Therefore, we can say that MT is a tool that, when a computerized text is introduced in a natural language, automatically translates the original text into another natural language, producing what is called a raw translation or MT output. Historically, MT has been based on two approaches, rule-based MT and data-based MT (Nitzke 2019), and its main architectures are described below.

2.2.1. Rule-based machine translation

Rule-based machine translation (RBMT) was the first MT system to be developed, which focused on word for word translation (Ibid.: 6-8). These first MT systems required two main actors to produce MT output:

- Firstly, a group of linguists or translation experts, who had to create a series of dictionary entries and linguistic rules (grammatical, syntactic, and stylistic) of the characteristics of the source language (SL) and the target language (TL).
- Secondly, computer experts. They wrote morphology and syntax analysis programs capable of reading, understanding, and representing the entries of the dictionary to

apply them in the TL. These computer experts could also be computational linguists or linguists with knowledge of computer programming.

RBMT was very useful for minoritized or low-resource languages, and it has been shown to offer adequate translations in several language combinations, especially in those of closely-related languages because of its grammar and syntactic similarities (as, for example, Spanish and Catalan) (Forcada et al. 2011). However, languages are complex, and lexical or syntactic ambiguity is one of the most difficult problems to solve (Ginestí and Forcada 2009). These ambiguities could be tackled by creating new rules. For example, the English term “sheet” can be translated into Spanish as “sábana” [e.g. cotton sheet] or “hoja” [paper sheet]. Context rules could be defined as follows:

```
<rule>
  <or>
    <match lemma=“bed” tags=“n.*”/>
    <match lemma=“cotton” tags=“adj”/>
  </or>
  <match lemma=“sheet” tags=“n.*”>
    <select lemma=“sábana” tags=“n.*”/>
  </match>
</rule>
```

In this context, a condition was added. If the English term “sheet” was accompanied by “bed” or “cotton”, it would be translated into Spanish as “sábana” instead of “hoja”. The problem with RBMT was that the cost of creating and compiling rules was very high because many rules had to be written to achieve a RBMT system that performed well, and this involved a lot of hard and complex work because languages have many exceptions and implications that are difficult to represent with written rules (Borja 2013). An example of a RBMT system that is still in operation today is Apertium (Forcada et al. 2011), but this kind of MT technology has been replaced by newer, more advanced MT architectures.

2.2.2. Statistical machine translation

With the proliferation of the Internet and globalisation, there is a large amount of text available online in many different languages that can be used freely. Statistical machine translation (SMT) is a type of data-based MT. This means that SMT is a type of computing system where an algorithm learns to automatically translate from existing translations aligned in the form of corpora (Hearne and Way 2011). SMT systems go through two different processes.

- Training: this first process involves extracting a statistical translation model from a parallel corpus (and thus associating the probabilities of a word in the SL being translated in a certain way in the TL). Furthermore, the training process also implies the creation of a language model from a monolingual corpus (i.e. a probabilistic dictionary that allows estimating the fluency of a text in the TL) (Koehn 2010).
- Decoding: this second process faces translation as a search problem. When an SL sentence is introduced as input into the SMT engine, the SMT engine looks for all possible translations according to the translation model. Then, the SMT system tries to rearrange these translations according to the language model, and finally provides as MT output the sentence in the TL that is "most likely" (Brown et al. 1990).

Hearne and Way (2011, 206) discuss that SMT is:

> probabilistically plausible; rather than focusing on the best process to use to generate a single optimal translation for a source sentence, SMT focuses on generating many thousands of hypothetical translations for the input string, and then works out which one of those is most likely.

To illustrate this process in a simple way, firstly, texts should be segmented into sentences. Punctuation marks are used as markers to indicate the end of a sentence so that it is automatically recognized what should be separated and what is a separate sentence. For example, if we have the text "The dog is green. The cat is brown", the full stop will be used as a separating element between two sentences. The same will happen with the Spanish version "El perro es verde. El gato es marrón".

Secondly, texts must be aligned. For the system to work properly, it is necessary that the sentences in the SL are aligned with their homonyms in the TL. Therefore, “The dog is green” should be aligned with “El perro es verde”, as well as “The cat is brown” with “El gato es marrón”. The alignment of all these sentences in bilingual texts (or corpora) is done with a single objective: to create the translation model and thus know which word in the TL may correspond to a certain word in the SL (Koehn 2010). Once the sentences have been aligned, a series of probabilistic operations take place, explained below, with the intention of discovering the probable translations of the words of the sentence:

The dog is green	The cat is brown	The dog
El perro es verde	El gato es marrón	El perro

At this point, a process called initialization takes place. All the words have an equal chance of aligning with each other. That is, in the first sentence, “The dog is green”, the word “The” could correspond to “El”, “perro”, “es” or “verde”. However, thanks to a process of iteration, in which the probabilities are compared with the adjacent sentences, by seeing that “The dog is green” corresponds to “El perro es verde”, that “The cat is brown” corresponds to “El gato es marrón” and that “The dog” corresponds to “El perro”, we can see how the probability of one word being the translation of another increases. After making this comparison and this simple probabilistic analysis, we can see that “The” is the SL equivalent of “El” in the TL for this sentence, as well as that “dog” goes hand in hand with “perro”, and “cat” with “gato”, to give some examples. This latter process of word for word alignment is called convergence. Thanks to these alignments, we can obtain a probabilistic bilingual dictionary, that is, we obtain the probabilities that “cat” is “gato” —expressed in [p(cat|gato)]—, thus achieving our objective: a translation model. These p (probability) values are assigned a weight, which will increase or decrease according to the probability that the word in the SL is correct in the TL. However, an additional element must be considered: the correctness or naturalness of the sentences in the TL.

To this end, the SMT engine must be trained with a monolingual text in the TL, that is, instead of doing a probabilistic analysis of bilingual corpora to obtain the translation probabilities (as

has already been done to obtain the translation model), a probabilistic analysis of the monolingual text in the TL must be done to predict the natural order and set the standard of the words and segments. After this process, a monolingual language model is obtained, which marks and sets the grammar, syntactical and standard rules of the TL. As with bilingual corpora, the larger the monolingual language model is and the more sentences and segments it has, the better the result will be.

The process that follows is called “decoding”. At this stage, the SMT system looks for all possible translation hypotheses and values them according to their score and final weight. The hypothesis with the highest score (considering both the language model and the translation model together) will be the option chosen by the SMT engine and will be shown as the MT output. Examples of SMT engines are Moses (Koehn et al. 2007) and MTradumatica (Martín-Mor 2017). SMT systems have been further developed and improved with the introduction of new algorithms and using different statistical methods. One of these examples is phrase-based SMT (PBSMT) (Zens, Och, and Ney 2002), where the system not only included the statistical probabilities that a single word X in the SL could be translated as Z in the TL, but also included contextual information about the statistical translation probabilities of surrounding words.

2.2.3. Neural machine translation

In 2014, a new approach to data-driven MT development was implemented by Sutskever, Vinyals, and Le (2014) in the form of neural machine translation (NMT). According to Koehn (2017), neural network models had been proposed before (Castaño, Casacuberta, and Vidal 1997; Forcada and Ñeco 1997). However, none of these models could be trained with amounts of text like those available now and, moreover, the processing power of machines and computers at that time was not as high as it is today. These advances in processing have made it possible to “resurrect” the old idea of neural models, which created a new opportunity in the world of MT. Forcada (2017, 292) explains NMT as follows:

> The name comes from the fact that the neural networks (which should properly be called artificial neural networks) on which NMT is based are composed of thousands of artificial units that resemble neurons in that their output or activation (that is, the degree to which they are excited or inhibited) depends on the stimuli they receive from other neurons and the strength of the connections along which these stimuli are passed.

In a more detailed but still accessible explanation, the journey of translating text using NMT begins with the input of a sentence or a piece of text in the SL into the MT system. This system employs a sophisticated method to transform each individual word into a numerical format, which is commonly referred to as a "vector". When the system considers the entire context of the sentence, this numerical format evolves into what is termed a "contextual vector" (Pérez-Ortiz, Forcada, and Sánchez-Martínez 2022). At the heart of NMT lies a complex structure known as a neural network. This network is composed of a series of artificial neurons, alongside different layers of algorithms that work with a feature called attention mechanisms (Vaswani et al. 2017), which are pivotal in managing the focus of the translation process. The artificial neural network undertakes a multitude of computations, in the order of millions, with the aim of refining these vectors. These computations are intricate, designed to optimize and adapt the contextual vectors to convey the most accurate semantic and syntactic nuances of the original text. Once this intensive computational process is complete, the outcome is an optimized set of vectors. These are then utilized by another critical component of the neural network, known as the decoder, as in SMT. The decoder's function is pivotal; it uses these optimized vectors to construct and predict the most probable and contextually appropriate translation of the input sentence or text into the TL (Pérez-Ortiz, Forcada, and Sánchez-Martínez 2022).

NMT requires much larger linguistic corpora than previous MT systems, so that the contextual vectors can be adequately optimized during the MT system training phase. In addition, systems cannot be made with conventional computers, but require very powerful processors called graphics processing units or GPUs (Li Chuan 2020). As this is a very computationally demanding process, because millions of mathematical operations need to be calculated, the training process can take days, months or even years, depending on the machine used (Ibid.). Today, NMT is a technology considered as providing better MT output than SMT, the previous paradigm, in terms of translation quality (Castilho, Moorkens, Gaspari, Calixto, et al. 2017).

2.3. Post-editing

MT was introduced in the language services industry through post-editing, a process where mistakes in the MT output are identified and corrected (O'Brien 2022). To put things into perspective, in this PhD I use the term "language services industry" to refer to an industry estimated to be worth USD 69.3 billion in 2023 according to a report by Nimdzi (Hickey 2023), a market research and consultancy company that helps their customers to succeed in the global market. Companies in the language services industry offer a whole range of services that allow brands and companies to internationalise their products and expand throughout the global market. These services include, but are not limited to, translation, localisation or transcreation. Therefore, this section provides an overview of TPE, how it has evolved and its adoption in the language services industry. A summary of the body of literature around the topic is also included. The literature review of post-editing is further developed in Chapter 3, which presents IPE and describes in-depth the evolution and implementation of this newer form of post-editing.

2.3.1. Traditional post-editing (TPE)

As explained above, there have been many changes and improvements in the MT world in the last decade. The paradigm shift, from SMT to NMT (Bentivogli, Bisazza, et al. 2016), has had a major impact on the translation quality provided by these systems (more information on translation quality can be found in Section 2.4 below). However, MT remains an imperfect technology that makes grammatical, syntactic, or lexical errors. Despite MT being used directly in some use cases such as product user reviews (Popović et al. 2021), where accuracy and fluency of the text are not key and a mere comprehension of the source text is sufficient (i.e. MT for *assimilation* purposes) (Kenny 2022), the use of MT can have serious consequences if used directly in sensitive domains such as in medical and legal use cases (i.e. MT for *dissemination* purposes) (Vieira, O'Hagan, and O'Sullivan 2021). As a consequence, to achieve translations without errors, it is still necessary to involve translators to detect errors and implement the necessary changes and edits. Today, it is still strongly argued that raw MT output cannot be compared with translations by professional translators in terms of translation quality (Läubli, Sennrich, and Volk 2018; Toral 2020).

This translator intervention takes place through post-editing. O'Brien (2011, 197) describes post-editing as "the correction of raw machine translated output by a human translator

according to specific guidelines and quality criteria". This task traditionally follows the following process:

1. a digital text in the SL is sent to an MT system;
2. this system provides a proposal for a digital text in the TL; and
3. the translator modifies, changes, or alters the raw translation as necessary.

Post-editing can be carried out in multiple MT interaction interfaces, which can range from a word processing application such as Microsoft Word, or within a CAT tool (see Section 2.1 above). The main goal of post-editing is to translate more content in less time and, specifically in the language services industry, to increase translator productivity and reduce production costs (a review of translation productivity can be read in Section 2.5 below). Thus, with advances in MT systems, post-editing has become an increasingly common practice (ELIS Research 2023) and has become an established field of research in Translation Studies and MT research, which has subsequently split into multiple branches.

Post-editing research has been carried out from many different perspectives. For instance, to analyse users' interactions with MT during post-editing, different tools have been developed to record empirical data from user experiments. An example of such a tool is PET (Aziz, Castilho, and Specia 2012), which records the time a person takes to post-edit a text and logs the keystrokes, allowing productivity comparisons and discovering the post-editing effort involved in post-editing tasks (Krings 2001). Another tool is Translog-II (Carl 2012), which collects the same data as PET and, in addition, is compatible with eye-tracking technologies, allowing researchers to obtain information about the areas where users' attention is focused while post-editing (O'Brien 2006). This type of technology has encouraged Translation Studies to engage with more transdisciplinary studies and new branches of research have appeared, such as the one that focuses on analysing multiple aspects of Translation Process Research (Risku 2014; Carl, Bangalore, and Schaeffer 2016) and even on the different cognitive processes that take place when translating, revising or post-editing. This latter research branch can be described as Cognitive Translation Studies (Alves and Jakobsen 2020). However, Krings (2001), in his PhD dissertation on post-editing, suggested that post-editing effort could be measured on three different levels, namely the temporal, the technical and the cognitive level. This is the seminal work that guided most post-editing research to date in terms of evaluation.

These are only some examples of research perspectives on post-editing, but it is worth stressing that most attention on post-editing research has remained around quality and productivity (see Sections 2.4 and 2.5 for a detailed description of these concepts and additional, relevant literature). As mentioned above, this is because the goal of the language services industry was to introduce post-editing to reduce production costs and become more competitive. As a consequence, post-editing has also been analysed in terms of comparing the effort of post-editing with other types of computer-assisted translation (O'Brien and Moorkens 2014) or with other types of traditional translation (Guerberof-Arenas 2008). Additionally, the productivity of post-editing with an SMT or an NMT system has been researched (Sánchez-Gijón, Moorkens, and Way 2019). Also, there have been other studies that have focused on users and have researched the needs of translators when post-editing (Moorkens and O'Brien 2017) or measured their satisfaction (Cadwell et al. 2016). The adoption of post-editing also disrupted academic training, and therefore new training methods for translators were developed (Nitzke, Tardel, and Hansen-Schirra 2019), new professional profiles for translators like "language engineers" were proposed (Briva-Iglesias and O'Brien 2022), or even the usefulness of post-editing beyond the language services industry was explored, as is the case of foreign language students (Zhang and Torres-Hostench 2022), just to name some examples. However, the body of literature on post-editing is very extensive, and this section mentions some distinguished generic studies. The literature review on post-editing should be conducted considering the purpose of each research study and framework, and, accordingly, Chapter 3 includes a more in-depth review on IPE, a newer post-editing modality, which is the most relevant post-editing workflow for this PhD dissertation.

2.4. Translation quality

The situation described in the previous sections and globalization have transformed the language services industry and landscape enormously. Now, there is a vast amount of content to be translated globally and, in addition, clients are increasingly demanding shorter deadlines accompanied by smaller budgets (Moorkens 2017). However, anyone who needs a translation and entrusts it to a language service provider wants to make sure it is correct and adequate for dissemination. Here is where the concept of "translation quality" becomes very important. It is worth noting that defining "translation quality" is difficult because there is no general

agreement on a particular definition and there are many factors and elements that may influence what one person may consider as a quality translation (Rossi and Carré 2022). Two interesting definitions of this concept are those of Koby et al. (2014, 416), who first suggest in a broad definition that:

> A quality translation demonstrates accuracy and fluency required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account end-user needs.

However, they also claim that they are not convinced by this definition because it excludes other language services, such as transcreation or localisation. Therefore, they offer a second, more concrete definition, where:

> a high-quality translation is one in which the message embodied in the source text is transferred completely into the target text, including denotation, connotation, nuance, and style, and the target text is written in the target language using correct grammar and word order, to produce a culturally appropriate text that, in most cases, reads as if originally written by a native speaker of the target language for readers in the target culture.

In reading the definitions above, the following information can be extracted: a good quality translation is that where i) a text in the SL ii) is transmitted completely, fluently and appropriately, to iii) another text in the TL that iv) meets the requirements of the end user and the target culture. Though all these definitions may seem like they leave no room for misunderstandings, the key element that complicates an agreed definition is "the requirements of the end user". With the constant time and cost pressures of the language services industry, new use cases have appeared where a "good enough" translation suffices, such as in the software industry where unedited and raw MT is applied to the user interface to cut costs and speed up the release of certain products to the market (Schmidtke and Groves 2019). Consequently, there is no gold standard measure of quality (Way 2018), and, as commented above, the understanding of "translation quality" will vary depending on what the client prefers, whether it is a translation for assimilation or dissemination (Ginestí and Forcada 2009). However, translation quality is a concept widely adopted in Translation Studies, MT research and the language services industry, and needs to be measured and evaluated (Sánchez-Gijón 2014).

2.4.1. Translation quality evaluation

Language service providers, the users of MT systems or even MT developers need to corroborate that the translations meet minimum quality criteria or that the translations they offer are adequate to the specific end user requirements. To achieve this, translation quality evaluation is essential. This topic has received so much attention in recent years that it is considered a standalone field or branch of research (Secară 2005). Due to the widespread use of MT, researchers from both industry and academia have been celebrating one major international event on this subject since 2006, the Workshop on Statistical Machine Translation (WMT), which ran from 2006 to 2015. After the appearance of NMT, the name of this event changed to Conference of Machine Translation (but still kept running under the name WMT) (https://www.statmt.org/). This event focuses on written MT and normally has two circuits that run in competitions: one that discusses the development of state-of-the-art MT systems and rank the best performing ones, and another that discusses translation quality evaluation, and which are the best practices for translation quality evaluation. There are two forms of evaluation widely accepted by academia and industry: automatic evaluation and human evaluation.

2.4.1.1. Automatic evaluation

When creating an MT system, normally the goal is to be able to translate more content in less time when compared with translation without MT aids. To know whether this will be possible, developers and computer scientists in charge of these tasks need to corroborate and check if their MT systems perform well or not, as well as if the modifications that have been introduced have served to improve, rather than disimprove, the system. In an industry where time pressure and urgency are constant, ideally, this evaluation should be done simply and quickly, to keep pace with society, industry, and the market demands. According to Martín-Mor, Sánchez-Gijón, and Piqué (2016, 40):

> research on MT quality evaluation focuses on tuning quality indices by comparing MT raw translations with human reference translations (also known as a gold standard), or a comparable corpus of text in the target language. If a human translation of the same text exists, each segment is compared in terms of number of editions (insertions, deletions, and substitutions) necessary to convert each segment of the raw translation into the human reference translation. [translation by the author]

In many cases, automatic quality evaluation is the only method used to declare the superiority of one MT system over another, which in turn guides MT development and research (Marie, Fujita, and Rubino 2021). This may be detrimental to the field because automatic metrics have been shown to not correlate very well with human judgements, which are considered the gold standard in translation quality evaluation (Mathur et al. 2020). Although many automatic quality evaluation metrics have emerged, this section only reviews the most common and best performing today, as per Kocmi et al.'s (2021) evaluation.

BLEU (Papineni et al. 2001) is the most used automatic evaluation metric by the MT community, and Marie, Fujita, and Rubino (2021) reported that it had been used in 98.8% of the studies they analysed. BLEU focuses on the order of words or groups of words and calculates how often words or phrases (up to a set of 4 words) match both the human reference and the raw MT output. One of the problems with this metric is that it uses a human sentence as a reference. However, the same sentence in the SL can be translated correctly in multiple ways in the TL. It is therefore possible that a correct sentence or translation may be rated negatively.

Newer automatic evaluation metrics that have been reported to correlate slightly better with human evaluations are chrF (Popović 2015) or COMET (Rei et al. 2020), but they continue to fail to consider language-specific phenomena, such as the particularities of a translated text that has not been written in that language (or translationese) (Zhang and Toral 2019; Graham, Haddow, and Koehn 2020). These automatic metrics also do not work properly when MT systems provide high-quality MT output, or when two systems offer similar translation quality (Mathur, Baldwin, and Cohn 2020). Therefore, human evaluation of translation quality remains the gold standard (Freitag et al. 2021), but it is expensive, is very difficult to reproduce and is very time-consuming, so it is necessary for MT engine developers to have access to a series of quick, cheap, automatic evaluations, which make it easier for them to know whether the engine has improved or not in the system development process, even though their limitations are well known.

2.4.1.2. Human evaluation

Despite being the best practice for translation quality evaluation (Freitag et al. 2021), human evaluation is not free of complications and "there are many design decisions that potentially affect the validity of such a human evaluation" (Läubli et al. 2020, 653). In addition, the increasing quality of the new MT engines is introducing new challenges and difficulties in assessing quality (Rossi and Carré 2022). Therefore, designing a good methodology to evaluate MT has been one of the main objectives in the translation technology community over the past decades, and multiple methods of human evaluation have been proposed (Moorkens et al. 2018). The most important ones are briefly described below.

Human evaluation via relative ranking is one of these evaluation methods. In this first method of evaluation, evaluators get the original sentence in the SL and several choices of MT systems in the TL. Then, they assign a quality order and relatively rank which systems work best. For example, system A is the best, system B is the second best and system C is the worst (Koehn and Monz 2006; Bojar et al. 2018). A drawback of this evaluation method is that information about the extent to which system A is better than system B is not obtained, nor are the ways in which one system is better than another.

Another method for human evaluation is conducting a direct assessment, where evaluators obtain the original sentence in the SL and view the translated sentences (produced by MT or translators) one at a time. Evaluators then must assign a score from 0 to 100 to each translation. This method allows them to know which engine is better and, moreover, to know to what degree it is better. In addition, to avoid problems of subjectivity and scoring between the different evaluators, instructions for standardising the evaluator criteria are normally prepared (Graham et al. 2013). In 2022, this evaluation method was established as the standard evaluation methodology of the WMT evaluation campaign (Kocmi et al. 2022). Yet, even if the translations were assessed on a 0 to 100 score, the evaluators were shown a scalar metric from 0 to 6, so that evaluation results were more consistent (Ibid.).

Arising from two EU-funded research projects, QTLaunchPad and QT21 (https://www.qt21.eu/), the Multidimensional Quality Metrics (MQM) framework was created. MQM is a translation quality evaluation model that homogenised multiple previous assessment models and works by identifying the errors in the text. MQM divides errors into different categories and various levels of penalties, depending on the severity of the error in

its context of use. The MQM model has evolved over time hand in hand with the language services industry and was updated with initiatives such as TAUS' MQM-Dynamic Quality Framework (MQM-DQF) (O'Brien 2012a; Görög 2014). Currently, the widely used version is the MQM Core (https://themqm.org/), which covers errors of seven classes (Accuracy, Linguistic conventions, Design, Locale convention, Style, Terminology and Verity) with four different penalty levels (None, Minor, Major and Critical). A more detailed overview of MQM-Core can be found in https://themqm.org/. It is one of the most used human evaluation methods, as well as one of the best recognized ones in academia and industry (Freitag et al. 2021), but it is extensively time- and cost-consuming.

The last human evaluation method for translation quality evaluation reviewed in this section is the Adequacy and Fluency assessment. It is a well-established evaluation method in the MT community (Koehn 2010), and both the assessment of adequacy and fluency normally take place at the same time. Adequacy is considered as "how much of the meaning expressed in the gold-standard translation or the source is also expressed in the target translation" (Linguistic Data Consortium), and needs to be assessed by displaying the target translation with a reference translation or the source text. Normally, adequacy can be scored on whether the target translation expresses none, little, most or everything of the meaning of the source sentence or reference translation. An appropriate adequacy evaluation requires the knowledge of at least two languages. On the other hand, Fluency is understood as to what extent the translation is "one that is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker" (Linguistic Data Consortium), and can be assessed by only displaying the target translation. Thus, an adequate fluency evaluation only requires monolingual knowledge. Normally, fluency assesses whether the target translation is incomprehensible, disfluent, or if the fluency of the target text is good or flawless. A clear distinction is made between adequacy and fluency because a translation may be flawlessly fluent (that is, the translation may be perfectly formed and follow all the target language rules), but may have adequacy problems (e.g., may contain only partially the meaning of the source sentence). In the adequacy and fluency evaluation, a 4-point Likert scale is normally preferred to a 5-point Likert scale to avoid the evaluators' tendency to select the central point

and force them to evaluate the translation towards the positive or the negative end (Rossi and Carré 2022).

After this review, we can observe that there are multiple methods for evaluating translation quality, regardless of whether this translation is produced by a translator or a machine, and that each method is more appropriate for obtaining different information. As human evaluation is expensive, non-reproducible and time-consuming, the evaluation method must be selected depending on the goals of the research and the budget available. If we only want to know whether one system is better than another, perhaps a relative ranking evaluation is sufficient. However, if we intend to have the most granular results possible and to know in depth where a translation fails and the severity of the failures, MQM analysis is the most complete method of human evaluation of translation quality. However, MQM is very time-consuming and expensive. Therefore, the adequacy and fluency evaluation emerges as a standard option that provides information on which system is the best, the difference between one system and another, while being less time- and cost-consuming.

The most important element in human evaluation today is that evaluators must be professional translators with expert knowledge of the languages being evaluated (Läubli et al. 2020). It has been shown that human evaluation with crowdsourcing volunteers, people with basic knowledge of a language or students do not provide accurate results because these evaluators tend to rate wrong translations as adequate and overlook errors (Freitag et al. 2021; Moorkens et al. 2018).

2.5. Translation productivity

Since the emergence of the first translation technology applications, translator productivity has been an interesting element to consider (Elliston 1978). Knowing how much content a translator could translate in a specific period of time makes it possible to find out how much their time was worth and to calculate payment.

As a consequence, it is undoubtedly with the introduction of translation technology applications that research began to investigate whether they helped translators to work faster and translate more content in less time, without significant impact on quality. This happened with CAT tools, which not only allowed translators to work faster, but also allowed

the resulting translation to be of higher quality by allowing for greater terminological consistency (Bowker and Fisher 2010). When post-editing was in its infancy, it was common to analyse whether the introduction of MT improved translator productivity. Some of the early studies on post-editing productivity compared to non-computer-assisted translation reported that there was no statistically significant difference between these two workflows (Carl et al. 2011; Garcia 2010). However, the authors indicated that the participants in the study did not have post-editing experience, which could substantially impact the results. With the general adoption of post-editing in the industry, new studies appeared that indicated that post-editing allowed translators to be more productive if compared with their non-computer assisted translation productivity or their computer-assisted translation productivity (Guerberof-Arenas 2008; Plitt and Masselot 2010).

Currently, although there is no generic agreement or standard measure of the productivity improvement provided by post-editing over human translation (Terribile 2023), it is widely accepted that MT substantially helps to translate faster in terms of words per hour (WPH) (e.g., Kosmaczewska and Train 2019; Sánchez-Gijón, Moorkens, and Way 2019), and that this will depend on multiple factors such as the experience translators have in post-editing (Guerberof-Arenas 2008) or the quality of MT systems, which is highly dependent on the domain of the text or the language combination being worked on (Koponen 2016).

This Chapter has presented a literature review of state-of-the-art translation technologies at the date of writing, as well as an overview of the general key concepts in the language services industry. The following chapters will analyse more in-depth the key concepts and areas of this research work.

CHAPTER 3. INTERACTIVE POST-EDITING

This chapter provides a thorough description of IPE, a post-editing modality that is the central topic of this PhD thesis. Section 3.1 introduces IPE. The historical context of IPE follows in Section 3.2 and explains how this post-editing modality has evolved over time and presents the different user evaluations of IPE and the lessons learned from these evaluations. Then, Section 3.3 covers a discussion on IPE and the potential this post-editing modality may have in today's language services industry.

3.1. Introduction to interactive post-editing (IPE)

With the boom in the use of MT, a rising interest in improving the quality MT offers and in increasing translators' productivity through post-editing (as described in Chapter 2), research has explored different ways for introducing MT in translation production workflows. Section 2.3.1 covers the use and application of TPE, but there is a post-editing modality that has received less attention than TPE and is worth exploring.

When talking about IPE, a certain level of terminological chaos can be observed in the literature because multiple terms are used for the same or very similar concepts. Terms are used such as "Interactive Machine Translation (IMT)", "Interactive Translation Prediction (ITP)", "Interactive Post-Editing (IPE)" or "Adaptive MT". Therefore, to facilitate understanding and homogenise the use of terminology in this study, the following terms have been used:

- "Interactive MT (IMT)" is used to talk about a technological feature that can be applied to different MT architectures. MT architectures such as RBMT, SMT or NMT calculate the probabilities that a text in a SL will be translated in a certain way in the TL through different algorithms (see Section 2.2). In contrast, an IMT system takes into account the input of the translator and updates the MT output proposals in real time (Barrachina et al. 2009). In other words, an IMT system has a text prediction and translation completion proposal feature activated. Thus, in this PhD thesis, the term IMT is used for this interactivity feature provided by different MT workbenches, regardless of the underlying architecture of the MT system.

- "Interactive Post-editing (IPE)" is used to talk about a post-editing modality that uses IMT. Unlike TPE processes (see Section 2.3.1), where the MT output is static, in IPE processes, the MT output is updated in real time as the translator writes. IPE tasks are conducted within an IMT system.
- “Adaptive MT” is used to talk about another technological feature that can also be applied to different MT architectures and systems. Regardless of the underlying MT architecture, a system with adaptive MT learns from the corrections of the translators and fine-tunes the MT system in real time with validated proposals (Green, Heer, and Manning 2013; Denkowski et al. 2014; Bentivogli, Bertoldi, et al. 2016). Translators using a system with adaptive MT do not need to correct the same mistakes repeatedly because the MT output is updated considering the translations that the translator has already validated.

Thus, IPE involves an intrinsic interface interaction that works in a similar way to the functions of predictive keyboards that are frequently included in some mobile tools or programs, such as Gmail. In TPE processes, the MT engine offers a static MT output proposal, and the translator identifies errors and modifies them. In IPE, the MT system suggests different MT output proposals as the translator writes in real time.

To use an example, in Figure 3.1 (Peris and Casacuberta 2019), we can see the iterative process of an IPE task from English into French. “*They are lost forever”* is a sample sentence that works as the source text. “*Ils sont perdus à jamais”* is one correct translation of the source text and works as the target text. “*IT-”* means each iterative step the IMT system takes during an IPE task. “*MT”* refers to the raw MT proposal at each iteration, and “*User”* means the feedback introduced by the translator. The boxed word is the amendment that the translator makes to the MT proposal. The MT output fragments validated by the translator are marked in green.

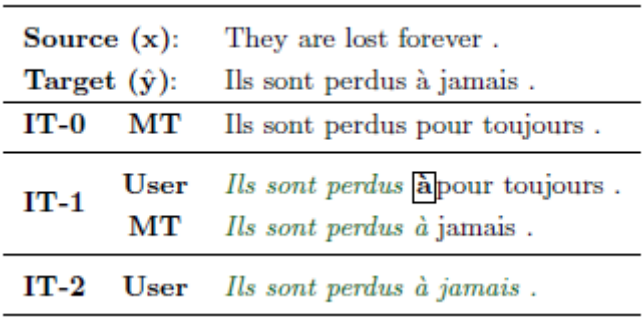

Figure 3.1. IPE process. Reprinted from "Active Learning for Interactive Neural Machine Translation of Data Streams" (Peris and Casacuberta, 2019).

Specifically, when the English sentence "*They are lost forever"* is sent to the MT system, the system suggests "*Ils sont perdus pour toujours"* in French as a first MT proposal, which is not the appropriate translation for the source text. Then, as we can see in "*IT-1"*, the translator accepts "*Ils sont perdus"* as a correct translation, but then introduces "*à"* (the boxed word). Here, the IMT system accepts the validation of the translator, re-runs its algorithm, and produces a new MT proposal, "*Ils sont perdus à jamais"*, which the translator accepts. Thus, in an IPE task, the workflow is like interacting with the predictive text of a mobile phone and considers both the input sentence and the corrections of the translator. Time is a key factor when re-running these IMT systems used in IPE tasks. Thus, with today's algorithms and available computing power and capabilities, the time for re-running the IMT system and offering new MT proposals decreases, and, thus, IPE has become more attractive (Peris and Casacuberta 2019).

Although IPE is a relatively new post-editing modality and has not been given the same attention as TPE, some research on the topic has already been undertaken from different points of view. On the first hand, from the user perspective, a comparison of TPE and IPE was conducted to analyse which was the best post-editing modality for being introduced into translation production workflows by looking at translation quality and translation productivity (e.g. Alabau et al. 2016; Sánchez-Torrón 2017). On the other hand, from the technical perspective, analyses were conducted to see which were the best techniques to run an IMT system (e.g. show the updated MT proposals completely or only partially, etc.) so that translators doing IPE tasks could benefit the most from the IMT feature (Bender et al. 2005;

Barrachina et al. 2009; Koehn and Haddow 2009; Peris and Casacuberta 2019). In this Chapter, I leave aside the technical perspective, and put the attention on the user studies conducted through the historical development of IMT and IPE.

3.2. Historical context

To understand the birth of IMT and IPE, we have to go back to the MIND system (Kay 1970), which was the first machine-assisted translation system that learned from the input of the user. Its main objective was to allow translators to translate very complex texts, with the help of the machine, thanks to different questions related to word order, pronominal references, or prepositional and sentence constructions. The MIND system would automatically pose questions about the source text, and then consider the answers of the translators to these questions to offer MT proposals. Therefore, it can be said that, in this system, the translator was the one who assisted the machine to disambiguate the source text so that the machine produced a target text. There were multiple systems at that time following the same logic (Brown and Nirenburg 1990; Blanchon 1994), but none of them was commercially viable because the question-and-answer process was too time-consuming, and the traditional method of translation was still preferred.

In 1993, Church and Hovy (1993) analysed the methods of MT evaluation, as well as the path MT research was taking at that time with systems like MIND. In their study, they proposed a feature that they thought could be very useful for translators and the MT field: a sort of word prediction system that would have a "Complete" button so that, when the button was pressed, the system would complete the rest of the word that the translator had started to type. The goal of this feature was that translators could write texts much faster and more productively (Ibid.). This feature had a great relevance in the IMT systems that were later developed for IPE tasks.

3.2.1. TransType and TransType 2

Foster, Isabelle, and Plamondon (1997) indicated that the question-and-answer process that systems like MIND implemented required a lot of effort, so it was still more efficient to translate without computer-assisted translation aids. They took Church and Hovy's feature

into consideration and proposed a system called TransType (TT_0[1]), which is considered to be the first IMT system in the context of computer-assisted translation. TT_0 offered one-word completions to the text the translator wrote by using SMT models. Foster, Isabelle and Plamondon had previously worked on TransTalk (Brousseau et al. 1995), which was a speech recognition system that worked through the dictation of the user and used the same text prediction feature. Two years later, Foster, Isabelle, and Plamondon (1997) used the term "modern IMT" for TT_0 and referred to systems like MIND as "classical IMT". They argued that this new IMT system could save up to 70% of a translator's keystrokes in producing the text. This represented a major paradigm shift in machine-assisted translation systems that learned from the input of the user, as this meant a change from systems in which the translator made it easier for the machine to understand the source text by answering a series of questions — the machine is therefore at the centre of this process— to systems in which machines assisted translators in their work and increased their productivity in producing the target text —where the human becomes the key factor in this interaction. The focus of the interaction also changed from understanding the source text to forming the target text. Langlais and Foster (2000) did an evaluation of TT_0 by running an automatic word completion session to calculate the number of keystrokes that a hypothetical user could save when using the proposed IMT system. In this session, a simulated user wrote the target text character-for-character, accepting the proposed completion of the system as soon as it was useful. Results showed that around 66% of the number of keystrokes could have been saved in comparison with translation without computer-assisted translation aids. Nevertheless, it must be stated that this evaluation was fully automatic, the results were theoretical, and the authors were aware of the limitations. During the translation process, a translator may not always choose the first valid translation completion proposal and may insert or delete characters or words on multiple occasions before producing a final translation in the TL.

[1] Though the authors of TransType only use the term "TransType", its different versions are assigned a number in this literature review. This way, readers can understand its chronological evolution.

3.2.1.1. TransType: Evaluation 1

In 2000, Langlais, Foster, and Lapalme (2000) presented TransType 1 (TT_1), the first prototype of an IMT system (based on the work by Foster, Isabelle, and Plamondon (1997)), in the framework of the TransType research project. This research project was funded by the Natural Sciences and Engineering Research Council of Canada. They wrote that TT_1 worked as follows:

> A translator selects a sentence and begins typing its translation. After each character typed by the translator, the system displays a proposed completion, which may either be accepted using a special key or rejected by continuing to type. Thus, the translator remains in control of the translation process and the machine must continually adapt its suggestions in response to his or her input. (Langlais, Foster, and Lapalme 2000, 1)

In TT_1, instead of using one-word completions like in TT_0, the system could offer multiple-word completions, and had improved language models and a more realistic user interface, like today's CAT tools (see Figure 3.2). After the release of this prototype, a series of evaluations of TT_1 were carried out.

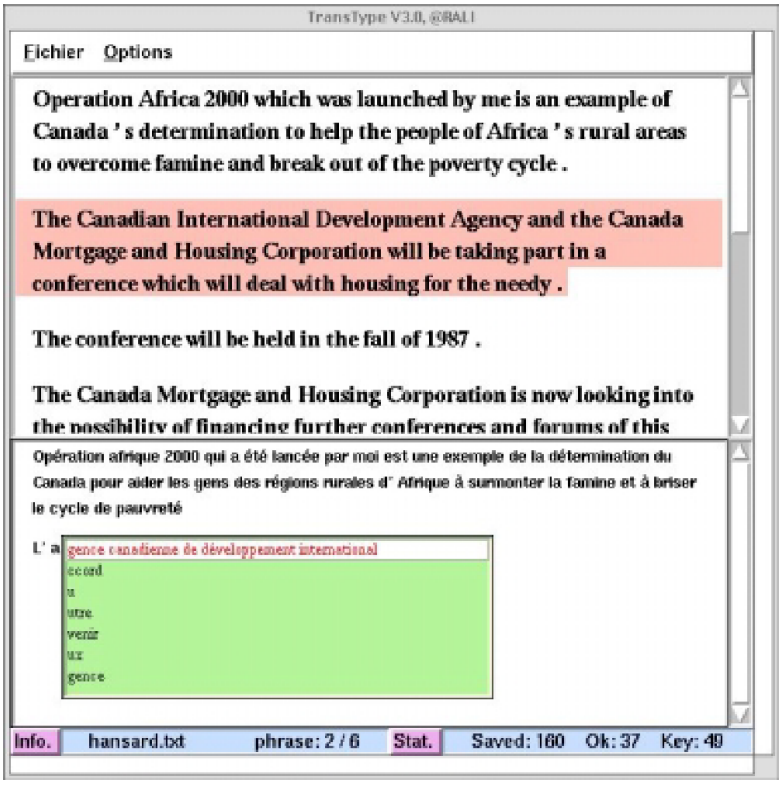

Figure 3.2. TransType (TT_1) interface. This figure shows an example of an interaction with the tool. Source text segments are in the upper part of the screen. In the lower part of the screen, translators type and a dropdown menu with the completion proposals appears. Figure obtained from Langlais, Loranger, and Lapalme (2002).

Langlais et al. (2000) did the first user evaluation of TT_1 (TT_{1a}[2]) by using ten voluntary translators (four professional translators and six students) in a three-step study.

- In the first step, translators had five minutes to get used to TT_{1a}'s text-editor and get to know its GUI and operations, which were the same as a normal text-editor. This first step yielded the "natural" typing speed of the translators.
- In the second step, translators had 20 minutes to conduct an IPE task with TT_{1a}'s IMT feature activated with completions at a word-level.
- In the third and last step, translators had to do IPE tasks with TT_{1a}'s IMT feature activated with multiple-word completion proposals.

Finally, the authors did a 10-minute feedback survey to collect the suggestions and feelings of the translators. After the study, authors stated that nine out of ten translators thought they worked faster when doing IPE, but only one really did so. Results also showed that most translators were less productive when doing IPE than using only the text-editor, as general productivity went down by 35% in IPE in comparison with the non-computer-assisted translation typing speed. The authors suggested that this happened because translators had to read the multiple MT completion proposals and then decide whether to accept them or not, instead of just writing the target text.

3.2.1.2. TransType: Evaluation 2

Langlais, Lapalme, and Loranger (2002) did the second human evaluation of TT_1 (TT_{1b}) through a slightly different three-step study with nine translators.

- The first step consisted of a period from five to eight minutes where translators had to translate in TT_{1b}'s text-editor without any type of computer-assisted translation aids (this stage measured the "natural" typing speed as in TT_{1a}'s evaluation).
- In the second step, translators had from 15 to 20 minutes to conduct an IPE task.

[2] There are two different human evaluations for TT_1. Thus, in this literature review I will use TT_{1a} and TT_{1b} to facilitate the comprehension of the reader.

- The third and last step consisted of a period from five to eight minutes where translators had to translate via IPE together with a special lexicon/glossary. This was a newly added feature.

Though Foster, Isabelle and Plamondon (1997) calculated that TT_0 could save around 66% of the keystrokes of the translator after an automatic evaluation, Langlais, Lapalme and Loranger (2002) found in TT_{1b}'s user study that translators only saved 31% of the keystrokes because they did not take into account most of the completion proposals. According to the questionnaire responses, almost all translators stated that they thought TT_{1b} improved their translation productivity. Yet, the study revealed that their IPE productivity went down by 17% (an improvement in relation to the 35% decrease of TT_{1a}) if compared with their non-computer-assisted translation productivity.

In addition, as the second step of the evaluation study was composed of 20 minutes, the authors compared the first 10-minute period against the second 10-minute span. This productivity loss decreased from 17% to 10%, which indicated that, when translators got acquainted with the IPE modality, their productivity increased. This also suggested that translators involved in IPE may require a longer learning curve because they are learning to use the system. By analysing the cognitive load (by studying the pauses translators took), results showed that short completion proposals (i.e., two to three letters) distracted users because they had to read them, and the authors suggested that completion proposals must be long enough to imply a time saving, though they had not studied the appropriate length at that time of the study.

3.2.1.3. TransType 2

Some years later, TransType received funding from the EU for an R&D project from 2002 until 2005. The project was then renamed to TransType 2 (TT_2) (Esteban et al. 2004; Macklovitch 2006) (see Figure 3.3). This time, the participants were three university research labs (RWTH in Germany, ITI in Spain, RALI in Canada), an industrial research partner (XRCE in France), an administrative coordinator (Atos Origin in Spain) and two translation companies (Société Gamma in Canada and Celer Soluciones in Spain).

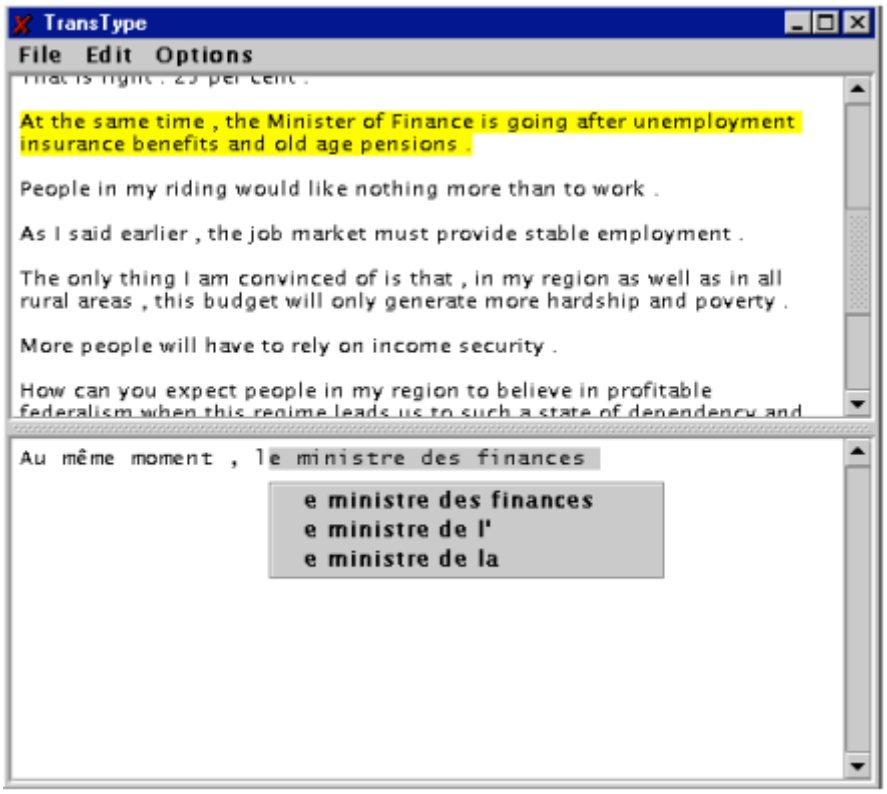

Figure 3.3. TransType 2 (TT2) interface. This is an example of an interaction with the tool. As in Figure 3.2, source text segments are in the upper part of the screen. Translators have to write the target for the chosen source segment in the lower part of the screen. A dropdown menu is shown with the completion proposals. Figure obtained from Langlais and Lapalme (2002).

The evaluation of TT_2 consisted of five evaluation rounds (ER1 to ER5) that took place in the premises of the participating translation companies, and two language combinations were studied: English to French and English to Spanish. Each ER lasted two weeks, where translators spent half-days translating texts ranging from 2,000 to 2,500 words using TT_2. The system used log-files that time-stamped every translator action, so researchers could see the time translators spent typing, pausing, as well as their operations (e.g., cut, paste, delete actions). In addition, TT_2 had special sessions that the authors called "dry-run sessions", where translators used TT_2's text-editor without any computer-assisted translation feature activated, so the baseline translation productivity figures of the users could be obtained. The ERs took place as follows:

ER1 and ER2 were preparatory, so that translators could familiarise themselves with the new features TT_2 included and get used to the GUI they would be working with.

ER3 included four professional senior translators (two per evaluation site). It was the first time that translators interacted with all the features of TT_2 in real-life working conditions. ER3

lasted five half-days. One half-day was used as a dry-run, and the other four half-days were used for translating using all the features of TT_2. In these last four days, translators tested two different configurations of the system: one offering shorter, multiple completion proposals (i.e., proposing one-word completions and two-word completions) for each word that translators typed; another offering only one completion of the whole sentence that translators had started to type. Translators suggested they felt more comfortable with the latter configuration, as they had to read fewer proposals and they therefore lost less time reading and evaluating the completion proposals. Three out of four translators surpassed their dry-run productivity in at least one text in an IPE task with TT_2, though the authors do not state by what factor.

In ER4, a senior translator was added to each evaluation site. Thus, this ER evaluated the translations of six users (all with the same profile, senior translators). ER4 consisted of ten half-days instead of five, as TT_2's team wanted to increase the amount of data they were collecting. These ten working days were divided as follows: (i) one half-day as a training to refresh translators because ER4 took place some months after ER3; (ii) one half-day as a dry run, to obtain the normal typing speed of the six users; and (iii) eight half-days of IPE with all the features of TT_2. Results of ER4 revealed that five out of six translators surpassed their dry-run productivity in at least seven of the eight texts in the IPE modality. In addition, 50% of translators surpassed their dry-run productivity in all texts. To ensure that the resulting translation quality was good, an independent revision was requested, which demonstrated that the six translators produced deliverable quality translations. In ER4, the average productivity gain in the IPE modality was at 20% if compared with translators' dry run productivity.

Regarding ER5, the final evaluation round, the evaluation protocol was almost the same as in ER4. The only new aspect was that Macklovitch (2006) added a second half-day dry-run near the end of the 10-days period, to counter the argument that dry-run figures were measured as a baseline in ER4. ER5 data demonstrated a constant average productivity ratio between ER5 (996 words per hour) and ER4 (1,005 words per hour). Yet, Macklovitch stated that there was a methodological problem in the second half-day dry run. If only the first dry run of ER5 was considered as the baseline productivity figure, the average productivity gain during ER5's translation sessions while doing IPE amounted to 12.53%. Nevertheless, if we only considered

the second dry-run of ER5 as the baseline figure, translators' productivity decreased by 23.18% when doing IPE in comparison with non-computer-assisted translation. The author stated that this was caused by a methodological mistake because the text used in the second dry-run had different complexity (i.e., shorter sentences, easier terminology) than all the other texts. As a conclusion, and according to the author, gains in productivity ranged from 12.5-to-20% through all the ERs if this last second dry-run of ER5 was not considered.

Moreover, TT_2 had a shortcut that stopped the clock ticking and allowed users to introduce comments because one of the goals of TT_2 was to know what the translators thought of using the system during the interaction. Macklovitch highlighted two general comments given by translators. In the first place:

> When the system's initial prediction on these sentences was not to the translators' liking, they would modify it a first time; and later, when that same sentence re-occurred within the file, they found they had to make the same corrections over again. This was something they did not at all appreciate, as they made very explicit in their comments. […] the system needs to incorporate a simple string matching and repetitions processing capability like that found in most commercial translation memory systems. (Macklovitch, 2006, 4-5)

The second general comment was along the same lines as the first one, which suggested that the amendments translators made were not being considered:

> There is a good likelihood that TransType will reproduce the problem which the translator initially corrected every time it reoccurs, since the system's underlying language and translation models remain unchanged during a working session. This too, the participants found particularly frustrating. "Why can't the system learn from my corrections?" they asked over and over again. (Macklovitch, 2006, 5)

Thus, translators were asking for an MT system that learned from the corrections, and that could take into account the amendments users made when going through previously translated text: that is, a system with adaptive MT. As suggested in Section 3.1, adaptive MT features would not appear until 2013.

Finally, after stating that TT_2 offered a 12.5-20% productivity increase in comparison with non-computer-assisted translation, the authors concluded that translators would not use this tool in their daily work, and that MT engine performance was not the only important aspect, but

also the fact that translators had to correct the same mistakes repeatedly. TT_2 research was then abandoned.

3.2.2. CAITRA

Based on the TransType and TransType 2 projects, Koehn (2009) developed a new IMT system named CAITRA. It was an online, web-based tool, which was accessible via the Internet. CAITRA's goals were to continue what the previous TransType's research projects started, that is, to explore the benefits of IMT aids to translators, to analyse user behaviour in IPE, and to develop new types of computer-assisted translation assistance. CAITRA was powered by the open-source SMT system Moses (Koehn et al. 2007).

CAITRA was developed with the user in mind because it tried to be an easy-to-use tool, and translators or researchers could easily set up a project or upload a file into a text box, and the system itself would pre-process the file, divide the text into segments and then show the source text in the same way a CAT tool does. In addition, CAITRA had a series of features that could be easily activated or deactivated according to the preferences of the user. These were as follows:

On the one hand, CAITRA included a text prediction and word completion feature, which Koehn (2009) named as IMT too, following the term coined in TransType. This feature used a similar text prediction and word completion system to that of TransType but had an improved SMT engine. The completion proposals only included a few words to avoid overloading the translators, so they did not have many proposals to read or evaluate. Koehn stated that neither the optimal length nor the best location for the proposals had been studied at the moment of the launch of CAITRA, which typically proposed the completions in less than one second. In addition, CAITRA offered up to ten completions, and translators could directly click on them to get them automatically inserted in the target text box. Completion proposals were colour-coded and received an automatic score, based on the probability of being the correct translation. This figure was extracted by the SMT model.

CAITRA also included a TPE feature. In case researchers using CAITRA preferred to use the TPE approach or wanted to measure the TPE typing speed and compare it against the IPE typing speed, they could activate this second feature. When activating the TPE feature, the raw MT

proposal was already inserted in the target text box when the translator selected a segment (see Figure 3.4).

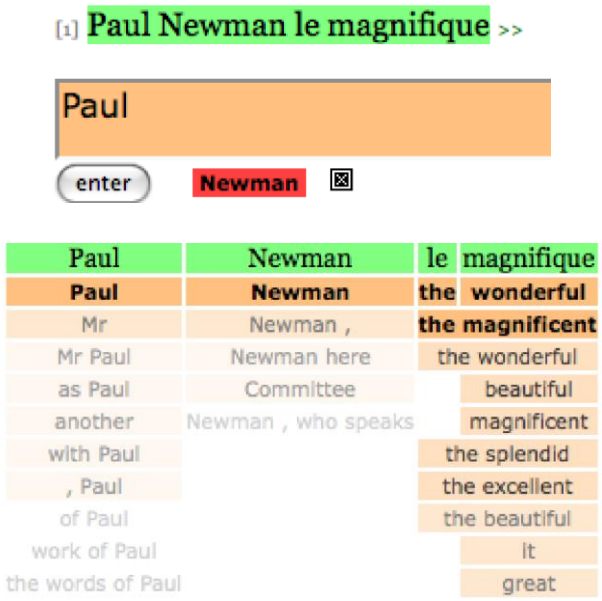

Figure 3.4. Screenshot of the interface of CAITRA. In the upper part of the image, the source text and the text box for the target text can be observed. A 1-word completion proposal appears in red: "Newman". In the lower part of the screenshot, the table with the colour-coded and scored MT options can be seen. If these options had more probability to be correct (according to the translation model), they appeared on top of the table; if the probability was inferior, they appeared in lower cells. Image retrieved from Koehn (2009).

Finally, the last and third feature that CAITRA included was key and time logging. The tool tracked every keystroke and mouse click of the translators and assigned a timestamp to these actions to facilitate the analysis or the study of the behaviour of the users, regardless of whether they did TPE or IPE tasks.

3.2.2.1. CAITRA: user evaluation

The first user study of CAITRA was also carried out by Koehn (Koehn 2009a). The study hired ten university students, who were paid a fixed amount to participate as non-professional translators in a French-to-English translation experiment. These ten students were divided into two groups: five native speakers of French (with a university-level of English) and five native speakers of English. As CAITRA was a web-based tool, students were allotted a period of two weeks to complete their assignment whenever they could. Koehn comments “this is also the first direct comparison of post-editing and IMT methods” (Koehn, 2009a: 4). In other

words, it was the first direct comparison of TPE and IPE because all the previous evaluations analysed IPE against non-computer-assisted translation. The ten students translated the same texts, which were composed of 192 sentences from French news stories into English, in five different conditions. To do so, all the texts were divided into five blocks of about 40 sentences and 1,000 words each. This meant that all students were translating into English, although five out of ten were not English native speakers.

In the first condition, students had to use CAITRA unassisted, that is, as if the tool only included a normal text-editor. In the second condition, participants had to translate the texts of the block through a TPE approach. In the third condition, CAITRA offered different MT translation completion options in a table, as shown in Figure 3.4, which students could use as a reference to boost their productivity and reduce their cognitive load. In the fourth condition, the IMT feature was turned on, and participants could accept the completion proposals by pressing the tab key or reject them by continuing to type. In the fifth and last condition, condition three and four were mixed, both showing the table with different MT options, and offering the IMT proposals activated, so that participants could do an IPE task.

As CAITRA automatically logged the keystrokes and time spent on each sentence, Koehn (Koehn 2009a) did a more profound study on the following variables: typing speed, translation quality, and assistance.

On typing speed, the average time per input word obtained from CAITRA's log was computed. This was the measure analysed for translation productivity. This figure ranged from 3.1 to 3.9 s/word. Regarding quality, as 10 different participants would produce 10 different translations, and all of them could be valid and correct, human evaluators were used to assess them with the following instructions through a web-based evaluation tool. The instructions for the judges were as follows:

> Indicate whether each user's input represents a fully fluent and meaning-equivalent translation of the source. The source is shown with context, the actual sentence is bold. (Koehn, 2009b, 10)

Evaluators could only evaluate whether the translation was "Correct" or "Wrong", which may be a weak indicator of translation quality because, as stated in Section 2.4, translation quality

is not an absolute value, and a translation may have different mistakes, good translated chunks or different levels of issues. In the study, Koehn only stated that judges were fluent both in French and English, but there is no mention of their translation skills or experience, which raises significant questions about the results. Language fluency is a weak indicator of language knowledge and translation skills. Further quality evaluation methods have been proposed later, taking into account all the aforementioned aspects from Section 2.4 (e.g. O'Brien 2012; Görög 2014).

The main research question of Koehn's paper was "Do translators produce better translations and are they faster than when unassisted?". However, it is worth stressing that obtaining reliable results from students using a tool they were completely unfamiliar with in a task that they had little to no experience in may be difficult. In the results, eight out of ten students were faster and better with TPE; six students were faster and better in the sentence completion plus options condition; and four students were faster and better in the IPE modality. Only two students achieved no gains with any assistance.

By taking a closer look at the results concerning the pauses, students were divided in three groups: (i) slow participants, who improved substantially (both qualitatively and productively) when assisted; (ii) fast participants, who also improved slightly when assisted; and (iii) "refuseniks", who did not use the assistance at all, and saw almost no quality or productivity gains. Some students typed only 10% of the sentence and managed to complete the rest with the sentence completion feature. As an additional, interesting aspect to comment on, a learning curve is observed through the experiment because participants increased their translation productivity (in WPH) after they familiarised themselves with CAITRA. Yet, it must be considered that the sample text was small, and no generalisable conclusions could be demonstrated.

To sum up, on average, participants were faster by 16% when using translation options, by 27% when using sentence completions, by 25% when doing IPE (that is, combining options and completions), and by 39% when doing TPE. Thus, though a productivity gain was demonstrated when doing IPE tasks, TPE was still faster than IPE. After the experiment, participants were asked to rank whether they found helpful and useful the different conditions of assistance in two surveys. TPE received the worst score, while IPE was valued the highest. Paradoxically, even if TPE was ranked low in terms of enjoyment and usefulness,

it proved to be more effective than all the other assistance types in terms of translation productivity. Again, it is worth noting that the participants evaluated were students, and some of them were even non-native English speakers, which was the language into which they were translating. These details may influence their evaluation because non-professional or crowdsourcing translators tend to accept more translation errors because they lack translation knowledge (Castilho, Moorkens, Gaspari, Sennrich, et al. 2017; Toral 2020), as commented previously on Section 2.4.1.2 on human evaluation of translation quality, raising serious concerns about the methodological validity of this user study.

3.2.3. CASMACAT

In 2011, the European Union funded a research project from 2011 to 2014 whose aim was to develop a new IPE workbench. The main research partners from this new R&D project were three universities —University of Edinburgh, Copenhagen Business School, and Universitat Politècnica de València— and one language service provider —Celer Soluciones. In 2013, Alabau et al. (2013) presented CASMACAT, an open-source workbench with the aim of investigating new types of computer-assisted translation for professional translation use.

As with previous IMT systems, CASMACAT had a series of different configurations, and users could try which one to turn on and use. The default mode was a regular CAT tool that used an SMT engine and allowed translators to post-edit MT or TM segments with a TPE workflow. The advanced mode was named “intelligent autocompletion” by Alabau et al. (2013) and was mainly an IPE system. For this type of feature, CASMACAT developers use the term “interactive translation prediction” (ITP).

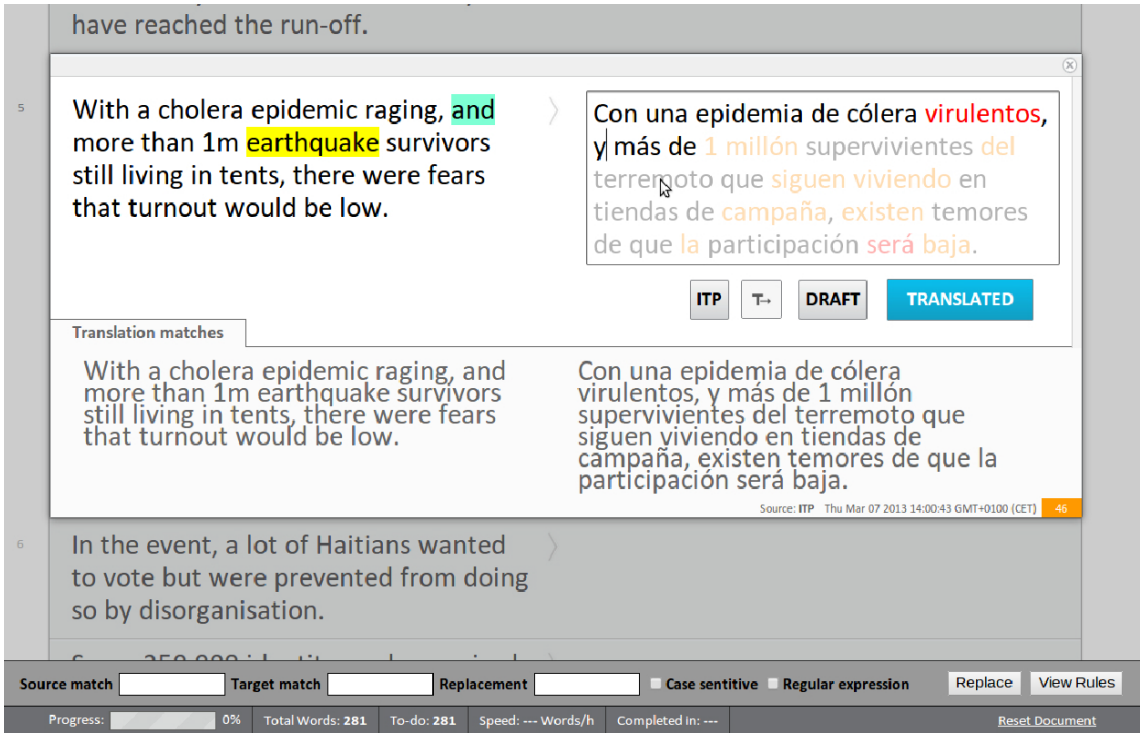

Figure 3.5. Screenshot of CASMACAT's graphic user interface. This image shows the user interaction with all the features of the system turned on: i) predictive text and sentence completion proposals; ii) coloured confidence measures; and iii) segment alignment information. Image retrieved from Alabau et al. (2013).

CASMACAT also offered different, specific CAT tool features, as shown in Figure 3.5. The first one was confidence measures, which indicated with a red and orange colour the probability that different chunks of the proposed translation completion were likely to be incorrect or dubious, respectively (one of the early implementations of a technology currently known as quality estimation (Specia, Raj, and Turchi 2010)). CASMACAT also included typical CAT tool features such as Search and Replace and sentence segmentation or alignment. Regarding the IMT feature, there was a new prediction length functionality added, which only showed text completion proposals until the system found a word with low confidence measures (that is, coloured in red or orange). Nevertheless, the user could press a button to see the complete sentence completion proposal, avoiding this prediction length feature. The already post-edited text was shown in black, while the rest of the proposal had a faded-grey colour, as seen in Figure 3.5. As the main goal of researchers was to assess the productivity of translators while using this new workbench, CASMACAT also included an eye-tracking, key logging, and replay feature, which registered the activity of the user in a detailed way and allowed for the replaying of the translation session of a specific user. This allowed for the study of the translation process and not only the final translation.

3.2.3.1. CASMACAT: First Progress Report

The first user evaluation of CASMACAT was presented as the first progress report sent to the European Union for the R&D project (Alabau *et al.*, 2013). In this study, the authors carried out a satisfaction survey among 16 translators that had performed post-editing tasks with CASMACAT to record what they thought about the workbench. In this report, Alabau et al. stopped using the term "ITP" and started using "IMT" once again. This study was performed on a web-based platform containing the CASMACAT system because it was easier for the translators to access the test. The GUI was based on an open-source, web-based CAT tool named MateCat, which originated from another R&D project funded by the EU (Federico et al. 2014). The project results from CASMACAT and MateCat were later merged, and the resulting combination is now a commercial project under the name MateCat managed by a language service provider called Translated.[3]

The main aim of this first study was not centered on finding out which post-editing modality was more productive (either TPE or IPE) like in previous studies. Rather, this first CASMACAT study aimed to analyse the human-computer interaction and tested four different IPE configurations and features to analyse user satisfaction with the aim of evaluating their potential for including them in a production-ready CAT tool. System 1 was a basic IMT workbench offering text prediction and word autocompletion for IPE tasks. System 2 also offered coloured confidence measures. System 3 included prediction length control, and System 4 included all the possible features of CASMACAT, as defined above.

Sixteen volunteer translators took part in this study, and all of them had a degree in translation studies and used CAT tools regularly, but none of them had previously used an IMT system for IPE tasks. Nine out of sixteen participating translators had previous experience with post-editing. The fact that some translators had experience in post-editing tasks may be a design flaw, as this could affect their perceptions and acceptability of TPE tasks against IPE

[3] https://translated.com/welcome, last accessed on the 29th March 2021.

ones. The texts used in this first user trial were non-specialised pieces of news in the English-Spanish language combination.

As for the evaluation methodology, Alabau et al. (2013) conducted an introductory system usability scale (SUS) questionnaire to collect quantitative data on user satisfaction. Translators had to rate their post-editing task satisfaction from 1 to 5 after using each of the different system configurations. Score 5 denoted the highest satisfaction and score 1 the lowest. Translators also had an area to include comments on any aspect they wanted to highlight, if applicable. Translators always evaluated System 1 in the first place, which was the baseline IPE configuration. The post-editing tasks including Systems 2, 3, and 4 were assigned randomly to minimise the order effect on user satisfaction due to the learning curve or fatigue effect, and the translators had no time limit to perform the evaluation. Once the translators had performed all the IPE tasks and completed all the questionnaires, a last questionnaire was then used to capture their thoughts on the system globally.

After analysing the results, System 3 (IPE with prediction length) was rated highest with a satisfaction score of 3.3 points out of 5. In the comments, translators stated that these features allowed them to stay more in control of the post-editing process. None of the translators rated System 1 (the baseline) above 3. There was no mention of statistically significant differences in these results. The assessment order may be of interest for further evaluations, as System 1 was evaluated at the start, and the order effect may have caused translators to rate it lower. A warm-up session could help avoid this effect, according to the authors of the questionnaire. Translators also commented that, when they familiarised themselves with the system, it became less cumbersome and complicated, an aspect that is in line with the studies of CAITRA (Kohen 2009), Barrachina *et al.* (2009) and Casacuberta et al. (2009). Again, this user study was different to those carried out up to that time, as Alabau and colleagues had the intention of evaluating the satisfaction of users when interacting with an IMT workbench through IPE, and not its productivity.

3.2.3.2. CASMACAT: Second Progress Report

In 2014, Sanchis-Trilles et al. (2014) presented the second progress report of the CASMACAT project, where they exhibited the results of the second field trial. In this second user study,

they compared TPE and IPE in CASMACAT. In this report, authors used “ITP” again to refer to IPE.

Nine professional translators and four reviewers participated in this second field trial. All of them were Spanish native speakers and worked regularly on post-editing. The translators worked with news pieces from English into Spanish. Every document had around 1,000 words. For TPE tasks, texts were translated on an SMT system and then loaded in the CASMACAT workbench. Translators received clear MTPE guidelines to make their post-editing criteria homogeneous, so they all had clear instructions on what to or not to edit while having full-publishable quality in mind.

In this report, Sanchis-Trilles et al. evaluated different configurations: (i) one doing TPE; (ii) another doing IPE (ITP, according to the authors; IPE1 in this section); and (iii) a third one using IPE with advanced interactive features (AITP, according to the authors; IPE2 in this section). In this third configuration, translators had all the aforementioned features of CASMACAT available, and they could choose which ones to turn on or off. Before proceeding with the post-editing tasks, the CASMACAT workbench was introduced to the translators participating in the study, and they had time to familiarise themselves with all the features before doing the trial evaluation. The trial consisted of three sets of three texts (thus, nine different texts), and each translator processed each text at least once under one of the three different conditions.

Dataset 1 was processed on the premises of the translation company participating in the study, and translators’ eye-tracking activity was recorded. Datasets 2 and 3 were processed virtually, and each translator worked at their home (their usual working condition, as they were freelancers). CASMACAT logged the keyboard and mouse activity of the translators performing the post-editing tasks in these two latter datasets. After each session, translators had to fill in an online questionnaire. Sanchis-Trilles et al. assessed and evaluated the data collected by using three different parameters: translation productivity, keystrokes, and gaze data.

As for translation productivity, the study used three made-up measures: (i) Kdur, the total translation time per segment, without taking into account pauses of more than 5 seconds, normalised by the number of characters in the source segment; (ii) Fdur, the total translation

time per segment, without taking into account the pauses of more than 200 seconds, normalised by the number of characters in the source segment; and (ii) Tdur, the total duration of the translation time. Results showed that TPE had the shortest processing time for Kdur and Fdur (i.e., offered higher productivity). Under the IPE configuration, translators were 5% slower in relation to TPE when considering the Fdur value. Hypothetically, authors suggested that this happened because translators were more used to TPE than to IPE. Authors therefore expected the processing time values to decrease (or translation productivity to increase) when translators got used to the system, and, as the first dataset was acquired in the premises of the LSP participating in the study and Dataset 2 and 3 were acquired virtually, Sanchis-Trilles et al. expected translators to reduce their processing time at home for IPE1 and IPE2 because translators already knew what CASMACAT offered after post-editing Dataset 1. It is true that, when working from home, Kdur and Fdur values for IPE1 and IPE2 dropped most, but TPE still had the lowest processing time values for Kdur and Fdur. Final results on average processing time per segment in terms of Kdur were 21.7 s (TPE), 27s (IPE1), and 29.6 s (IPE2). However, no statistically significant difference is mentioned or checked for these results in the study.

Regarding keystrokes, the study considered the number of actions of the users, understood as insertions and deletions. Per segment, on average, IPE1 (123.6) required fewer operations than TPE (131.3) and IPE2 (132.6).

On gaze data, eye fixations were used to record where the attention of the user was. Generally, translators focused more on the target window rather than on the source window, as would be expected; both IPE1 and IPE2 recorded more attention on the target window, as interactivity and constant changes in the target section may be an aspect that draws user attention. Again, authors do not state whether results on typing activity were statistically significant.

Translation quality was also studied in this trial, but only for Dataset 1. The post-edited texts were reviewed by professional reviewers, and quality was measured having “edit distance” in mind, meaning the number of operations to transform the original text into the reviewed one. This edit distance was calculated on a per-word basis, that is, the number of words needed to be amended in a post-edited sentence to become the reviewed sentence. As the post-edited texts had different lengths, all user operations (insertions, deletions, substitutions, and

corrections) were normalised to obtain true percentages. This way, all the systems could be compared, regardless of the size of the texts. The best edit distance score was obtained by IPE1 (9.4), while IPE2 (9.7) and TPE (10) got the second and third rankings, respectively. It is worth noting that these results only reflect the post-editing tasks of Dataset no. 1, and users therefore were still getting used to the different system configurations that CASMACAT offered. Once again, it must be stressed that these values were similar and that Sanchis-Trilles et al. did not mention any attempt to calculate whether these results presented a statistically significant difference.

User feedback was collected via ad-hoc questionnaires. Translators had to rate their user satisfaction after using the different configurations from 1 to 5. These results showed the perception users of CASMACAT had when doing the post-editing tasks. Globally, IPE was rated better than TPE, as IPE2 received a 4/5 and IPE1 a 3.89/5 score, while TPE only obtained a 3.78/5. Despite the good score given to IPE systems, seven out of nine translators stated in the questionnaire that they would have preferred to translate with non-computer-assisted translation tools. This can be linked to the negative perception that translators have of MTPE and technologies, in line with later studies by Guerberof (2013), Gaspari et al. (2014), and Moorkens and Way (2016).

To sum up, in this second trial study of the CASMACAT workbench, Sanchis-Trilles and colleagues demonstrated that IPE reduced the number of keystrokes required in comparison with TPE. Yet, translation productivity was slightly lower. The findings of this study suggested that productivity may increase when users familiarise themselves with the IPE system, so an IPE approach may require longer learning time than TPE. This result is supported by the fact that translators had experience in TPE but had never interacted with IPE before. To reduce the influence of familiarity vs. novelty in these post-editing modalities, a longitudinal study to analyse long-term translator performance and interaction with the IPE workflow would be required.

3.2.3.3. CASMACAT: Third Progress Report – The Longitudinal Study

In 2014, the third year and last progress report for the CASMACAT project was presented (Alabau et al. 2016). In this third and last report of CASMACAT, Alabau and colleagues performed two different studies: a longitudinal study and a third field trial investigation. This section describes the longitudinal study.

The results of the second field trial (the previous section in this literature review) suggested that translators needed more time to get used to IPE. This longitudinal study had the aim of investigating whether translators improved their performance when using IPE over time, and to what extent. Five professional translators participated in this longitudinal study, who worked alternatively with TPE (as a baseline configuration) and IPE for six weeks. Texts translated were news pieces, and there were 24 source texts of 1,000 words each, which had to be translated from English into Spanish. The main research question of this longitudinal study was: "Do translators become faster when familiarising themselves with using IPE?". Translators had to post-edit four texts per week. Authors counterbalanced each of the MTPE conditions to avoid text and tool-order effects. The first and last week of the study, translators worked on the premises of the translation company participating in the study, and they used eye-tracking systems. Working in the lab also helped to establish clear guidelines at the beginning. Weeks 2 to 5 were carried out at home, in the usual working environment of the participating translators.

This longitudinal study analysed post-editing behaviour by considering three parameters: Kdur, Fdur —which were the same as in the previous field trial— and Pdur. Pdur was the total translation time per segment, without considering the pauses of more than 1 second, normalised by the number of characters in the source segment. Results showed that post-editors used more keystrokes in the IPE configuration rather than in the TPE one. This result was to be expected because an IPE system the predicts the text and offers translation proposals while the translator types. Consequently, translators did more manual insertions in IPE than in TPE, where MT output is already inserted in the target segment. By contrast, in TPE, translators did more manual deletions.

The second element analysed in this longitudinal study was the most important one for the authors: the learning effects. After the six weeks of the longitudinal study, translators became substantially quicker when using the IPE approach, while there was no significant change in the TPE condition. These results may be attributed to the fact that translators were already used to TPE and could not actually "go faster" because they had reached their human limits. Taking into consideration the hypothetical assumption that there was a linear relationship between the time spent using CASMACAT and Kdur, Alabau and colleagues measured hypothetical regression lines based on simple linear models (see Figure 3.6). Following the

learning curve in an IPE approach, and the stable and linear productivity of the TPE approach, the authors of the study suggested that translators would become more productive in IPE than in TPE between weeks 9 and 10.

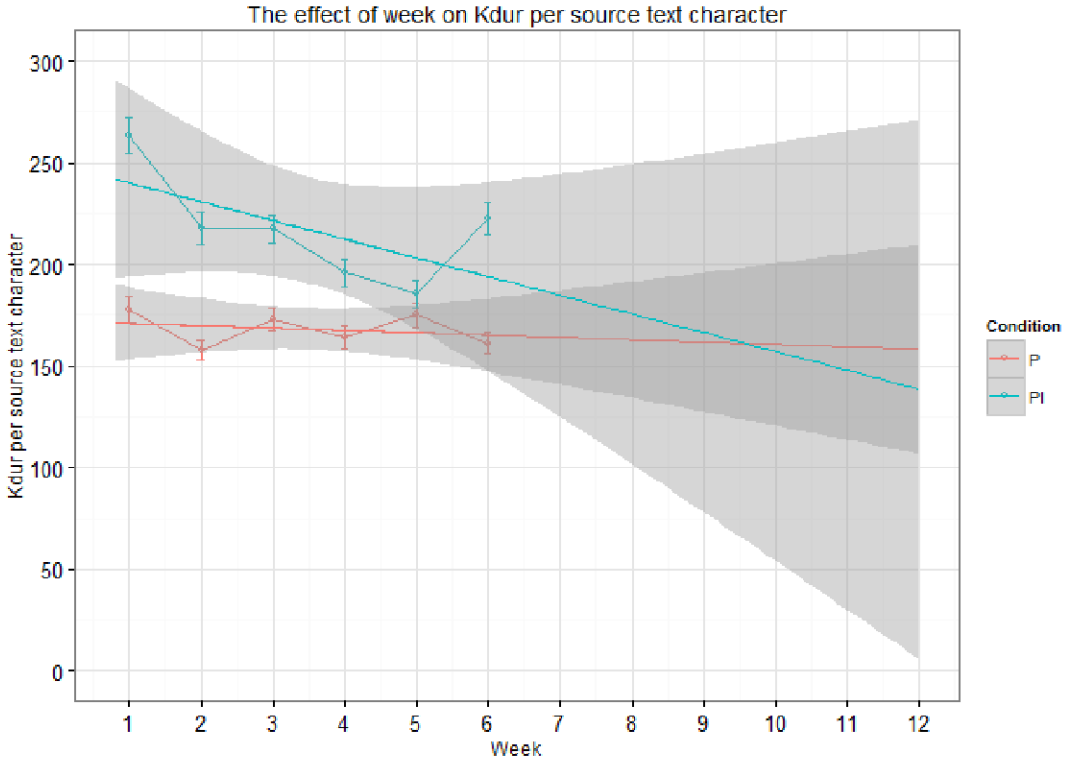

Figure 3.6. The image shows the effect of week on Kdur per source text character. TPE is the red line (P by the authors) and IPE is the blue line (PI by the authors). Grey areas are the hypothetical regression lines, which show that translators would become more productive under the IPE approach between weeks 9 and 10. Image retrieved from Alabau et al. (2016).

Having a more in-depth look at Figure 3.6, an increase in Kdur time can be seen for IPE in week 6, where translators had to post-edit in the premises of the translation company, while being eye-tracked. Authors suggested this increase is due to the lab effect and the difficulty of some texts processed in that week. Text difficulty was calculated using Translation Edit Rate (TER) (Snover et al. 2006), a metric for automatic evaluation of translation quality that consists of measuring the number of changes needed to the post-edited translation so that it exactly matches a reference translation. Having these two aspects in mind, Alabau and colleagues presented a new hypothetical projection, only taking into consideration the Kdur of those weeks when translators worked from home (their regular working method, as they

were freelancers for the translation company participating in the study). In this theoretical situation, according to the authors, translators would have been more productive in an IPE setting rather than in a TPE setting by week 6.

As the final part of the longitudinal study, an ad-hoc questionnaire was also used to collect feedback from the users at the end of the six weeks. Four out of five participants stated they preferred TPE over IPE. In addition, the authors highlighted the following comments on user perceptions, which deserve due attention: "post-editing with interactivity demands a controlled typing speed and this is difficult to achieve when you are an experienced touch typist" from one translator, and "I have to retrain myself on typing for IPE purposes" from another one (Alabau et al. 2013, 19). Three out of five participants confirmed that they would be willing to use CASMACAT in their future post-editing tasks. Also, the study authors suggested that IPE acceptance varied a lot in relation to the experience of the translators: the least experienced one (1-year professional experience) was positive about IPE features, while the most experienced one (27-year professional experience) had negative views of the system.

3.2.3.4. CASMACAT: Third Progress Report – The Third Field Trial

After having concluded the longitudinal study, Alabau and colleagues added an additional feature to the CASMACAT interactive approach: adaptive MT. This feature allowed the SMT that powered the CASMACAT workbench to personalise its translation completion proposals, which was one of the problems that users commented on in previous IPE user studies like that of CAITRA (Koehn, 2009). In the CASMACAT version with adaptive MT, when the translator validated a segment, the MT system introduced the validated segment into the translation model, and thus the subsequent translation completion proposals would take this validated segment into consideration. The main goal of this third field trial was to evaluate whether translators would benefit from a web-based, adaptive IPE approach in comparison with TPE.

Seven professional translators participated in this field trial, and four of them had already participated in the longitudinal study. This time, the texts post-edited were specialized in the medical domain, and there were two source texts containing around 4,500 words each. It must be noted that it is not known whether the translators participating in the study were

experienced or specialized in the medical domain. Regarding the evaluation methodology, as in previous studies, there were two post-editing modalities used: TPE and adaptive IPE. The participants had to perform the post-editing tasks in the premises of the translation company and were eye-tracked. Translators had to post-edit each of the texts in a single session. Yet, the authors did not state whether translators had to translate both texts (9,000 words) in one day or in more days (one session per day). It is worth stressing that translating 9,000 words in a day may be difficult, even with computer-assisted tools, so fatigue may be an important element to consider in such an experiment. Going back to the evaluation, translators were given time to familiarise themselves with CASMACAT, as three of them were using the workbench for the first time. Finally, the post-edited texts were later proofread and edited by different reviewers. Alabau et al. used the Fdur, Kdur and Pdur measures again.

Regarding productivity, there were two elements analysed: keystroke activity (the number of manual deletions and manual insertions) and processing time (with Fdur and Kdur). In terms of keystroke activity, on average, translators saw their keystrokes decrease in IPE if compared against TPE, both in deletions (from 70.71 to 36.94 keystrokes) and in insertions (from 79.53 to 68.73 keystrokes). The authors of the study claimed that this important decrease was statistically significant, and that the main cause was because of the IPE approach with an adaptive SMT engine.

In terms of translation time, there was no statistical significance. After the authors replayed the screen recordings of the post-editing tasks, they found that translators spent more time looking for terminology on the Internet in IPE than in the TPE. Therefore, they only preserved the time spent in the CASMACAT interface and computed the results again. After this amendment, there was a statistically significant decrease in time spent to complete the post-editing task both in Fdur and Kdur values for IPE. It is worth bearing in mind that terminology look-up is an important part of the translation process and, if IPE caused more of this, this extra time must be factored in, and, therefore, the results may have not been statistically significantly different after all.

After having described and analysed all the studies done with CASMACAT, the following conclusions can be drawn. Translators were estimated to need, on average, six weeks to get acquainted with the novel features of the IPE system to be more productive with such an approach if compared against TPE productivity. It must also be noted that translators' typing

behaviour is crucial, as it probably must change in regard to a TPE workflow, where speed is the most valued factor. In an IPE setting, overtyping may be negative because users will not benefit from all the possible advantages of the IPE features. It could be interesting to research whether translators with slower keyboard activity become familiar with IPE faster and can overtake their TPE productivity before the sixth week, as well as if they accept the system features more eagerly. If this was confirmed, probably special training in IPE should be provided. Nevertheless, a comment regarding a methodological issue of the previous studies should also be stressed: in this last report, the third field trial, translators only used adaptive learning MT capabilities in the IPE workflow, which may not benefit TPE. What would happen if we tested an adaptive MT system in TPE and IPE? For the comparison to be fair in the current context of this PhD thesis, translators should perform a TPE task with a state-of-the-art, adaptive NMT engine, as well as an IPE task with a state-of-the-art, adaptive NMT engine.

3.2.4. Lilt

In 2012, the Computer Science Department at Stanford University developed a new IMT system called Predictive Translation Memory (PTM) (Green, Wang, et al. 2014; Green, Chuang, et al. 2014). The developers of this system suggested that previous IMT systems were violating some of the principles for mixed-initiative user interfaces that Horvitz (1999, 160) proposed, no. 8 of which is: “minimizing the cost of poor guesses about action and timing”. As a consequence, PTM was designed to be a fast, responsive system, operated mainly via the keyboard, where users could do everything the workbench offered by just typing or pressing a combination of hotkeys. These aspects were crucial if taking into consideration the comments that translators had made on previous interactive user studies and the working situation of freelance translators — they are usually paid by the number of source words translated, and thus they have to work fast, and are usually touch typists (Carl 2012).

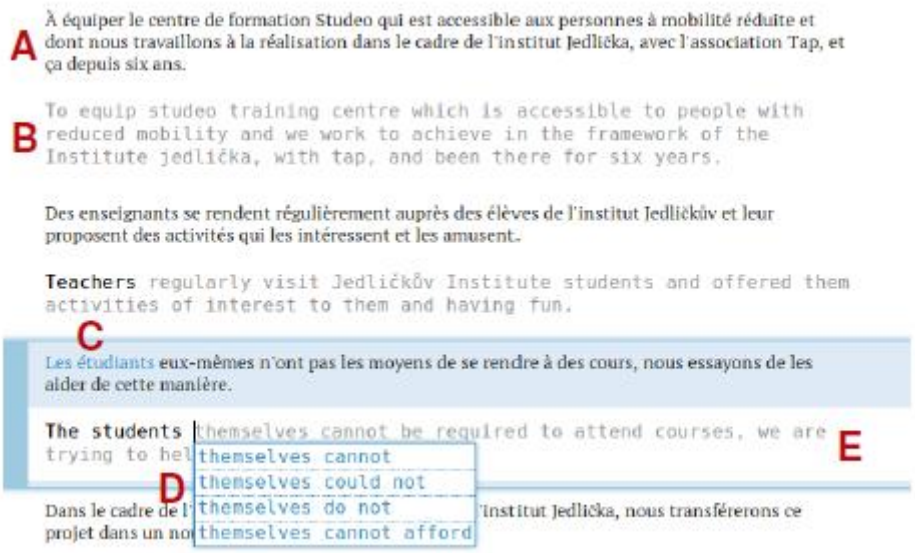

Figure 3.7. Interface screenshot of PTM. Image retrieved from Green, Chuang, et al. (2014).

As novel features for an IMT workbench, PTM recognized which words of the source text the translator had already typed and highlighted them to facilitate the comprehension of the user. According to the authors, all the features of PTM were designed taking into account Horvitz' principles for mixed-initiative user interfaces ("interfaces that enable users and intelligent agents to collaborate efficiently (Horvitz, 1999, 7)), mainly with the intention of developing significant value-added automation (a computer-assisted translation feature to reduce the typing effort of the translators), inferring ideal action in light of costs, benefits, and uncertainties (time saving and increased productivity), and maintaining working memory of recent interactions (adaptive, learning capabilities that personalise in accordance with the actions of the user). PTM used a different segment distribution (grouping source and target text vertically) instead of the usual left-right segment distribution in most CAT tools. In Figure 3.7, we can observe the source text in French, and the target text in English. In the source text box, the system highlighted in blue the already validated translation by the translator. In the target text box, we can see the already validated text in black, the MT text in grey, and the translation completion proposals in a dropdown menu, which could be accepted directly by pressing a hotkey.

Green, Wang, et al. (2014) also performed a user study of PTM's prototype. This study evaluated three different domains (software, medical texts, and news domain) in two language combinations: French-to-English and English-to-German. It should be noted that, in 2014, phrase-based SMT engines had better performance for the FR-EN language pair.

Participants were professional translators hired through ProZ, and all of them had previous professional experience post-editing MT output with other commercial CAT workbenches. This user study evaluated quality through BLEU+1 (C.-Y. Lin and Och 2004), a sentence-level variation of the automatic evaluation metric BLEU (Papineni et al. 2002), while reckoning that human evaluation was the gold standard, and leaving this latter type of evaluation for future work. Automatic results will not be mentioned because of the disadvantages they pose (more information on this in Section 2.4.1.1). Thus, only the human evaluation results are reported in this section, which are the best practices for translation quality evaluation (see Section 2.4.1.2).

In terms of translation productivity, the measures analysed considered the processing time per sentence. In the French to English combination, mean time for TPE was at 46 sec/sentence, while for IPE was at 63.3 sec/sentence. IPE was 18% slower. In the English to German combination, mean TPE time was at 51.8 sec/sentence, while IPE time was 63.3 sec/sentence. In this language combination, translators were 22.1% slower with IPE. After analysing these results more in-depth, Green et al. affirmed that translators saw their translation productivity increase with IPE over the course of the post-editing session. This last comment was in line with previous IPE user studies that suggested that the novelty factor of IPE may have a negative effect on translators' productivity on studies including only one interaction.

After the post-editing tasks, a questionnaire was distributed. In the questionnaire, translators suggested that they thought they could be faster using IPE than using TPE after enough training, but that IPE was more intensive and demanding because they had to read multiple translation proposals that were changing over time. As a final parameter evaluated in this study, instead of focusing only on quality and time, the study also focused on the HCI that took place when using IPE. The goal was to try to figure out what words came from the translation completion proposals or from the typing of the translators. In the French to English combination, 71% of the words came from translation completion proposals, and 18% from typing. In comparison, in the English to German combination, 65% of the words came from the completion proposals, and 34% from translators typing. This also corroborated the view that MT proposals were better for the first language pair, as translators used more translation completion proposals than in the second language combination.

Some years later, in 2016, one of the developers of PTM founded a Lilt – a language service provider that works with a specific CAT workbench powered by IMT and offers IPE services.

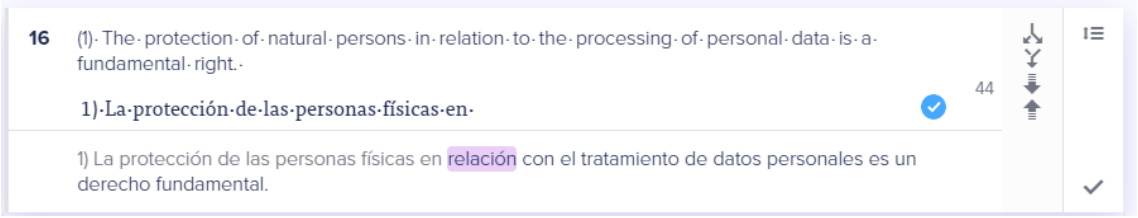

Figure 3.8. Screenshot of Lilt's graphic user interface

As Lilt was not only a research system, but a commercial one, the workbench presented a more visual graphic user interface (see Figure 3.8). Nevertheless, the underlying technology was the same as that of PTM. Lilt's workbench worked with two main hotkeys. On the one hand, Enter or Tab, which inserted the following word proposed by the MT system, which in the figure above would be the highlighted "relación". On the other hand, Shift+Enter or Shift+Tab, which accepted the whole translation completion proposal that the MT system offered. Since Lilt uses proprietary software, features like keystroke logging and time tracking are not available for normal users like in the previously mentioned open-source workbenches, and thus it is more difficult to evaluate this state-of-the-art interactive, adaptive workbench.

Nevertheless, in 2020, an in-house study of Lilt's user behaviour was presented at the iMpacT 2020 Workshop of the Association for Machine Translation in the Americas (Kovacs 2020). It is worth stressing that this is the first NMT study reviewed in this chapter, as previous workbenches still did not use the NMT architecture. The main goal of this piece of research was to know whether translators were using Lilt's MT translation completion proposals, and, if so, to what extent. Data was collected from Lilt's freelance translators from August to September 2020. The amount of post-edited segments or words is not known. Kovacs stated that, after evaluating MT translation completion proposals offered by their system, 46% of the proposals were correct. Nevertheless, translators still typed 58% of the text, and proposals were only accepted at a word-level 21% of the time (by pressing Enter or Tab to accept one-word completion) and at a sentence level 17% of the time (by pressing Shift+Enter or Shift+Tab to accept the whole translation completion proposal). One of the questions that Kovacs posed was whether translators were using Lilt as an interactive system for IPE or were

accepting the whole translation proposal to then perform a TPE task. By taking a closer look at this aspect, translators used the interactive proposals four times more in an IPE workflow than using the whole translation proposal for performing a TPE task. Yet, 17% of Lilt's users were still used to the TPE workflow and accepted the whole translation completion proposal to perform a TPE task.

3.2.5. Additional user evaluations of IMT systems for IPE tasks

Previous sections of this chapter contain IPE studies centered on IMT workbenches in this new field. These IMT workbenches received funding from universities and/or international bodies and were developed and assessed by their own developers. In addition to the studies from the system developers, further research on IMT and IPE has been undertaken. These research studies are presented in the following section.

3.2.5.1. Alves et al.'s (2016) study on CASMACAT

Most studies comparing TPE and IPE aimed to see whether there was any translation productivity difference when using these post-editing modalities. Nevertheless, Alves et al. (2016) changed their focus and studied cognitive post-editing effort specifically, that is, the amount of effort expended in a post-editing task (O'Brien 2006). Alves and colleagues were interested in the cognitive processes taking place in the translators' mind while post-editing (Krings 2001), and therefore conducted a different study that is also worth attention in this literature review.

The authors conducted a user study with 16 professional translators with at least five years of experience. Translators had to translate two different specialized clinical texts from English to Brazilian Portuguese (one with 17 segments and the other with 19) in TPE and IPE. However, no length nor complexity text control was carried out. In previous similar studies, it had been suggested that this may cause problems when gathering data or comparing different texts or post-editing modalities. The workbench used for this study was CASMACAT, together with an eye-tracker. In terms of methodology, translators first filled out a questionnaire with their professional experience and post-editing knowledge. Then, they post-edited the two clinical texts in the TPE and IPE modalities.

Alves et al. (2016) focused on gaze activity, specifically the number of eye fixations and its duration and average, as well as on type and number of edits in each of the post-editing modalities. The type of edits analysed in this study were entirely different to edits studied in the previous IPE studies. Following Sperber and Wilson's (1986) Relevance Theory and their conceptual/procedural information distinction (Wilson and Sperber 1993), instead of using the keystroke logger of CASMACAT, Alves et al. manually annotated three different types of edits: (i) procedural information, those edits relating to information that could be encoded in non-lexical categories (negation, tenses, determiners, etc.); (ii) conceptual information, edits relating to information that could be encoded in lexical categories (noun, verb, adjective); and (iii) hybrid encoding, edits relating to lexical items that included both of the above type of edits.

After the translators did the post-editing tasks, Alves et al. studied all the eye-tracking recordings and manually annotated the type and number of edits. In terms of results, there was no statistically significant difference in the total post-editing time between TPE and IPE. Thus, the authors suggested that, even with little training in IPE, this new form of post-editing could be a viable and alternative solution to TPE in terms of productivity.

Based on the different types of edits and the different tasks in TPE and IPE, Alves et al. suggested that both tasks involved different types of cognitive processes. In TPE, translators must read static MT output, detect the issues, and then amend them. In IPE, translators start writing their translation while they can accept translation proposals offered by the system. Therefore, in the IPE task, translators had more eye fixations than in TPE, but these fixations were shorter on average, suggesting that the cognitive effort in IPE was lower than in TPE.

3.2.5.2. Daems and Macken's (2019) study on Lilt

In 2016, the emergence of NMT started to gain strength because some researchers reported that the new NMT systems offered better MT quality than SMT (Bentivogli, Bisazza, et al. 2016; Toral and Sánchez-Cartagena 2017). This resulted in most MT service providers moving from using SMT to NMT.

Daems and Macken (2019) were interested in Lilt's IPE workflow. They already carried out a user study using Lilt's SMT system in order to compare it with another MT system. Lilt

subsequently changed their MT system (from SMT to NMT), and Daems and Macken therefore replicated the SMT user study with the newly implemented NMT system.

This evaluation consisted of two rounds of experiments on English-to-Dutch medical texts. On the one hand, one post-editing evaluation using the SMT engine of Lilt. On the other hand, one post-editing evaluation with the NMT engine of Lilt. Their main research goal was to compare translation quality and productivity offered when working with an adaptive, interactive SMT system against working with an adaptive, interactive NMT system. In addition, the authors gathered information about the perceptions that translators had of IPE. Four freelance professional translators participated in each of the experiments (eight for both studies), their experience varied from two months to fifteen years, and all of them had limited to no experience in the medical domain. Two out of the eight translators had never used a CAT tool. It is worth stressing that the wide experience difference of the translators and their lack of experience with CAT tools were important factors that were not controlled and that may have acted as important confounding effects.

To prepare the post-editing tasks and fine-tune the NMT system for the medical domain, the English-Dutch EMEA corpus (Tiedemann 2009), a medical corpus from the European Medicines Agency, and a manually-created medical termbase were uploaded to Lilt. Here, fine-tuning the NMT system on Lilt and not fine-tuning the SMT system may also translate into a biased comparison.

Before proceeding with the post-editing tasks, translators first worked on a test text to get used to the interface and features that Lilt offered. Then, they translated short texts of 20 segments for the SMT and NMT experiments. Camtasia was used to record the screen during the post-editing process, and Inputlog recorded the keystrokes of the translators. Finally, the users had to respond to a survey regarding their experience on using the platform.

Instead of using the typical BLEU or MQM-DQF metrics to assess MT quality, the SMT and the NMT systems were evaluated using the fine-grained error taxonomy and annotation guidelines of Tezcan, Hoste, and Macken (2017). One of the authors manually annotated 60

segments using the BRAT rapid annotation tool,[4] using the previous materials as a reference and using segments that did not offer exact or fuzzy TM matches, so these segments were raw MT output. The SMT output had 78 errors, while the NMT output had only 55. Yet, it must be taken into account that in IMT settings, only the first MT output proposal can be evaluated, as the proposal changes while translators type during IPE.

Regarding translation times, the time translators spent post-editing their texts was comparable for SMT and NMT (including the time spent looking up in external resources). In order to measure the post-editing effort required in different post-editing workflows (IPE or TPE), the measures introduced by Barrachina et al. (2009) and later used by Ortiz-Martínez et al. (2011) and Peris and Casacuberta (2019) were implemented. These measures were as follows:

- KSR or keystroke ratio: the total number of keystrokes divided by the total number of characters in the reference translation.
- MAR or mouse-action ratio: the number of pointer movements divided by the total number of characters in the reference translation.
- KSMR or keystroke and mouse-action ratio: the sum of KSR and MAR.

These measures were simulated, that is, calculated hypothetically by a computer and not extracted directly from the participants' data. They were normally calculated using a human translation as a reference. The underlying problem of this method is that the results of this evaluation are fully automatic, and translators do not always produce a translation using the least number of edits possible to meet a reference translation. In addition, there are also multiple possible translations that could be used as a valid reference translation, so these numbers may not be a direct or fair indicator of post-editing effort. In their study, Daems and Macken use Barrachina et al.'s (2009) measures, but amend their application, as they use the post-edited segments of the translators (the already accepted segments) as the reference translation, treating this as a fairer metric and indicator of post-editing effort (Popović, Arcan, and Lommel 2016). It should be stressed that the computer only calculates the minimum

[4] http://brat.nlplab.org/, last accessed on the 8th of April 2021.

number of edits necessary, and as explained above, automatic metrics have different problems for assessing translation quality (see Section 2.4.1.1). In the study, differences in KSR, MAR, and KSMR do not seem to differ much in any of the scenarios, neither in the SMT nor in the NMT. In addition, there is no mention of statistically significant differences, as the texts post-edited were too short to obtain statistically significant results (only 20 segments). To obtain more data to analyse, Daems and Macken also present the results of two automatic post-editing effort evaluation metrics, (H)TER (Snover et al. 2016) and CharacTER (Ling et al. 2015), showing that the interactive NMT system required fewer edits than the interactive SMT system for both metrics. In terms of post-editing effort, they conclude as follows:

> We demonstrated that technical effort calculated on the final product as expressed by TER and CharacTER scores does not always reflect actual effort as expressed by the KSMR scores. This corresponds to findings of Daems et al. (2017) that "product effort measures do not necessarily measure post-editing effort the way process effort measures do". When developing new MT systems or translation tools it would be good to take these differences into account. (Daems and Macken, 2019, 14-15)

Finally, in terms of perceived usability, Daems and Macken carried out a survey to assess whether translators preferred either the SMT proposal, the NMT proposal or neither of them. To accomplish this, the survey showed one source sentence together with four options: the SMT proposal, the NMT proposal, a "neither" option, or a "no preference" option. The MT proposals were randomly presented for each sentence, and no information on the type of MT system creating them was given. The results of the survey showed that the NMT proposal was the option chosen most often (18/40 times), together with "neither" (14/40). The SMT proposal option was considerably less preferred.

3.2.5.3. Sánchez-Torrón's PhD Dissertation on CASMACAT

In her PhD dissertation, Sánchez-Torrón (2017) performed two empirical studies with English to Spanish translators to research different productivity aspects in TPE and IPE. For the sake of this Chapter on IPE, only the IPE study will be described and analysed. As this PhD thesis also occurred after the major implementation of NMT engines, Sánchez-Torrón did the first study comparing TPE and IPE with a NMT engine. In this study, the term used to refer to the

technology behind the text prediction and translation completion proposal system is once again "ITP", and the workbench used was CASMACAT.

Eight English to Spanish translators participated in the IPE study (seven had already participated in a previous TPE study by the same researcher), and they were hired through ProZ,[5] one of the biggest translation job portals. The requirements of the job advertisement were that translators needed to have at least two years' translation experience and to be familiar with CAT tools in general. To avoid any type of bias in the participant selection, they were hired on a first-come-first-served basis.

The main objectives of this study were to discover whether IPE was an efficient alternative to TPE in productivity terms, and to investigate whether translation productivity increased as translators familiarised themselves with the IPE system. To achieve these objectives, similarly to previous research on IPE, a longitudinal user study was carried out.

The methodology that Sánchez-Torrón followed in her PhD thesis was as follows. First, a pre-task questionnaire was used with the aim of obtaining information about the translators relevant to the main task, namely, their translation, MT and PE experience, and the languages they worked with. This first questionnaire was followed up by a warm-up task, where translators could familiarise themselves with the CASMACAT workbench. During this warm-up task, translators had to post-edit four sentences following the TPE modality, and four additional sentences using the IPE modality. CASMACAT's logs of the actions of the translators were then checked to see whether they were recorded correctly. This was followed by the main task. The main task was divided into eight translation sessions (S01 to S08) within a period of around four weeks. In the first translation session (S01), translators used the CASMACAT workbench to perform a TPE task, which was used to obtain the baseline productivity data of each translator. From S02 to S08, translators had to perform IPE tasks to study the possible learning effects resulting from the continuous use of IPE. Translators post-edited eight different texts, one per session, which were controlled for length and syntactic complexity to avoid possible effects on translation productivity (Lin 1996; Green, Heer, and

[5] https://www.proz.com/, last accessed on the 12th of April 2021.

Manning 2013; Mishra, Bhattacharyya, and Carl 2013). The post-edited texts covered different topics, so that the post-editing of one text would not pose an advantage in post-editing any other text. Finally, after the main task, translators had to fill in a post-task questionnaire, where their perceptions of an IPE system after using it were gathered.

Sánchez-Torrón studied IPE taking into consideration three major aspects. In the first place, translation productivity or processing time. The amount of time that translators took during the post-editing tasks was recorded by the CASMACAT workbench at a segment level. Though CASMACAT did not measure the time spent outside its interface, Sánchez-Torrón measured this time and added it to the active time of the segment. The time spent looking for external sources or terminology should be considered, as it may be possible that one workflow required more terminology lookup than another. The processing time was calculated by dividing the time in seconds spent by the number of tokens of the segment. Tokens were used instead of words in this study to use the same productivity measure as that of the studies of Alabau et al. (2016) and Alves et al. (2016) to compare results.

The second aspect to be evaluated was translation quality. For this, a MQM scorecard with 106 issue types was used (Lommel and Melby 2014), and the author manually annotated the issues using the guidelines and decision trees provided in Burchardt and Lommel (2014). The Pass/Fail threshold in quality was established at 95%.

The third and final aspect evaluated was technical effort, which was measured via five operations that translators could perform while post-editing (all these data were extracted from the keystroke log that CASMACAT offered): (i) manual insertions: the total number of alphanumeric characters manually inserted by the translator divided by the number of source tokens; (ii) manual deletions: the total number of alphanumeric characters manually deleted by the translator divided by the number of source tokens; (iii) navigation and special key presses (key indicator, not taken into account in the previous IPE studies): all the operations not considered manual deletions or insertions, that is, the use of navigation and control keys, and the Tab key to accept the translation completion proposals; (iv) mouse clicks: all the clicks the translators did while post-editing; (v) tokens of MT origin: in order to know whether the MT proposals that the IMT system offered were useful and translators were accepting the proposals during the IPE tasks, the number of MT tokens accepted were divided by the number of source tokens. All these five measures were calculated at the segment level.

After every IPE task was performed and data were analysed, Sánchez-Torrón presented her results in the first user study of an IPE system using NMT. In terms of processing time, five out of seven IPE sessions presented mean values lower than the TPE session. On average, IPE reduced the processing time by 0.10 seconds per source token or a time decrease of 2%. In terms of technical effort, as IPE and TPE involve different processes, on average, four out of five indicators favoured IPE against TPE because their values were lower. The only indicator that had a higher value was Special Key Presses. Yet, this is a positive result for IPE because it meant that translators had used the translation completion proposals while doing the IPE tasks. Most of these results were statistically significant. Regarding translation quality on a segment level, MQM scores for both configurations were similar, and most of the segments had an acceptable level of quality (Pass). Yet, IPE obtained slightly higher scores in the Pass/Fail ratio, that is, more segments were considered a Fail in the IPE setting than in the TPE one. Sánchez-Torrón argued that this could be because the TPE task already populated the whole target text with MT output, while in IPE translators had to type the target text by themselves, and they therefore introduced typing mistakes that they did not amend. It should also be stressed that, in terms of sessions, almost all the TPE and IPE sessions were a pass, but one, which was the first IPE session of one of the translators (56 out of 57 sessions obtained the Pass score). When looking at the type of issues, the IPE workflow had many more fluency issues (224% more) than TPE but had also fewer adequacy issues (27% fewer) than TPE. Thus, temporal effort and technical effort seemed to favour IPE, while translation quality indicators did not differ much between conditions and were comparable. In terms of translators' perceptions, via the post-task questionnaire, five out of eight translators stated that they preferred IPE over TPE. Six of them also thought that they improved their translation speed with IPE as the study progressed. Also, clear advantage was seen in participants with experience in TPE, who benefited more from the IPE setting and features.

As a summary, the statistically significant results in Sánchez-Torrón provided insights into the potential of IPE to pose a viable alternative to TPE in terms of technical effort. When speaking of productivity, translators were marginally faster with IPE over TPE, also backing the idea of IPE being a feasible alternative to TPE. Regarding translators' perceptions, IPE was rated better than TPE, as translators stated that the interactive assistance was a favourable type of assistance that they would likely introduce in their workflows, a similar finding to the studies

by Koehn (2009) and Langlais, Foster, and Lapalme (2000). This is a contrast with the translator dissatisfaction or dehumanization that the TPE task implies, as Moorkens and O'Brien (2015) suggested. It is also worth highlighting that Sánchez-Torrón found no clear productivity increase in the IPE setting as the study progressed.

3.2.5.4. Torregrosa-Rivero's PhD Dissertation

The IMT systems analysed previously used a computational strategy called a "glass-box approach". In his PhD thesis, Torregrosa-Rivero (2018) proposed a computational strategy called "black-box approach" to IMT systems. As a parallel with the field of electronics, he stated:

> A black box is a closed, opaque system with well-defined inputs and outputs: what happens inside cannot be perceived, only the outputs can be observed. Conversely, a glass box is also a closed system with inputs and outputs, but it is transparent: the inner components and how they evolve as the inputs are being processed can be observed and studied along with the output value. (Torregrosa-Rivero, 2018, 47).

In glass-box IMT systems, the embedded MT engine is queried for translations and also for other features that the IMT systems offer (such as alternative translation proposals, confidence measures, etc). Glass-box IMT systems generate a new translation proposal for each keystroke (or prefix) inserted by the user. This means the translator must stop writing, read the new proposal, and then accept it if they deem the new proposal relevant. These aspects cause the computational load and requirements of glass-box IMT systems to be high.

In the IMT black-box approach proposed by Torregrosa-Rivero (2018), the systems pre-translate and obtain all the possible subsegments of a text when the document is loaded in the system, and therefore the translation proposals can be shown faster than in glass-box systems because black-box IMT systems do not have to recheck all the linguistic corpora each time the translator types a prefix. This would also help low-resource languages, making the black-box IMT system less dependent on huge parallel corpora. Yet, as a result, black-box IPE systems cannot offer whole-sentence translation proposals with good quality, but only small segments with limited context. This may be a drawback, as it has been seen in previous IPE studies that giving the translator too many options to read and choose from may be too

cognitively demanding, and longer translation proposals may be more productive for the post-editing task. In addition, if all the translation proposals are pre-translated when uploading the source file, there is the possibility that MT proposals may not adapt adequately to the text introduced by the translator because the system may have not considered every potential translation option. Thus, the glass-box IMT approach may be a double-edged sword, offering potential benefits, but also potential disadvantages.

For this PhD thesis, the open-source, web-based tool Forecat was developed (Torregrosa-Rivero, Forcada, and Pérez-Ortiz 2014) (Figure 3.9). In line with previous IMT systems, Forecat aimed to ease the coupling of a CAT tool for IPE tasks with an MT engine.

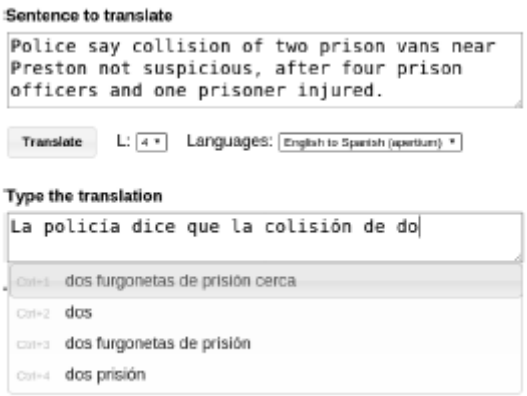

Figure 3.9. Forecat interface. In the upper part, the source text can be seen. In the lower part, the target text box can be observed, with dropdown menu showing up to four translation completion proposals. Image retrieved from Torregrosa, Forcada and Pérez-Ortiz (2014).

3.2.5.4.1. Evaluations with non-professional translators

Torregrosa-Rivero (2018) worked with two different types of black-box approaches: a heuristic approach and machine-learning based approach.

For the heuristic approach, Torregrosa-Rivero (2018) performed two different evaluations. In the first place, an automatic evaluation following the one done in Langlais et al. (2000), where the machine simulated a hypothetical user completing the translation task, and measured the keystroke ratio (KSR, the total number of keystrokes divided by the total number of characters of the reference translation). Two language combinations were evaluated, English to Spanish (offering up to 48% saving in keystrokes), and English to Czech (offering up to 31% saving in keystrokes). It should be stressed that this evaluation was fully automatic and reflected the

"hypothetical" scenario where a translator always accepted the best translation proposals during the post-editing of each sentence (as in TransType's early evaluations). In the second place, Torregrosa-Rivero also performed a user study of the heuristic, black-box IMT approach.

Eight non-professional translators (volunteer computer science students), who had no experience in translation, had to do some IPE tasks using the web interface of Forecat as the IMT workbench, connected with the RBMT engine Apertium (Forcada et al. 2011) in the Catalan-Spanish language pair. The user study was divided as follows. First, translators had two sentences to test Forecat and get used to the workbench. Then, the main task started. Ten sentences from the system training corpus were extracted (211 words) and were divided into two blocks. Five sentences were translated without computer assistance, and five with IPE assistance. As expected in this evaluation, all users translated faster with assistance than without. KSR values from the main task indicated that users could save from a minimum of 46% keystrokes to a maximum of 77% using the ITP features. Finally, participants completed a survey and rate their experience with a Likert scale. All users stated that the interface was easy to use (median answer of 5). They also indicated they would use the tool for future translations (median answer of 5) and thought that suggestions were useful and that the tool allowed them to translate faster than unassisted (median answer of 4.5). Yet, it must be noted that, as a big shortcoming, users were computer science students and not translators, and therefore had no professional experience in translation. In addition, Torregrosa-Rivero compares the productivity of IPE with the productivity of non-computer-assisted tasks. With this comparison, it is normal and easy to claim that there has been a productivity increase. In a real-world scenario, translators do not work without assistance, as they normally use translation memories or MT with different CAT tools. Then, like in previous IPE studies, a fair comparison would be to contrast TPE with IPE, which would really show whether it is feasible to introduce in a real professional workflow the black-box IMT approach proposed in Torregrosa-Rivero's PhD dissertation. It is also worth noting that translation quality was not assessed.

After the first evaluations of the heuristic, black-box IMT approach were carried out, Torregrosa-Rivero proposed a machine-learning approach, based on the newly adapted neural networks (Sutskever, Vinyals, and Le 2014). Technically speaking, he used a neural IMT

system trained with 15,000 English-Spanish sentences from the Europarl corpus, a collection of proceedings from the European Parliament. Then, he connected the neural IMT feature with the SMT system Moses, trained with more than 150,000 sentences of the Europarl corpus. This machine-learning based IMT system was also automatically evaluated, but this time in comparison with Thot (Ortiz-Martínez and Casacuberta 2014), a glass-box IMT system developed with the CASMACAT workbench. The automatically evaluated KSR values of the machine-learning based configuration of the black-box IMT system were lower than the KSR values of the heuristic approach, showing statistical significance. Yet, the limitations of the automatic evaluation should be noted once again. Then, Torregrosa-Rivero (2018) carried out a preliminary user study to test and compare all the previous systems. Eight computer science researchers participated in the study, who were Spanish native speakers with a limited working proficiency of English. The users of the study had no translation education or experience. It is worth stressing that this is once again a major sampling drawback. If we leave aside the computational and algorithmic aspect of the study, which is not discussed here, how applicable and reliable is the profile of the study participants for the language services industry? In other words, how applicable to the translation industry and reliable are the results of a group of computer science researchers who are not familiar with translation techniques, have limited proficiency of English, may have a limited knowledge of their L1, and have never used a CAT tool?

As per the IMT system for the IPE tasks, Forecat and Thot were integrated into the open-source CAT tool OmegaT[6] as plugins. The translation sessions were logged with the OmegaT session log plugin. Twenty English sentences with lengths between 15 and 25 words were post-edited into Spanish. These sentences were divided into four blocks of five sentences each, and had to be translated under different conditions: (i) induction, for the translators to familiarise themselves with the GUI and both suggestion models; (ii) unassisted task, with no suggestions offered; (iii) black-box IPE task, offering the system up to four suggestions ranked with a neural system; (iv) glass-box IPE task, re-running the system and offering new translation completion proposals every time a prefix was inserted by the user. The indicators

[6] https://omegat.org/, last accessed on the 15th of April 2021.

studied were the total translation time (measured in seconds), the size of the final accepted translation by the users (measured in characters), the total number of keystrokes, the KSR, and the translation speed (measured in characters per second). Regarding the results of this user study, in comparison with the unassisted task, on average, users saved 10% on keystrokes and were 4% faster with the black-box approach and saved 15% keystrokes and were 12% slower with the glass-box system. Black-box suggestions were less useful at saving keystrokes but allowed users to translate faster. After the IPE tasks, users were asked to sort the tasks according to their perceived speed of translation (black-box, glass-box or unassisted), and three preferred the black-box, while five deemed they were faster with the glass-box one. Perceptions did not correspond with the results, as most users preferred using glass-box rather than black-box systems, although the black-box approach offered better results. Once again, it should be stated that this evaluation was carried out with computer science students, who had no experience with translation, and therefore the results may not be applicable to the real-life translation industry. In addition, no comparison with TPE was made .

3.2.5.4.2. Evaluation with professional translators

Torregrosa-Rivero (2018) stated that all the previous evaluations were for testing his formulas and improving his black-box approach in a development stage. As has been observed in previous IMT system-related research, non-professional translators may incorrectly evaluate the translations because they lack translation knowledge. Due to the economic and resource limitations of a PhD researcher, Torregrosa-Rivero undertook two human evaluations with non-professional translators (i.e., computer science students/staff; with no earlier translation experience). Now, in the last human evaluation of his PhD dissertation, he hired eight professional translators, all of them Spanish natives, who were used to working with OmegaT. The system evaluated was the Forecat IMT system connected with OmegaT and Moses SMT. The text to be translated was extracted from the News Commentary Corpus and the United Nations Corpus. As the users translated the texts in one session and the whole text used the different MTPE modalities, the sentences were shuffled, not following any logical narrative, to try to avoid the learning effects of a continuous text on the MTPE modalities. All translation

suggestions (subsegments) were generated beforehand and were then kept in a cache, using the CacheTrans-OmegaT plugin, so the system could generate them instantaneously.

In terms of the evaluation methodology, all the sentences to be translated were split into six sentence blocks: three induction blocks (to get used to the different types of assistance) with 20 sentences each and three evaluation blocks with 100 sentences each. The translation task was divided as follows. First, a 5-minute induction session was carried out, where translators had no assistance whatsoever. Then, translators had 40 minutes for Task 1, which was translating without any type of computer-assisted translation aids. This was followed by a 15-minute break. After the break, a second induction session of five minutes was performed, now using a black-box IMT system for conducting an IPE task, offering up to four translation completion proposals. Then, in Task 2, translators had 40 minutes to conduct an IPE task with the black-box IMT assistance. This was followed by a 15-minute break for answering a short questionnaire. A new 5-minute induction session (the third one) was then carried out, using a black-box IMT system with a quality threshold, meaning that the system could offer up to four suggestions, showing only the suggestions that were ranked best by the neural IMT network. In the case where there were no high-scoring translation proposals, none would be proposed. Finally, the 40-minutes post-editing Task 3 followed, using the previous system configuration. The user study then finished with 15 minutes for answering a short questionnaire.

In terms of results, as logically expected, Task 2 (IPE with a normal IMT system) allowed the users to translate faster than Task 1 (non-assisted translation). Users also performed better in Task 2 than in Task 3 (IPE with an IMT containing a quality threshold). According to a hypothetical calculation of the keystrokes that a user could have saved, the author stated that this figure was between 25% and 65% over their obtained results. Yet, as in previous IPE studies, it has been shown that this type of automatic calculations gives much better results than those extracted with real post-editing process data. When user studies are undertaken, users do not tend to always accept the best possible solution, and their real processing time tends to be longer. In addition, user perception of the systems of Task 2 and Task 3 were quite negative, as the median result in the Likert scale (from 1 to 5) was at 2.37 and 2.25 respectively, none of them achieving a Pass (2.5). It is worth stressing that the overall results of this study were negative in terms of user perceptions of the system, and that no real

comparison had been done with TPE. The results obtained from Torregrosa-Rivero's (2018) PhD dissertation were easy to anticipate, that is, translators perform more keystrokes and translate slower unassisted than when having IMT assistance. This is no breakthrough. Thus, to really know whether the proposed black-box IMT approach was useful and a viable solution for the professional industry, it should have been compared with TPE, which is the norm and the current real-world workflow. It should also be noted that the user study methodology and task distribution was interesting, and that is the reason why these studies have been included in this literature review on IPE.

3.3. Discussion on interactive post-editing (IPE)

The previous sections provide an extensive review of studies to date based on IMT systems for IPE tasks. In this section, I summarize the main findings from this review and highlight the interesting questions that emerge from all these previous studies. One of the main findings is the relevance of selecting the appropriate variables to study, more specifically, users and texts.

As far as users are concerned, it has been shown that non-professional translators may lack translation knowledge and may accept mistakes (Castilho, Moorkens, Gaspari, Calixto, et al. 2017). Thus, the requirement for participants with translation knowledge is obvious if the research seeks to make valid conclusions on that cohort's interaction with such systems. Yet, this translation knowledge may have considerable influence on the translation task and may vary significantly between users. For example, more experienced translators may be more reluctant to adopt (new) technology aids, as they are acquainted with more TPE workflows, as seen in the CASMACAT workbench (Alabau et al. 2013; Sanchis-Trilles et al. 2014). Typing speed is also a relevant aspect that may influence the experiment. In IPE, if the system does not offer the translation completion proposals rapidly, touch typists may be hampered by the system, having to reduce their typing speed to wait for the system proposal, read it and then decide whether to accept the proposal or continue to type. Also, in the past, computing processing power was inferior to that of today, and current IMT systems may offer faster text prediction and translation completion proposals, allowing the IPE process to be faster than ever before.

It must also be stressed that many professional translators have introduced TPE tasks into their daily workflow for some months or years already, and this may be a disadvantage for comparing TPE (an expert or semi-expert task) with IPE (a novel approach to the task). We cannot assume that, in the studies outlined above, translators had started "cold" for the TPE tasks. In the same way, we cannot compare the productivity of translators that had been translating for 10 years already with TPE, if they have only had one hour of IPE training. This suggests that a longitudinal study should be carried out to take better account of such imbalances in experience and to investigate the possible learning curve of translators when doing IPE tasks. According to Diggle et al. (2002), longitudinal studies would allow us to observe individuals on multiple occasions enabling the direct study of change.

Another interesting finding of this review is that it is of key importance to control the text length and difficulty. Otherwise, results may be problematic, and trends may not be observed properly because processing speed for longer and easier texts was faster than for shorter and difficult texts, like in TransType's last evaluation. Similarly, we can also observe an effect if the translator sits for too long doing a certain task, as the cognitive demands increase and their productivity decreases. Therefore, to avoid this fatigue effect, it is interesting to perform the post-editing task of each text in a single, not-too-long sitting. In addition, to avoid participant-dependent speed in PE and the variability of results between translators, as in Cettolo et al. (2013), Koehn and Germann (2014), and Sánchez-Torrón (2017), all translators should post-edit all the source sentences, no translator should post-edit the same sentence twice, and all translators should be exposed to the same amount of output from all the different workflows (when comparing different systems or configurations). These latter aspects need to be controlled correctly and considered when designing the experiment.

Some studies have also suggested that, though users' emotions and enjoyment of the post-editing task is not very high, their productivity increased more than with other assistance types (Koehn 2009a). It may be worth discussing whether emotions have a special effect in human-computer interaction, on the productivity data, and/or the flow of the translation process, whether in TPE or IPE.

Finally, as most of the previous IMT studies were carried out before the paradigm shift from SMT to NMT, most studies did not include NMT. In the moment of writing this PhD dissertation, the fairest comparison would be to analyse the IPE task of a text with an adaptive

NMT system against the TPE task of an adaptive NMT system. With the current situation and the excellent quality that NMT systems offer now, it is possible that TPE would be more productive than IPE in language combinations with MT output of high-quality because not many changes have to be implemented. In the case of lower quality MT language combinations, IPE may be a more viable option. Yet, this must be evaluated and investigated. In terms of user satisfaction, TPE is not seen positively by translators because of the loss of user agency and the dehumanization of the translator (the MT engine proposes and the translator adapts to the MT) (Cadwell et al. 2016; O'Brien et al. 2017; Fırat 2021), but IPE may be a much more rewarding task (the translator proposes and the MT engines adapts to the translator); IPE has the potential to be more human-centered post-editing modality, respecting, empowering and augmenting translators during the translation process.

CHAPTER 4. TRANSLATION AS A FORM OF HUMAN-COMPUTER INTERACTION

This chapter provides an overview of human-computer interaction (HCI), an established field of research within computer science, and discusses its relationship with translation. Section 4.1 provides a summary of relevant aspects of HCI and is followed by the introduction of the two most relevant concepts within the field, namely usability (Section 4.2) and user experience (Section 4.3). Then, a review of the studies of HCI factors in Translation Studies can be found in Section 4.4. Finally, Section 4.5 presents and discusses the concept of human-centered, augmented machine translation (HCAMT).

4.1. Human-computer interaction (HCI)

Over the last decades, computers and information and communication technologies have developed substantially. Due to the rapid and exponential progression of the computing field, with improvements in algorithms and processing hardware, the main organizations and/or societies of the computing world (i.e., the Association for Computing Machinery [ACM] or the IEEE Computer Society) have been changing their focus of attention, hand in hand with technological developments (Association for Computing Machinery 2020).

Yet, it is worth stressing that computational aspects are not the only perspective from which computers or technological devices can be studied or analysed. One of these different perspectives has emerged as a key subfield within computing: human-computer interaction (HCI) (Dix 2003). Though there is no agreed definition of HCI, Hewett et al. (1992, 5) offered the following general and concise definition:

> Human-computer interaction is a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them.

The main goal of HCI is to focus on humans because they are the ones interacting with technological systems to achieve certain goals. Since there are many different systems or devices, the goals of users will also differ. Hewett et al. (1992) also suggested that HCI is a large interdisciplinary area, which could involve research fields like psychology (to study

users' behaviour or cognitive processes when interacting with computers or devices) or sociology and anthropology (to assess the relationship between technology and work), among others. There have also been different studies attempting to pursue the terms "man-machine symbiosis" (Licklider 1960) or the "augmentation of human intellect" (Engelbart 1962). Today, it is widely accepted that HCI is a key field in today's computing world (Dix 2010), and multiple conferences studying the interaction of humans with computers or intelligent devices take place every year and draw the attention of many researchers all over the world.

The interdisciplinarity of the HCI field and not having a clear definition of the term "HCI" has caused multiple discussions in the HCI world. For instance, Liu et al. (2014) analysed all the keywords of the 3152 publications available in the Computer-Human Interaction (CHI) Conference, one of the most renowned conferences on HCI and computer science. In their study, Liu et al. (2014) crawled all the keywords of these publications and divided them into two 10-year periods to analyse them in-depth to try to find a thematic core in HCI. The first period included the keywords of papers published between 1994 and 2003, and the second period included the keywords of the publications presented in CHI from 2004 to 2014. Liu and colleagues (2014) observed that the HCI field changed tremendously in the 20 years evaluated, influenced by the rapidly evolving technology and new technological devices. According to the authors, almost half of the core and backbone keywords of the first 10-year period (e.g., "world wide web") disappeared in the second 10-year period, and new keywords and topics of interest appeared (e.g., "mobile phone"). In addition, they suggested that, since the HCI field had no central topic, significant technological changes and the appearance of new technologies (e.g., studies on the design of a mouse were relevant for ergonomic mouse designers, but not to touchscreen designers) made the accumulation of knowledge or even the sustainability of the discipline more difficult. They therefore proposed to refer to HCI as a "field" rather than a "discipline".

Blackwell (2015) responded to the paper by Liu et al. (2014) suggesting that HCI researchers should not look for a thematic core to find the identity of HCI. Instead, Blackwell proposed that HCI researchers should deem HCI as a way of contributing and responding to other disciplines, defining HCI is an "inter-discipline". To defend this position, Blackwell (2015) suggested that HCI should focus on incorporating how knowledge was created in interdisciplinary encounters by considering science and technology studies, as well as the

sociology of knowledge. In addition, Blackwell (2015) commented on Galison's (1997) book, where the latter proposed that the HCI community worked as a "trading zone" where productive exchanges took place between engineers and technology designers, and researchers had to provide information on user experiences and behaviour. Following this "trading zone" point of view, Fincher and Petre (2004) argued that computer science research was also a trading zone between two different stakeholders: on the one hand, the mathematical and technical part of computing; on the other hand, the sector focusing on how this must be applied in education and industry. This discussion applies to the domain addressed in this PhD dissertation because HCI is a key component for understanding today's MT field. In the MT community, we have translators on the one hand, who are the people in charge of producing translations, and MT developers on the other hand, who are the people responsible for creating the different translation technologies for translators (e.g. CAT tools or different MT systems). We could even include a third party here: lay users of MT, who may not be translators, but only people using MT for gisting or assimilation purposes (Nurminen 2019). Thus, we can also use the "trading zone" analogy here, as technology and MT developers will develop tools and workbenches with new features, and translators or lay MT users will be the ones using them, sometimes having to discuss with language service providers and clients the use of specific tools. Yet, translators sometimes reject specific technologies or are reluctant to introduce them into their daily workflow, normally in addition to many other tools they have been asked to use. Do these new tools improve or worsen translators' conditions and workflows? These are aspects that have been considered for some time but ought to be discussed from this translator-computer interaction "trading zone" perspective. Going back to Blackwell (2015, 3), he suggested:

> HCI is not solely an "interface" field necessary where computer science must engage with the outside world, but an essential independent model that can be drawn on to maintain the discipline of computer science itself.

Liu et al. (2014) suggested that HCI needed a thematic stabilisation to consolidate as a discipline. By contrast, Blackwell (2015) referred to the comparative history of science, as it has been suggested that disciplines emerged in communities that shared an interest within a professional context instead of being determined by clear bodies of scientific knowledge

(Lloyd 2009). In the MT field, the introduction of MT into the professional translation workflow has disrupted the previous scenarios of the language services industry (see Chapter 2), and therefore new issues or negotiations between communities arose due to this innovation, mainly among language service providers and MT system developers on one side, and translators on the other.

To back his ideas and reflect the inter-disciplinary status of HCI, Blackwell (2015) compared the nature of the experiments that were being undertaken in the thematic review subcommittees of CHI[7] with the ones of the Crucible network, a research network of the University of Cambridge focused on interdisciplinary collaboration of technologists with the Arts, Humanities and Social Sciences. The lack of convergence shown in both bibliometric analysis (from Liu and colleagues and from Blackwell) raised the opinion that, probably, the purpose of HCI was not to create a stable body of knowledge, but to be a trigger for innovation, and to be "questioning, provocative, disruptive and awkward in relation to other disciplines" (Blackwell, 2015, 7).

Finally, Blackwell suggested that HCI should look for the expectation of innovation and new technological advancements. He also commented that interdisciplinarity and innovation were related with unexpectedness, and argued that new, innovative pieces of research arose when different teams of researchers from different disciplines and with different points of view came together to address one topic, which was of interest to all the participating parts. That was the moment where discoveries were made. Blackwell (2015, 9) later concluded:

> HCI is not about static knowledge, but ways of deploying and engaging with knowledge in a technological setting. If so, HCI should not aspire to be a discipline, measured through bibliometric convergence on core findings, but rather a mode of challenge and provocation – although one that is characterised by humility, playfulness, invention and rigorously honest reflection rather than confrontation between alternative disciplinary frames.

[7] http://chi2015.acm.org/authors/selecting-a-subcommittee/, Last accessed on 20 January, 2024

As discussed before, the transformation of the professional translation workflow by the introduction of new computer-assisted aids is now a reality (see Chapters 2 and 3). This fact has been studied from many different points of view in the literature, but mainly focusing on translation productivity or translation quality, as suggested in previous chapters. Less attention has been paid to translators themselves, to what they experience in these interactions with technology. Thus, following Blackwell, this PhD dissertation aimed at conducting a "questioning, provocative and disruptive" HCI-centered user study of this nature.

4.2. Usability

When looking more in-depth into HCI, we need to analyse what are the key elements of the interaction of a user with a technological device. According to Nielsen (1994) and Faulkner (2000), early computer interfaces were designed only for specialist users, but when hardware costs decreased, and the modern Internet and the personal computer appeared in the 1990s, computing and technological devices started to be used by many more users. Most of these new users had no specialist knowledge. Therefore, it was at this point when computer manufacturers started to focus on the ease of use by users, and the term "usability" came into play. Shackel (2009, 2) defined the usability of a system as:

> The capability in human functional terms to be used easily and effectively by the specified range of users, given specified training and user support, to fulfil the specified range of tasks, within the specified range of environmental scenarios.

In Shackel's (2009) definition, "easily" referred to a specified level of subjective assessment and "effectively" to a specified level of (human) performance. In a similar way, but in a more recent definition, according to the International Organization for Standardization, in their ISO 9231-11:2018 (ISO 2018) standard, "usability" is:

> The extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

In the ISO's 9231-11:2018 (ISO, 2018) definition, "effectiveness" refers to a certain level of completeness, accuracy, or quality to which specified tasks are completed (Bevan et al. 2016; Cowan 2011) and "efficiency" to the effort of the user in terms of time or cognitive resources (Bevan et al. 2016; Cowan 2011). We can see that these two terms refer to objective outcomes of interface interaction. According to Shackel (2009) and Cowan (2011), it is worth noting that users may opt for an interface that is more pleasurable to use, while sacrificing speed and accepting more errors. Users are the ones interacting with the different interfaces and they are the ones who say whether they would like to use specific interfaces in the future or even in their daily work. However, measuring the usability of a user interaction by measuring only objective outcomes may not be the best option. It is here that the last term, "satisfaction", comes into play. In the ISO's definition of usability, "satisfaction" refers to the personal perceptions and attitudes of the user towards the system or the interface (Cowan, 2011; Bevan et al. 2016).

In a more profound but similar definition of usability, in his seminal work, Nielsen (1994) deconstructed the concept of usability into five different attributes, which are explained below. The first attribute is 'learnability', which, according to Nielsen (1994), is one of the main elements in the HCI and usability context. When users try a system, their first experience is whether the system is easy to learn or not. Therefore, when using a system with good usability, the learning curve should be steep for novice users. This allows the user to achieve a reasonable level of proficiency in short periods of time, not having to spend much time learning all the particularities of the system. Ease of use or learnability may be the easiest attribute of usability to measure by having some users —who have never used the system before— use the system, and measuring how much time they take to acquire a certain level of proficiency. It is important to consider that users should be representative of the intended users, and the experience should be controlled. We cannot evaluate a post-editing CAT tool with some users with 10 years' experience in post-editing mixed with other users who have never performed a post-editing task.

The second attribute mentioned by Nielsen (1994) is the efficiency of use. This attribute also appears in Shackel's (2009) and the ISO's (2018) definitions. It refers to the level of performance of the user when the learning curve flattens out. Normally, involvement of "experienced users" is required and, for this, we must define what "experience" is.

Sometimes, “experience” can be assumed whenever a user has been using a system or interacting with an interface for some time already and has at least basic knowledge of it. There are different ways to find “experienced users”. One example may be to ask users to work for a specified number of hours and, after this time, using their efficiency as the base to recognize them as “experienced users”. Another way of finding experienced users is assessing the learning curve, and seeing at which point the learning curve flattens out, therefore establishing that point as the moment where users are experienced.

The third attribute presented by Nielsen (1994) is memorability, which is the fact that a system is easy to remember. Memorability is applicable to casual users, that is, users who do not use the system every day, or who have not used the system in some time (e.g., after holidays or sick leave). Normally, interface memorability is calculated with a group of casual users, assessing whether they have difficulties using the system after some time off.

Attribute number four is errors. An error could be defined as any action that did not achieve its intended goal. In the HCI context, errors may be thought of as the number of actions taken until a user achieves a certain task. If a user could perform a specified task in one action, but took ten, we could say that there were nine errors. In the translator-computer interaction context, errors may involve not using the system features properly (e.g., not accepting adequate translation completion proposals of an IMT workbench for IPE tasks). Therefore, error rates may give information on the usability of a system or interface.

The fifth and final attribute proposed by Nielsen (1994) is once again satisfaction, that is, how pleasant it is to use a system. This is a major attribute in usability, as the attitudes of the users toward systems are important for many aspects: if the user is not satisfied when performing certain tasks with the system, they are not likely to use the system again. In addition, previous research has found that human agency and the sense that users are the ones with control over the computer improved the attitudes of the user towards technology, and the interactions and systems were better valued in terms of satisfaction (Kay 1970). Subjective satisfaction questionnaires are a good way to approach user satisfaction evaluation. They are conducted after the experiment, once users have tested or interacted with the system. There are multiple, generally-recognized questionnaires on satisfaction in the HCI world. Likert scales are very common. Another type of measurement are semantic differential scales. In this case, users are shown two opposite terms (e.g., simple and complicated; pleasing and

irritating), and have to rate their interaction with the system, getting closer to or further from each of the concepts. When using these scales, it is recommended to have a baseline or multiple evaluations of systems, so results can be compared.

After analysing these three definitions of "usability", we can say that there are two main elements to be considered when talking about "usability". On the one hand, the objective attributes or performance (efficiency and effectiveness), which may give us indications on how successful the interaction of a user with an interface was; on the other hand, the subjective attributes or perceptions (satisfaction), which may indicate how the user has felt in this interaction and what they think about the interaction. Once this is clear, how should developers measure usability? Both objective (efficiency and effectiveness) and subjective (satisfaction) usability measures of the user interaction with a system can be designed and assessed (Cowan, 2011), and the following sections describe these processes in detail.

4.2.1. Usability engineering and testing

In the HCI literature (Nielsen 1994; Shackel 2009; Faulkner 2020), it is widely accepted that usability should not be considered as a one-step element in system design, but a multiple-steps process. This process is called "usability engineering", which takes place during the whole lifecycle of a product. During usability engineering, multiple elements of a system are assessed and evaluated through different iterations to see whether the system is usable. What is this process? Nielsen (1994) proposed a usability engineering model with different steps.

The first step is to know the user. According to Nielsen (1994) and Faulkner (2000), always, when trying to build a usable system, the first thing to do is to establish who the intended users are and how they will use the product. Defining the concept "user" is of key importance because intended users will change from system to system, and their needs or tasks will vary. Though system developers or researchers may try to think as if they were one intended user, results are not as effective as if actual users are asked. Therefore, individual user characteristics are important, and always engaging with intended users of the system is required. By knowing the background of the user (e.g., their educational level, age, work, and computer experience, etc.), we may anticipate what problems or needs they will have when

interacting with the system, but this is not enough, and actual testing should be carried out. To get to know the user correctly, it is necessary to analyse the task they will be performing. The goals, needs, aims, and troubleshooting of users should be known to get a better understanding of their interaction with the system. It is also worth stressing that user attitudes towards a system change after they use the system for a period of time. This may be caused because they learn to use it, or because they find new ways of achieving their goals. This learning effect is called the "coevolution of tasks and artifacts" (Carroll and Rosson 1992). Therefore, we should bear in mind that users' behaviour and/or performance may and will likely change after they get used to the system after some interaction. This supports the idea of doing longitudinal studies. For instance, when sampling translators for the studies of this PhD dissertation, it was important to choose wisely which users were needed, analysing in-depth their skills, abilities, and knowledge of the systems evaluated. First, we worked on a CAT tool, which is the regular working environment of professional translators nowadays. Therefore, it would not make sense if the translators participating in the study had no experience with CAT tools. Secondly, we worked with MT features incorporated into the CAT tool. Have the users of the study post-edited before? If yes, how many years of experience of post-editing did they have? As commented above, we could not compare novice users (without or with very little post-editing experience) with experienced users (with many years of post-editing experience). These were important elements that were considered when sampling the translators for the data collection part of this PhD and are further addressed in the Methodology chapter (see Chapter 6).

The second step of the usability engineering process proposed by Nielsen (1994) was doing a competitive analysis. Prototyping is a crucial part of the usability process because systems need to be developed in accordance with established usability guidelines and results from empirical user tests (Faulkner 2000). If the user test results are positive, it can be claimed that a system is competitive and helps the user perform tasks easily to achieve their goals.

Establishing the goals of the system from the very beginning is the recommended third step in usability engineering (Nielsen 1994). A system may have countless aims, and they would all depend on what the developers want to achieve or the profile and knowledge of the actual users. There are multiple usability parameters that should be considered beforehand, and setting the usability goals of the system is therefore an important task. For example, is this

system designed for translators with experience in MTPE, and are translators required to do legal post-editing tasks? Another option may be that users are required to do subtitling post-editing tasks. Depending on each case and situation, as well as the goal of the task, usability goals may differ substantially. In connection with the usability goals, it is also important to analyse the financial impact. For example, will this system allow translators to translate faster and with the same quality level if compared with existing tools? Will this new system allow for production costs to be reduced? This latter element is also a fundamental goal to be considered in industry settings.

The fourth step is empirical testing or interface evaluation. As has been previously mentioned, user testing is key. For this, Nielsen (1994) proposed establishing severity rankings. When evaluating a system, typically multiple errors of a different nature will appear. Yet, these usability errors may differ greatly, and, for this reason, it is important to determine the severity level of the errors. Before testing the system with intended users, a possible solution may be to send a list with the most common errors to usability experts or experienced users of the system (or similar types of systems) and ask them to rate the severity of these errors. As experienced users, they will be able to indicate whether the problems are common or not, and whether they are major or minor errors (Nielsen, 1994). This type of user testing allows researchers to spot usability problems and to improve the system by applying the solutions proposed by the users, and having additional information from experts before the final user testing takes place. This way, through different iterations, the usability of systems normally improves and enhances the user experience, but two last methodological considerations should be taken into account when testing usability: reliability and validity. In terms of reliability, individual differences may seriously affect the results of usability tests with real users. For example, in a post-editing task with different translators, it is not uncommon to find substantial levels of between-subject variability in terms of translation productivity (Terribile 2023). Therefore, it is recommended to work with a higher number of users, so that results are more reliable. Statistical tests can be performed to estimate the significance of the differences between-users (Pedhazur and Schmelkin 2013). On the other hand, validity helps us measure whether the usability testing of the system measures the actual usability of the system in a real-life scenario with real users. Some of the most common validity problems are not choosing the appropriate users for the test (e.g., computer science students for assessing

an IPE workbench, as they have never worked as translators and probably will never do a post-editing task with a CAT tool either), or not including time or social constraints or influences (see Sections 3.2 and 3.3 for a further discussion on potential methodological mistakes when conducting IPE studies).

As a final recommendation for the usability engineering and testing process, Nielsen (1994) recommended that usability should be considered also after the live implementation of the system. Continuing to observe user interactions with the system after its launch will help to learn new aspects that could be improved or enhanced in subsequent versions or releases.

4.3. User experience (UX)

In the last two decades, technological advancements increased exponentially, and the computing and intelligent devices world experienced huge changes with the introduction of new breakthroughs. Devices were not only usable, but also more complex and fascinating, including new features that aimed to create more emotions in users. These changes had the effect that, in the HCI world, the term "usability" started to lose traction at the expense of "user experience" (UX), which is one of the most researched concepts currently by researchers in HCI (Albert and Tullis 2022). UX seems to capture better all the changes that are taking place in this rapid and fast-evolving field of knowledge (Dix, 2010).

Forlizzi and Battarbee (2004) commented that, since the development of computing and the increase in popularity of PCs and devices, analysing the experiences of users when interacting with products became more important. Usability was first considered as the most important aspect to be taken care of in HCI, but, with new technological devices, researchers started to suggest that there were many more important aspects than usability alone. According to Forlizzi and Battarbee (2004, 1), when using interactive systems, it was important to study "all aspects of experiencing a product — physical, sensual, cognitive, emotional, and aesthetic". Usability did not cover all these important aspects, as it focused mainly on ease of use and design, and therefore the term "UX" started to gain popularity. Design teams became more multidisciplinary and involved people from different fields of science and knowledge (Forlizzi and Battarbee 2004).

Hassenzahl and Tractinsky (2006) later commented on the controversy and discussions that the term "UX" generated. As a new term appearing in the HCI literature, more and more researchers took it up, resulting in different definitions and use-cases. The first studies about UX were mainly programmatic, and tried to convince HCI researchers that, in an interaction, there were many more aspects to be analysed than only the task-related and satisfaction-related ones considered in the usability paradigm. Alben (1996, 1) regarded UX as "all the aspects of how people use an interactive product: how they feel using it or how well they understand how it works". Overbeeke et al. (2002) had a similar idea of UX, suggesting that ease of use was not the only important element in the interaction of a user with a product, but also emotions, which were extremely relevant for the experience. These simpler and earlier UX definitions were later superseded by more conceptual studies, and the concept of UX evolved subsequently. Hassenzahl (2018) proposed a new UX model, suggesting all the aspects that constituted UX and should be taken care of (the subjectivity of UX, the importance of satisfaction, and the context of the interaction). In a similar framework to the one of Hassenzahl (2018), Wright, McCarthy, and Meekison (2004) suggested that UX was a set of feelings, emotions and particular situations and contexts where the interaction takes place, resulting in different experiences for the user. Interestingly, Wright, McCarthy and Meekison (2004) commented that, unlike usability (and usability engineering), researchers cannot design or engineer UX, but should design *for* UX.

The International Organization for Standardization is also concerned with UX, and in its ISO 9241-11:2018 standard (ISO 2018), centered on the ergonomics of human-system interaction, UX is defined as a "person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service", a definition aligned with the previous UX frameworks and definitions presented above. This latter definition introduces an essential element that is worth stressing: perceptions and/or responses of users that matter most are not the ones that appear after using the system only, but also the ones that happen before the use of the product, that is, user pre-task perceptions. Pre-task perceptions have an important role in the UX. For instance, if translators are not happy with the idea of using MT systems for post-editing, their experience will probably be negative even before starting to use the system. Thus, considering both the pre- and post-task experience of translators will

be of vital importance. This issue is a key element tackled in this PhD dissertation (see Chapter 5).

All the above-mentioned definitions and conceptualizations for UX had the same goal: establishing a concept of what good UX is. Hassenzahl and Tractinsky (2006) also gathered all research trends in the UX field and described three different research trends that UX researchers were following at that time.

In the first place, UX should go beyond analysing merely instrumental use, which was previously the main topic of HCI research. What does this mean? Since the inception of research on HCI, most research focused mainly on the achievement of certain goals when performing a task (Law et al. 2009). Therefore, user-centered studies were putting all their effort and energies into evaluating the task, looking for and testing usability, but the appearance of different elements, also of key importance in an interaction, compromised this early focus of HCI research. Some of the most important elements that appeared later were beauty (or aesthetics) (Alben 1996), diversion or intimacy (Gaver and Martin 2000), or novelty and stimulation (Hassenzahl, Burmester, and Koller 2003). All these elements implied something more than just ease of use and linked the attributes of the project with the needs and values of the user. After recognising the importance of these proposed aspects, the problem was how to understand and define them? And then, how to translate them into quality of the product or system? These are questions that current UX studies still try to answer.

The second research trend in UX was focusing on the affective and emotional aspects of the interaction. According to Hassenzahl and Tractinsky (2006), emotions are critical for a group of human-computer interactions, such as subjective well-being (Diener et al. 1999) or decision-making (Loewenstein and Lerner 2003). For example, in the language services industry, a translator with specialist knowledge in MT will likely be more open to accepting and performing post-editing tasks (Alabau et al. 2013). This translator knows what the MT engine can offer, and their emotions will help them tackle the post-editing task more easily. If the translator has never worked with MT output before, they are likely to feel discouraged and to reject that task when offered less money for the same number of words than in a "normal" translation task (Fırat 2021). Affective computing is thus a focus in the UX field because designing systems that aid irritated or dissatisfied users may lead to a better UX,

enhancing the interaction of the user with the system (Cockton 2002). This trend is now studying positive emotions that users feel when interacting with systems, like joy, fun and pride (Blythe and Monk 2018). UX research dealing with emotions is also divided in two different ways considering emotions and affect (Hassenzahl 2018): one research line focuses on the importance of emotions as a consequence of product use (Kim and Yoo 2021; Hassenzahl, Burmester, and Koller 2003), and another line centers on the importance of emotions before the use of the product and as evaluative judgements (Norman 2007).

The third and last research trend proposed is focusing on the nature of the experience. This view on UX has two main aspects regarding technology use. The first one is the situation and/or context, and the second is the temporality. In this approach, an experience is a combination of elements (regarding the user and the product, e.g., their mood, expectations, or goals) that extend over a period of time and have a beginning and an end. All these elements interact with each other and are related to each other. Thus, it is important to analyse the UX within a particular and situated context and scenario, as well as during a certain period of time, to see whether the UX changes, evolves and improves or worsens during the interaction. For instance, the changes in the UX of a translator using an IMT workbench for IPE tasks may happen throughout the whole interaction, from the moment a translator receives a task assignment, to the moment they start using a new IMT workbench, to the scenario where they are recurrent users of such a workbench and their perceptions towards the system and their UX may change over time.

After this discussion on UX, taking into account the different definitions proposed by researchers in the HCI field and the various trends in research, we can suggest the defining aspects of UX. First of all, UX should have more than a task-focused and instrumental use, as UX should always take the context and the subjectivity of the user into account. UX relies vastly on the internal state of the user (their expectations and perceptions) and depends on the features of the system (its complexity, usability, functionality), as well as on the context where the interactions take place (in a real-life working scenario, in a social setting or in a volunteering situation). All these elements make UX a much more complex concept than usability, as different elements must be considered. Aligned with this brief summary, Hassenzahl and Tractinsky (2006, 6) concluded with a thoughtful sentence, suggesting that, instead of following the traditional HCI assumption where a “good system” is one without

errors, the main goals of HCI should be to "to contribute to our quality of life by designing for pleasure rather than for absence of pain". If transposed to the context of this PhD: Do IPE workbenches, MT systems and post-editing tasks contribute to pleasurable or painful experiences when translating? These are some of the questions addressed in this PhD dissertation (see Chapter 5).

4.3.1. User experience from different angles

Since UX is a broad concept including multiple elements to be considered (satisfaction, emotions, feelings, specific contexts, etc.), there are different types of perspectives taken to research UX. The most common approaches to the study of UX are: interaction-centered studies, product-centered studies, and user-centered studies (Albert and Tullis 2022). As the focus of this PhD is working with translators, the UX studies that are reviewed and analysed in this section are user-centered ones.

The user-centered approach focuses on helping product developers and designers to understand the people who will be using their products. In the literature, this has been studied from different angles and points of view, according to what the goals were of the researchers. Instead of focusing on the traditional goal- and task-oriented usability paradigm, new frameworks appeared, which concentrated on the wide range of elements considered in the UX paradigm. For example, Hassenzahl (2018) studied the fun experienced by the user when interacting with specific products or systems, and whether the interaction was fruitful in terms of enjoyment. Users' motivations and context of use were also studied (Mäkel and Suri 2001), and Jordan (2003) proposed that a good UX should look to provide users with usability, pleasure, pride and functionality, as every product or system should activate all these senses before, during or after an interaction happened. Mao et al. (2005) later commented on the growing importance of user-centered design for achieving a good UX, as researchers started to focus on this topic, and user-centered design started to be a focal point in the main conferences on HCI. Thus, we can say that user-centered design is a key element in HCI, whether we talk about a website that leads a user to their goal with just a mouse click —instead of with nine mouse clicks— or if we talk about a translator using a CAT workbench including MT and TM functionalities that do not consider what their needs and objectives are (O'Brien et al. 2017). Negative experiences may result in users abandoning such systems to

look for other products offering better user experiences, or directly not adopting such systems in their daily workflows (Albert and Tullis 2022).

In 2006, taking into consideration all the previous frameworks and models of UX to date, Roto (2006) suggested a new holistic model, proposing that user-centered UX was a group of building blocks. Figure 4.1 presents Roto's UX framework. Firstly, we have the system, which was defined as "all products, services, and infrastructures that are involved in the interaction when using the examined product" (Roto 2006, 3). In an IMT workbench for IPE tasks, we could talk about the TM and the MT features being vital to the translation completion proposals, which are exclusive to these types of workbenches. UX is a combination of all the attributes of the system, so an IMT workbench will evoke different user experiences if compared when translators perform TPE or IPE tasks, as translation completion proposals may appear or disappear depending on the MTPE modality.

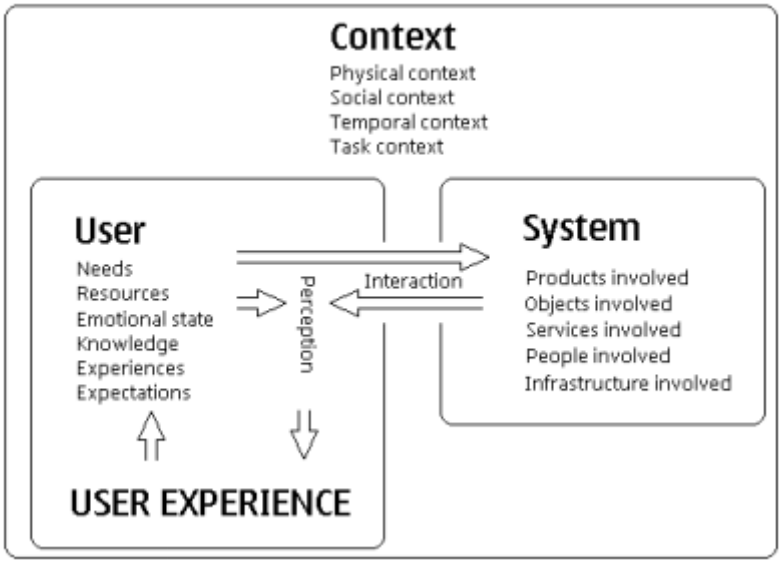

Figure 4.10. Building blocks for user-centered UX proposed by Roto (2006). Image retrieved from Roto (2006).

Secondly, Roto (2006, 3) talks about the context, which is "the physical, social, and temporal factors, and (optionally) the task context for the experience". This second component contains four types of contexts that should be analysed together. The physical context is everything the user can see or feel; for example, the office or desk where the translator is working, its ergonomics, the temperature, or the lightning (Schilit, Adams, and Want 1994). This supports some situated, embodied cognition work undertaken in the Translation Studies

field (Risku and Rogl 2020). The social context includes the expectations and influence that other people have about the user. In a real-life working scenario, project managers will expect to receive a high-quality translation within a limited deadline. Or, in a lab experiment, what expectations do researchers have when a translator participates in a study? Therefore, analysing the UX of translators in their normal context of interaction with technologies may be the best situation to obtain valid data. The temporal context refers to the period of time that a user can dedicate for the system under certain context restrictions. For example, if a client needs an urgent translation, will the translator perform light- or full-post-editing of a text to meet a tight deadline? The task context is also regarded as a key element of the context component, as the situation where the task is performed is also relevant.

Finally, the user is the most important component of Roto's (2006, 3) approach to UX, where the focus is on "the mental and physical state of the person who interacts with the system". Users are without any doubt the most important element in user-centered UX. In this component, we should analyse the personal experiences and expectations of the user, which will affect the interaction with the system. Is the translator used to post-editing? Does the translator know the system? What does the translator think about post-editing? This will influence vastly how the translator interacts with the system and the resulting UX after the interaction.

4.3.2. User experience evaluation

The wide acceptance of UX in the HCI world and its widespread use has undoubtedly had some consequences. As seen above, different researchers have focused on different experiential elements included in the "UX" concept (see Section 4.3.1). The questions that arise now are: how can we measure these experiential elements? Understanding and interpreting emotions is difficult (Forlizzi and Battarbee 2004), but researchers claim that, even if no generally accepted measures for UX exist, UX could be measured in different ways (Roto, Obrist, and Väänänen-vainio-mattila 2009). This section provides a summary of the most used UX measures and evaluation methods to date.

Bargas-Avila and Hornbæk (2011) reviewed 51 publications on UX from 2005 to 2009, to find which was the most common UX evaluation method in the field during that time. The data

sources used were three well-known scientific repositories: the ACM Digital Library, ISI Web of Knowledge and ScienceDirect. In this analysis, the authors found that most evaluation methodologies were qualitative scales and questionnaires, which were built on previous usability studies. Emotions, enjoyment, and aesthetics were the most analysed factors in the studies reviewed.

Some years later, Law, van Schaik, and Roto (2014) performed a second analysis and inspected 58 studies. The examined period was from 2010 to 2012. In this second analysis, most studies measured UX together with other cognitive factors (e.g. learning efficacy when reading) (Mumm and Mutlu 2011) and behavioural factors (e.g. task completion time). The methods employed by all studies were validated questionnaires or scales (e.g., AttrakDiff, the User Experience Questionnaire, Self-assessment Manikin, Flow State Scales, or PANAS, among others), and some studies measured psycho-physiological measures such as heart rate and keystroke patterns to link them with confidence, hesitance, relaxation, etc. (Epp, Lippold, and Mandryk 2011). Interestingly, both Bargas-Avila and Hornbæk (2011) and Law, van Schaik and Roto (2014) found that all research analysed was measuring UX in a non-work-related context. As mentioned previously (see Sections 3.2 and 4.2), in the framework of this PhD, analysing and studying non-work-related contexts may lead to biased and unreliable results because translators would not interact with the product in the actual way they do in their normal routines and working days.

Traditionally, usability was measured with stopwatches or logging keystrokes or user actions because these were objective measures that could provide information about the number of clicks or errors, or even about the time users took to execute a certain task (see Section 4.2.1). UX is subjective, and therefore these objective measures were not appropriate. This subjectivity meant that UX could not be measured by just observing users performing a task; UX should go beyond the objective level (Obrist, Roto, and Väänänen-Vainio-Mattila 2009). In addition, not only the “satisfaction” concept included in the “usability” definition was of importance here, but also users’ motivations, expectations, emotions and lived experiences (Kankainen and Suri 2001; Kaye 2007). In another review of UX evaluation methods, Vermeeren et al. (2010) also highlighted the importance of field evaluation instead of lab evaluation. Since the most important thing to analyse is what the user experiences,

investigating UX in the most realistic scenario possible is key, backing up once again the idea of analysing translators working from home or in an office, instead of from a lab.

Roto, Obrist and Väänänen-vainio-mattila (2009) performed a categorization of different UX evaluation methods, so that researchers could know which ones to use depending on their study: of the samples they studied, there were lab tests, field studies, surveys, expert evaluations, and a mix of these methods.

The first type of UX evaluation method proposed by Roto, Obrist and Väänänen-vainio-mattila (2009) was lab studies, which were a common and popular method in usability evaluation. In this type of evaluation, users are asked to perform a task with a product, and to think aloud during the interaction. This way, researchers may observe and analyse some actions and emotions of the user while carrying out the task. This is a good method for assessing UX in early phases of product development, to know early what users think of any interaction. An example of a lab study is the Tracking Realtime User Experience (TRUE) method (Ganglbauer et al. 2009), which includes psycho-physiological measurements (Hazlett 2006; Mandryk, Inkpen, and Calvert 2006), and these types of methods sometimes require a controlled setting and specialized equipment such as eye-trackers and key loggers.

The second method proposed were field studies. Again, since that context has a great importance and effect on UX, it is relevant to analyse user experiences in real-life scenarios (Fields et al. 2007; 2008). Field studies can also be longitudinal to see how the experience of the user evolves during the interaction with a product or system (Kujala et al. 2011). This latter method may be interesting for a longitudinal study on the UX of translators doing IPE tasks, so we can reduce the experience difference that TPE vs IPE tasks entail.

Another feasible method in UX evaluation is to carry out surveys. They are a convenient way of getting feedback from real users and, due to the fact they can be done online, surveys allow for reaching a much bigger group of participants than a lab study. Roto, Obrist and Väänänen-vainio-mattila (2009) mentioned AttrakDiff™ (Hassenzahl, Burmester, and Koller 2003), Emocards (Desmet 2002) and the User Experience Questionnaire (Laugwitz, Held, and Schrepp 2008) in this context.

Expert evaluation is another method for evaluating UX. To cut costs in recruiting participants from the exact target group to interact with the evaluated system, a common step in early

prototyping of the product is to have usability and field experts evaluate the product with usability heuristics (Nielsen 2010). It is recommended to go over this step in every project before doing the user test, in a pilot experiment, which would allow for detecting and correcting minor usability issues that would ruin the expensive and time-consuming user test. The most common method is to use a heuristics matrix or to focus on a specific experiential element of the interaction and then have an expert rate this element (e.g., enjoyment in the interaction).

The final UX evaluation method proposed by Roto, Obrist and Väänänen-vainio-mattila (2009) was a combination of different evaluation methods mentioned above, like mixing objective observation data with a keylogging program while collecting subjective feedback from the user through questionnaires, interviews or surveys. This way, richer data can be collected from users and interactions, and therefore the analysis may bring more robust results after triangulation.

As a conclusion for this section, in the HCI field, there has been substantial discussion about the methodology for measuring UX, and different methods have been proposed depending on the objective of each researcher or study (Obrist, Roto, and Väänänen-Vainio-Mattila 2009). Some examples put the attention on the Hedonic Quality (HQ) of a product, and pay closer attention to more subjective emotions, hedonic elements or sensations (Hassenzahl, Beu, and Burmester 2001), while others have focused on the Pragmatic Quality (PQ) of a product, paying closer attention to a mix of subjective and pragmatic elements (Vermeeren et al. 2010). Therefore, the most appropriate UX evaluation method will depend on the goals of each research project, and whether a more HQ or PQ focus is desired. However, the conclusion that has been reached is that questionnaires are the most viable tool for collecting UX measures, and there are different questionnaires that are most commonly used in terms of UX in the HCI world, specifically AttrakDiff and UEQ (Law et al. 2009).

4.4. Studies of HCI factors in Translation Studies involving MT

The previous sections of this chapter, containing definitions, frameworks, and evaluation methods of key concepts of the HCI world, allow for easily differentiating usability from UX. These two terms are central to the study of today's interaction with digital devices, products,

or systems. Current HCI studies focus on a wide variety of topics, such as intelligent personal assistants, user speech production, and communication with interactive voice or text response systems (Clark et al. 2019), just to name a few. If we look closer at these topics, we can see that there is one in-depth element that is common to all of them: language. Whether users talk to intelligent personal assistants like Siri or Alexa, or communicate with interactive voice response systems, one of the most important factors that enables these interactions is language. Some studies demonstrate that the lack of language coverage for intelligent personal assistants hinders usage and hamper accessibility depending on the native language of the speaker (Wu et al. 2020). This highlights the importance of communication, language, and translation in making all these technologies available to the wider public. Nevertheless, as (O'Broin 2011; 2012a; 2012b) commented a decade ago in the magazine *Multilingual*, the translation or MT fields have not paid attention to the critical role of translation and language in enabling the interactions and experiences of users with intelligent and digital devices. In this section, a literature review of HCI factors studied in Tanslation Studies and the MT community is presented. To conduct this review, proceedings of all major venues and conferences in Machine Translation (AMTA, EAMT, MT Summit, etc.) and HCI (CHI, ACMMM, ACM UI) have been analysed by looking for specific keywords and specific words in the titles or abstracts, namely "UX", "experience", "usability", "HCI", "human-computer interaction", "interaction". From the articles found, abstracts have been read to see whether they studied any HCI factor in relation to translation and MT. Also, this search method was used in Google Scholar and other scientific repositories such as ScienceDirect, Scopus and Web of Science. Of all the articles found, only the ones that appear in sections 5.1 and 5.2 were relevant to this PhD research.

4.4.1. Usability in Translation Studies involving MT

Traditionally, a popular branch of research in Translation Studies is the study of the translation process (Alves and Jakobsen 2020), focusing mainly on the particularities of the translation task by using keylogging and eye-tracking methodologies (as alluded to previously in Section 2.3.1). This allows researchers to focus on the analysis of the time spent in performing certain translation or post-editing tasks, the number of keys pressed, etc. Data obtained from translation process research is objective, and there may be some elements that cannot be

investigated or fully observed or studied without taking subjective feedback or user perceptions into consideration (Bundgaard 2017).

Thus, special attention is also paid to subjective feedback in Translation Studies, which allow for introducing usability into translation process research. A literature review demonstrates that two main approaches to analysing usability can be observed. The first approach looks for direct feedback from **translators** when interacting with MT. For instance, Etchegoyhen et al. (2014) claimed that subjective feedback from translators is an important aspect when considering the usability of MT in a translation workflow, and that translators' perceptions should also be accounted for along with other typical indicators, such as translation productivity measurements. In a study with 19 translators, Etchegoyen and colleagues (2014) ran a questionnaire to capture what translators thought about their post-editing tasks in a subtitling environment, and the process and the usability of introducing MT in a subtitling workflow was negatively rated overall (the results showed an average of 2.37 on a 5-point Likert scale, where 1 was regarded as a poor post-editing experience, and 5 as an excellent experience). Similarly, Torres-Hostench et al. (2017) analysed the usability of a mobile app for post-editing comparing keyboard and voice input. The usability was assessed through think aloud protocols and discussion groups between five participants and five observers. Later, Teixeira et al. (2019) conducted two usability experiments of a desktop interface to conduct post-editing tasks via touch and speech input, and collected the satisfaction of participants via questions and open comments, though most participants preferred the traditional keyboard and mouse input modality for post-editing.

The second usability approach is the most common in the literature, and changes its focus from translators to **end-users**, that is, the readers of machine translated texts. By following this second approach, the end-user is the central point of the study, and the evaluated elements are the effects that a machine translated text has for usability and acceptability from the end-user (reader) point of view (Suojanen, Koskinen, and Tuominen 2014). In this latter line of research, other researchers also analysed end-user comprehension of machine translated text (Roturier 2006; Stymne et al. 2012), but they only considered the subjective reception of the final text (i.e., how usable was the final text) and did not consider objective measures, such as productivity. Therefore, the "effectiveness" and "efficiency" factors of the

ISO definition of “usability” were not assessed, and we conclude therefore that they studied usability only partially.

In a more recent study, Doherty and O’Brien (2014) assessed objective and subjective elements, more specifically, the usability of unedited MT output through eye-tracking and a post-task questionnaire looking at goal completion, satisfaction, effectiveness, and efficiency. This study did not deal with translation tasks, but with the impact that machine translated instructions had on users interacting with online documentation. They therefore followed the second approach to usability in Translation Studies mentioned above. The goals of this research were to study whether goal completion, satisfaction, effectiveness and efficiency were affected by the instructions the users read to complete a certain task. One set of instructions was originally written in English, and the other set of instructions was machine translated into Spanish, French, German and Japanese, without any type of editing. Thirty native speakers participated voluntarily in the study, and they had to perform specific tasks with the help of instructions written in their mother tongue. There were 15 English, 4 French, 3 German, 4 Spanish and 4 Japanese native speakers; only the English speakers read the original instructions, while the rest of the speakers read a raw machine translated set of instructions. A post-task questionnaire was used, including 12 items in a 5-point Likert scale, and results showed that the English participants rated their instructions higher than any other group of participants who read the machine translated instructions. The rating of some questionnaire items such as comprehension, end-user satisfaction or the value of the instructions to complete the required task showed statistical significance. From this study, the authors concluded that end-users’ reception of texts was worse in machine translated texts than in originally written texts.

Later, in her PhD thesis, Castilho (2016) analysed the acceptability of machine translated text for end users. Here, acceptability was understood as a combination of usability, quality, and satisfaction. In a first pilot study, 18 Brazilian Portuguese native speakers were asked to interact with a product by reading some instructions. These native speakers were divided into two groups; one group had to read the raw MT instructions, without editing, while the second group read the post-edited instructions. The goal was to perform certain tasks, such as setting up an automatic calendar within a time tracking product. Participants’ actions were monitored with an eye-tracking tool by looking at eye fixation time, count and duration. Also,

participants had to complete a 5-point Likert scale questionnaire on their level of satisfaction after performing the required tasks by rating the instructions provided. Results showed that post-editing increased usability and user satisfaction in comparison with raw MT instructions (Castilho et al. 2014). In a second part of her PhD, Castilho (2016) also studied the usability of people interacting with a spreadsheet program. In terms of usability experiments, eye-tracking software was used and, following the ISO definition of usability, effectiveness (via goal completion), efficiency (in terms of task time and goal completion) and cognitive effort (via fixation duration and count, and visit duration and count) were considered. Two groups of users were studied: translators and people who read the final translation. Participants interacted with different translation modalities (raw MT output, post-edited texts, and human translations), which were also analysed. The final instructions were considered usable if participants could use them satisfactorily in the intended context of use. In addition, end-user satisfaction was also studied in terms of how pleasant it was to use such texts as instructions. For satisfaction, both a post-task questionnaire (in a Likert-scale way) and a satisfaction survey (where users had to answer YES/NO to whether the information was useful) were carried out. Results showed that post-editing significantly improved acceptability of target-text readers in all the language combinations analysed (Castilho and O'Brien 2017).

4.4.2. User Experience in Translation Studies involving MT

With researchers from Translation Studies looking each time with more interest to the HCI world because of the digitalization of the translation field (O'Brien 2012b), the term UX started to appear in some studies of MT.

Bowker (2015) was the first person to reportedly study UX after reading O'Broin's (2011, 2012a) reflections of UX in the language industry. As a consequence, Bowker (2015) carried out a study to analyse whether the UX and the translatability of a website text were correlated. In this study, Bowker did not follow the ISO definition of UX, and instead regarded UX as "things such as whether a product is easy to figure out, whether it is difficult to accomplish simple tasks, or how it feels to interact with that product (e.g. satisfying, frustrating)" (Bowker 2015, 3). Two sample texts were then re-written by following UX-oriented guidelines (e.g., to avoid jargon and write simple sentences) or translatability-oriented guidelines (e.g., use the active voice and write short sentences). Then, text readers

were asked to provide feedback on their "experiences" when reading these texts. One-hundred and seven people were asked to evaluate the source-language texts. 61% of participants preferred texts written following the UX-oriented guidelines, in comparison with the 39% of the translatability-oriented ones. Then, the sample texts were machine translated, and three professional translators evaluated the quality of the resulting MT raw output in terms of adequacy and fluency. All translators rated the translatability-oriented text higher than the UX-oriented text. The final study of the paper was carried out on the target-language. A recipient questionnaire was undertaken, asking native target-language readers to state which MT output they preferred. 62% of participants preferred the text written by following the translatability-oriented guidelines. Bowker finally concluded that, as translatability increased, the UX of source-language readers decreased, while the UX of target-language readers increased. Bowker and Ciro (2018) carried out a second study, with a bigger sample, which also used Bowker's (2015) definition of UX, but the focus was still on acceptability and readability of texts from an end-user perspective, more specifically for readers of MT content. Therefore, we can say that these two studies were more usability-focused than UX-focused, in line with previous studies of usability and in accordance with the latest ISO definition of UX considered in this literature review (ISO 2018) (see Section 4.3).

In a smaller-scale experiment, Matusov, Wilken, and Georgakopoulou (2019) undertook a user experiment with two translators, who worked in the audiovisual translation field, and subtitled a program with MT assistance. According to the authors, translators were asked to rate "their MTPE experience", but they only had to rate if they liked post-editing MT output in the subtitling task with a 5-point Likert scale, and the average obtained was a 3. As we have seen in the theoretical UX section, UX should include pre-task perceptions, measurement of a wide range of emotions while performing the task, and the experience after doing the task, so we cannot claim to have addressed UX by just offering a simple Likert scale after interacting with MT. Though the authors claim to have addressed UX, this study should also be classified as a usability-oriented study. We could even say that the study by Matusov and colleagues (2019) looked at usability in a rapid and superficial manner, only using a simple Likert-scale questionnaire after users performed a task with a tiny sample (two translators), which did not allow for the use of statistical analysis.

More recently, Guerberof-Arenas, Moorkens, and O'Brien (2021) published a paper called "The impact of translation modality on user experience: an eye-tracking study of the Microsoft Word user interface", researching the impact of different translation modalities on the "user experience". Yet, considering strictly the ISO definition of UX, Guerberof-Arenas, Moorkens and O'Brien (2021) mainly measured usability, considering the following elements: (i) Effectiveness, which was calculated through task completion. The more tasks users completed within the allotted time, the more effective they were; (ii) Efficiency, which was measured after analysing the number of tasks users completed in relation to the time it took to complete the tasks. The lower the time spent to complete the tasks, the higher the efficiency of the user; and (iii) Satisfaction, which was measured with the IBM computer usability questionnaire (Lewis, 1995), where users had to rate a series of statements on a 7-point Likert-type scale. Gaze data were then replayed, and translators were interviewed through a think-aloud protocol. Guerberof-Arenas, Moorkens, and O'Brien (2021) concluded that effectiveness was not significantly different when observing different translation modalities. Yet, the efficiency and satisfaction values showed statistical significance depending on whether users worked with a computer program translated by a human or a machine. Once again, these users did not interact with the MT assistance themselves in an editing/translating workflow, but worked with machine translated content. Interestingly, the previous experience of participants impacted substantially on how well they performed the tasks, even if the MT output was not correct. This is also applicable to IPE, as seen in Chapter 3, because previous experience in post-editing tasks results in less resistance to the introduction of (new) technologies in translation workflows and reduces translators' reluctance to technology (Alabau et al. 2013; Sanchis-Trilles et al. 2014). The participants of Guerberof and colleagues' study (2021) were not people interacting with MT, and the goals of the study were to evaluate whether the translation modality affected the acceptability or usability of an application (MS Word). Once again, there is no analysis on UX specifically as per the ISO definition of UX, and we could therefore say that the authors undertook a usability-oriented study. This is also supported by the fact that the subjective questionnaire used was named "IBM computer usability satisfaction questionnaire" (Lewis 1995).

As can be seen in the review of literature thus far, most researchers claim to have researched UX in relation to MT, but none of them has actually followed a current UX-methodology.

Instead, usability methods and questionnaires have been used, focusing mainly on efficiency and effectiveness, neglecting user satisfaction. In the theoretical section on UX (see Section 4.3), we have seen that there are many more aspects to be considered when researching UX, which have not been applied in the above studies of this section.

Koponen et al. (2020) have been found to be the first researchers to actually focus on UX involving MT, as per the ISO standard definition of UX. Koponen et al. (2020) collected feedback from 12 translators who produced subtitles through MTPE. It is worth stressing, however, that these translators had no experience in combining subtitling with MTPE. Translators' feedback was collected through a UX questionnaire and later with semi-structured interviews. The main research questions sought to know whether translators thought that these subtitling tasks were positive or negative, to understand their experience, and to study how this experience could be improved. Koponen and colleagues (2020) slightly modified the User Experience Questionnaire (UEQ) created by Laugwitz, Held, and Schrepp (2008) to adapt it to a subtitling task, with the aim of collecting the perceptions of translators after performing the post-editing tasks. The UEQ uses 26 opposite adjective pairs (e.g., easy/difficult) on a 7-point Likert scale to rate the experience of users when interacting with a product or system. The authors of this study amended the original UEQ and reduced the UX questionnaire to 13 different adjective pairs. Then, they converted the UX results and established experience thresholds from a range of -3 to +3, where average scores between -0.8 and +0.8 were neutral, scores below -0.8 were considered negative, and scores over +0.8 were deemed positive. As for the UX results, most evaluated adjective pairs were neutral, and none of the ratings assigned by the translators fell below the -0.8 value. The most negative adjective pair overall was "limiting/creative", which is normal in an audiovisual translation setting because subtitle segmentation and character limit play an important role in this translation domain. Semi-structured interviews were then carried out to collect more feedback from the translators and how they experienced the MTPE tasks in such a subtitling workflow. The interviews were later transcribed, anonymised, and analysed thematically with the software Atlas.ti, with a special focus on identifying positive and negative comments. Koponen et al. (2020) identified 143 total statements, where 55% were negative, 29% positive and 15% neutral. Most negative statements focused on specific spotting or segmentation issues related to the subtitling domain. In terms of the positive statements analysed, the vast

majority were acknowledging the good results and solutions that the MT output offered, more specifically about useful lexical options or terminology use. Out of the 143 statements, 42 stated that MTPE had a strong impact in the subtitling process, and most of these comments were negative, as translators perceived that MT reduced their productivity and the creativity of the resulting post-edited output was lower than if translated without MT assistance. Yet, these were only user perceptions, which may not correlate with empirical data and measurements, where quality or productivity need to be measured and triangulated with these subjective data (as suggested in Section 4.3). Interestingly, translators proposed different improvements to the systems used, though they were not asked about them. From Koponen et al.'s study (2020), we can therefore conclude that the experience of subtitlers was not negative (none of the elements evaluated fell below the –0,8 value). Using the amended UEQ version allowed for identifying the friction points of subtitler-MT interaction (e.g., segmentation, creativity), and allowed for identifying areas of improvement in the development of translation technology tools for achieving a better UX.

Building on the previous study by Koponen et al. (2020), Karakanta et al. (2022) conducted a new study with 22 subtitlers, who were asked to post-edit automatically generated subtitles. Then, subtitlers' UX was collected with Koponen et al.'s amended UEQ, and results were also normalised to UX scores ranging from -3 to +3. The authors shared that the subtitlers' UX was neutral or positive, as 12 of the 13 pairs of adjectives analysed had an average UX score above 0, and the only pair of adjectives with an average negative UX score was very close to 0. Finally, Karakanta and colleagues included a questionnaire with open questions to investigate subtitlers' opinion about the MT quality and subtitlers' perceptions of automatic segmentation and automatic subtitling. The answers to these open-ended questions were then coded with thematic analysis (Braun and Clarke 2006), and analysed further. The thematic analysis results showed that the main issues of automatic subtitling originated from failures in speech recognition, which caused error propagation, translations out of context and inaccuracies.

Although Koponen et al. (2020) and Karakanta et al. (2022) were the first to use UX evaluation methods from the HCI field, the UEQ they used underwent major modifications. A validated questionnaire with 26 adjective pairs was adapted to a questionnaire with only 13 adjective

pairs. These amendments may compromise the validity of the questionnaire (Taherdoost 2016), and further methodological consideration should be undertaken.

4.4.3. Ergonomics, situated interactions, and hedonomics in Translation Studies involving MT

Besides "usability" and "UX", there is a third important concept in the HCI world that has caught the attention of translation and MT researchers: "ergonomics" (or human factors). According to the ISO 9241-11:2018 standard (ISO 2018), named Ergonomics of Human-System Interaction, a more recent definition of ergonomics is:

> [a] scientific discipline concerned with the understanding of interactions among human 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 (ISO, 2018)

Ergonomics therefore tries to understand better the interactions between humans and computers, and not only consider system/task performance, but also human well-being. Traditionally, the study of ergonomics focused on preventing people's pain in their workplace but has evolved to a discipline devoted to improving the physical and cognitive environment of people by changing, designing and redesigning elements of their surrounding environment (Salvendy and Karwowski 2021). User interfaces should be easy to use for users to have an appropriate interaction, and it is the same with translation technologies. For example, translation memories are supposed to ease translators' cognitive effort (Muñoz Martín 2012) because translators are automatically offered already (partially) translated segments.

When speaking of ergonomics, Doherty and King (2005, 2) commented that "[s]ystem development projects have typically been viewed as exercises in technical change, rather than socio-technical change", not taking into account the thoughts and experiences of users, in line with Olohan (2011, 6), who added that "the human and organizational aspects are not addressed at all, or only implicitly, […] when the system is being developed". Not listening or paying attention to the users of a system may be considered a critical issue in other industries, but why does this not happen in the MT community and, more broadly, in the translation technology fields? This may have serious consequences. Thus, what translators think while they translate is not the only thing that affects performance, but also how translators interact

with their environment and context. This leads us towards considering current translation workflows as “situated” activities.

Suchman (1987; 2007) was one of the early researchers who started theorizing about the situatedness of activities and cognition. In translation studies, situated cognition was first introduced by Vienne (1994), in relation to the functional approaches to translation, suggesting that translating was not only a problem-solving task with a specific text, but there were many other important aspects, like the context of the translator, the type of assignment and text, etc. More recent research on situated cognition has also explored practice theory (Olohan 2017), ergonomics (Ehrensberger-Dow and Heeb 2016), and embodied, embedded, extended, enacted, affective (4EA) cognition (Risku and Rogl 2020).

In the contemporary language services industry, which implies, without any doubt, a form of human-computer interaction, Ehrensberger-Dow and Heeb (2016) studied the impact of the translation context (e.g., the physical conditions of the translation workplace or the use of language technology) on the cognitive aspects of the translation process by recording the screen of the translator, as well as the room where the translator was working. They concluded, “If translators are overly constrained (e.g., by the tools they use), they may adjust their cognitive processes and actions to fit those constraints instead of searching for creative solutions to the problems that TM and their other language technology tools cannot properly deal with” (Ehrensberger-Dow and Heeb 2016, 13). Ehrensberger-Dow and O’Brien (2015) studied “cognitive friction” and the role of ergonomics in complex translator-computer interactions, with a focus on translation workplaces for both in-house translators and freelance translators, and found different common issues (e.g. sitting for too long translating) and commented on ergonomic solutions to relieve these problems. Some other research has been carried out on ergonomics by taking into account the importance of the situated interaction of translators with their environment (tools, computers, offices, etc.), and researchers have studied how ergonomic issues in translators’ workspaces impacted their efficiency and the quality of their resulting translation, specifically at the moment where a strong digitalization of the translation process was taking place, and TM and MT technologies were merging to “assist” and “ease” translation workflows (Ehrensberger-Dow and Massey 2014; Ehrensberger-Dow et al. 2016; Ehrensberger-Dow 2014; 2017; 2020;). In addition, the ergonomics of specific tools have also been analysed, by focusing on the interaction of

translators with CAT tools, suggesting that tool developers actually did not take into account users' perspectives when designing such tools (Lagoudaki 2008), which may be irritating in some real-life scenarios (O'Brien et al. 2017). These ergonomics studies also support Doherty and King's (2005) and Olohan's (2011) comments on the non-inclusion of user feedback in system development.

In this context of ergonomics and pain relief, Hancock, Pepe, and Murphy (2005, 1) coined the term "hedonomics" as "[the] branch of science and design devoted to the promotion of pleasurable human-technology interaction", and introduced this concept into the human factors/ergonomics lexicon. The main purposes of hedonomics were (1) to promote pleasant and enjoyable human-computer interactions and (2) to promote well-being through technological augmentation (Oron-Gilad and Hancock 2017). Here, again, appears the 'human augmentation' concept proposed earlier by Engelbart (1962) (see Section 4.1). While ergonomics is more focused on preventing pain, hedonomics is centered on the promotion of pleasure. Yet, Oron-Gilad and Hancock (2017) highlighted the importance and difficulty of setting the boundaries and balances between affective- and productivity-driven frameworks. A hedonomic interaction is not a utopia where people interact with systems without demands or constraints. Although productivity and performance are very important elements in today's society for economic reasons, user satisfaction and well-being should also be analysed and considered in human-computer interactions, and, ultimately, be taken into account in system development. Hancock, Pepe, and Murphy (2005) proposed a framework for the study of hedonomics. In the development of any product or system, according to Maslow's (1958) model of optimization of human satisfaction, there are some low-level needs that need to be satisfied before dealing with other higher-level needs. In this framework, safety for users should be the first concern, and only after ensuring the product or system is safe, functionality becomes a priority. Then, only after ensuring the lowest-level need, we can move forward to the next level need. When functionality is achieved, usability should be considered. Once a system is usable, we can then move to promoting pleasurable experiences to the user. It is therefore of utmost importance to set the hierarchy levels in the translator-MT interaction. The hierarchy may change depending on the context, as we deem the translation activity as a situated form of human-computer interaction, but users' pleasure in the interaction should be looked for and considered regardless of the context this interaction takes place in. As the

review in Section 4.4 shows, the experiences of translators in modern translator-computer interactions have received very limited attention.

4.5. Human-centered, augmented MT (HCAMT)

In this context of increasing attention towards the user in human-computer interactions, recent technological developments have led to a surge in popularity of AI, and its adoption and influence have increased exponentially. The most prominent example of these developments is the launch of ChatGPT in November 2022,[8] which captivated the general public's interest in AI and expanded their utilisation beyond academic and industrial contexts, facilitating their integration into the daily lives of non-experts, as explored and demonstrated by Yue et al. (2023).

Amidst the growing excitement surrounding the potential of AI, with heightened attention from the media, academic circles, and the industry, recent research has begun to explore its transformative impact across various spheres of life, ranging from education (Kasneci et al. 2023) to software development (White et al. 2023), or translation (Jiao et al. 2023; Lyu, Xu, and Wang 2023; Castilho et al. 2023; Briva-Iglesias, Camargo, and Dogru 2024), among other professional domains. Concurrently, numerous concerns have been voiced regarding the possible negative consequences of these emerging technologies in the workplace, including job displacement or disruption due to increased automation (Eloundou et al. 2023), the hazards associated with adhering to AI-generated guidance (Oviedo-Trespalacios et al. 2023), and the ethical (Zhuo et al. 2023) and privacy (Sebastian 2023) challenges that are likely to emerge soon, underscoring the need for more stringent regulation and oversight of these technologies (Hacker, Engel, and Mauer 2023).

This brings into question the goal of developing AI technologies and how to adopt them. According to Shneiderman (2022a), the main goal of developing traditional AI technologies has been the creation of an intelligent agent that emulates human behaviour and acts as an autonomous system that automates human tasks. As a consequence, a novel technology design framework has gained a foothold recently: human-centered AI (HCAI), where instead

[8] https://openai.com/blog/chatgpt, last accessed 09/02/2024.

of human replacement, the aim is to produce a powerful tool that augments human capabilities, enhances performance, and empowers users, who are at all instances in supervisory control of such systems (Shneiderman 2020a; 2020b; 2020c; 2022b). A key element in the HCAI framework is that of “augmentation” of the human intellect. Human performance is constrained by cognitive load and augmentation seeks to overcome this limitation (Alicea 2018), to amplify intelligence (Stanney et al. 2015) by deploying technologies related to human perception and cognitive performance. Raisamo et al. (2019) also deal with augmentation as technologies that enhance human productivity or capability, or that somehow add to the human body or brain. This shift, moving from emulation to empowerment, aligns with the concept of Intelligence Amplification (IA), placing humans at the centre of AI technology (Shneiderman 2020a). This reorientation, emphasizing the synergy and collaboration between humans and machines, heralds a new era where AI becomes a partner rather than a substitute. In the language services industry, this human-centered augmented approach to translation has been recently proposed by O’Brien (2023). In this context of exponential technological developments, we consider that it is essential to follow and adopt a human-centered, augmented approach to MT (HCAMT).

Ethical principles play a crucial role in the design and deployment of HCAMT tools and workflows. Therefore, ethical considerations such as reliability, safety and trust become essential in this context (Shneiderman 2020c). In the language services industry, these issues underscore the importance of a conscious decision-making process in designing MT workflows, considering both the values of developers and the impact on translators and end-users (Moorkens 2022). In addition, the recent commentary towards “augmented translation” workflows, which involve combining human capabilities with AI-driven technologies to overcome human limitations (O’Brien 2023), also gain strength in this context of inevitable translator-computer interaction. Therefore, considering these ethical principles is imperative to ensure that the translation tools developed under the umbrella of HCAMT not only enhance productivity but also uphold ethical standards and respect the rights and roles of all stakeholders involved (Moorkens 2022; Briva-Iglesias and O’Brien 2023; O'Brien 2023).

As a conclusion, in the language services industry, HCAMT's role is not just about enhancing technological capabilities but also about transforming the users' perception of technology. By

positioning AI or MT as a tool that augments human skills rather than replaces them, HCAMT fosters a paradigm where technology adoption grows not just through its efficiency but also through its capacity to empower users, whose control must remain paramount (Shneiderman 2020b; O'Brien 2023). In essence, the successful implementation of HCAMT in the language services industry may lead to sustainable, diverse, and ethically sound development in MT systems and other technological tools through a wide variety of users and use-cases (Briva-Iglesias and O'Brien 2023).

CHAPTER 5. RESEARCH RATIONALE AND RESEARCH QUESTIONS

In the previous chapters, we have presented a comprehensive literature review of translation technologies in the language services industry and of interaction with MT. First, state-of-the-art translation technologies have been reviewed and their application in contemporary translation production workflows has been discussed (Chapter 2). Secondly, IPE has been analysed in depth, its historical evolution described, and the reasons why its adoption in today's industry could be interesting have been discussed (Chapter 3). Thirdly, the field of HCI and its most relevant concepts have been presented from a transdisciplinary point of view, with a focus on the concepts of usability and UX and on the language services industry. In addition, studies of HCI factors applied by the Translation Studies and MT communities have also been reviewed, finishing with the introduction of the concept of HCAMT (Chapter 4).

Through this literature review, we have observed that today's language services industry requires, without any doubt, human-computer interaction. Computers and technology are now a vital element in current commercial translation production workflows, and translators are bound to use these digital tools to be competitive in today's language services industry. Nevertheless, in modern translator-computer interactions, we have also observed that the MT community has focused mainly on productivity and quality for increasing automation, neglecting to focus on how users interact with current technological systems. This has had serious repercussions on human factors and, consequently, we have seen an increase in translators' rejection of TPE (Torres-Hostench et al. 2016; Macías 2020), the dehumanisation of the translator and the commodification or uberisation of translation workflows (Fırat 2021), as well as the non-adoption of post-editing as a technology (Cadwell, O'Brien, and Teixeira 2018). These are not the only worrying studies on the direction of current translator-computer interactions; other recent studies reveal an increase in translators' fears of MT or AI (ELIS Research 2023) or the use of algorithms as the only element to manage translators' participation in current commercial translation production workflows (Moorkens 2023), which leaves translators as a simple cog that can be easily replaced in a large machine, if needed (Moorkens 2020). In addition, we have observed that technology adoption has followed a process of human adaptation (Winner 2007; Vallor 2024), that is, technology has been developed first and users have been asked to adapt to this developed technology at a later stage. It is at this stage when we raise the question on whether the process should be

the opposite: first, we understand what users need from technology, and then we develop new technologies to meet these needs and augment users.

Hence, we consider that the translation technology and MT communities should re-visit their technology design, development and adoption focus, and should shift their attention from the translation productivity- and quality-first approach towards a HCAMT approach. This HCAMT approach is built on the basis of the HCAI (Shneiderman 2022b; 2022a) framework. The HCAMT approach and the focus on translation quality and productivity do not have to be mutually exclusive and should be applied together. HCAMT technologies should be developed considering the needs of its users as a central point, but also their productivity and quality if applied to different translation production workflows. MT users vary substantially (Nurminen 2019), and therefore covering them all is out of the scope of this PhD thesis. The overarching aim of this research is to lay the basis of a new methodology to foster the development and adoption of HCAMT tools, systems and workflows, using today's language services industry as a specific use case. By narrowing the scope of the work to a specific use case, this PhD research will explore whether IPE may be a better alternative to TPE by analysing MTUX, translation productivity and translation quality. The remainder of this chapter will elaborate on this rationale and introduce the main research questions to be addressed.

5.1. Operationalising MTUX

From the literature review in the HCI domain, we can easily conclude that users' feedback resulting from the use and the anticipated use of any product should be crucial in the development of any technological tool or system, regardless of whether this feedback takes the form of a usability evaluation, an ergonomic evaluation, or a UX evaluation. The latter form of user feedback assessment, the UX evaluation, is deemed to be the most appropriate to undertake from our standpoint, as it includes all the previous types of evaluation given that "UX" is a broader concept that incorporates usability, ergonomics and hedonomics. All the studies analysed in Section 4.4.1 aimed to study "usability" to some extent by measuring some of the concepts included in the ISO definition (ISO, 2018): effectiveness, efficiency, and/or satisfaction. It is worth highlighting that only some researchers focused their studies on the real users interacting with systems (Etchegoyhen et al. 2018), and most studies on

usability focused on the readers (end-users) of machine translated texts and the acceptability of such machine translated information. Acceptability of a machine translated text is, without a doubt, of relevance for the translation and MT fields. However, when considering acceptability, we are only paying attention to the user reception of a static text, and there is no actual interaction of a user with any type of product or system. We could therefore argue that these studies are more similar to text reception studies (Hall 1980) and deal with how readers understand and interpret a text produced via different modalities (by a human without technological assistance, post-edited or direct raw MT output) than to usability testing. Furthermore, some studies only researched usability vaguely because satisfaction was not included in their projects (for example, Roturier (2006)), and only Doherty and O'Brien (2014) and Castilho (2016) examined the three concepts within the usability paradigm.

This allows us to suggest that the most important element of user-computer interaction has been traditionally forgotten by the MT community: the users. The "satisfaction" term of the ISO's definition of usability has been neglected in most studies reviewed, and there is a whole world beyond if we consider ISO's definition of UX, which establishes UX as a "person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service". If we apply this definition of UX to our use case, we can extract two pieces of information that have utmost importance in this PhD dissertation. On the one hand, "person's perceptions and responses resulting from the use of a product, system or service", that is, what do translators experience when using MT? How do they feel after engaging with such a product or system? Is this a rewarding task, and do they feel that their productivity is being augmented or reduced by such a system? On the other hand, the importance lies not only in the resulting experiences after system use, but also on "person's perceptions and responses resulting from the […] anticipated use of a product", which will undoubtedly affect how the translator faces such an interaction with MT. Do translators like post-editing MT? How does a translator feel before using MT as an aid? Do translators think that post-editing tasks are a threat to their profession? From the literature review in section 4.4.2, we have found that most studies reportedly analysed UX, but only performed some basic Likert-scale questionnaires (Bowker and Ciro 2018; Matusov, Wilken, and Georgakopoulou 2019) or used usability questionnaires and did not consider users' responses or perceptions before, during

and after task completion (Guerberof Arenas, Moorkens, and O'Brien 2021). The only research project found to apply actual UX methods as per the ISO definition was Koponen et al. (2020) and Karakanta et al. (2022) in the audiovisual translation domain, but they only took into consideration users' post-task perceptions or responses, and with substantial modifications to a validated questionnaire.

Therefore, this PhD study pursues a concept of UX that has not been fully engaged with to date by the Translation Studies and the MT communities and intends to raise awareness of the importance of UX in these fields. Building on Roto's (2006) framework of UX, we consider that user-MT interactions are a situated activity that should consider the context where the interaction takes place, the system employed, as well as the experience of the user. Consequently, we propose the concept of **MTUX as a person's perceptions and responses resulting from the use and/or anticipated use of MT.**

This study aims to be a catalyst for a step change within the MT community and beyond to consider MTUX as a crucial concept within MT research. If the MTUX of translators (or any other user of such tools) is not evaluated or considered when designing these technologies, interactions will undoubtedly not offer the best MTUX possible, which should be the ultimate goal of translation technology tool designers and MT system developers in order to establish a strong relationship with users and products. Therefore, MTUX should be an indispensable element in MT studies (to date, neglected), which should without any doubt include HCI methods (Dillinger and Lommel 2004) and a HCAMT approach. In this work, we consider MTUX as a holistic approach to traditional UX, where different elements need to be considered in the interaction of users with MT: pre-task perceptions and post-task perceptions.

User pre-task perceptions should be analysed to see whether they impact user performance (in terms of quality or productivity). In addition, user post-task perceptions are also key to identify pain points in the interaction and see how tools can be improved to offer better user experiences. If we consider these two elements, we will be able to develop better HCAMT tools, systems and workflows. But not only these subjective elements should be studied, objective elements (productivity and quality) should also be considered to triangulate performance data with the final post-task perceptions resulting from the MTUX evaluation. In a very competitive market like the language services industry, time, quality and costs are

key to ensure profitability and business feasibility. However, the users have been neglected to date. This work aims to show that improving productivity and quality is not mutually exclusive from respecting the user. The focus of the translation technology and MT communities should be to develop HCAMT technologies and workflows that lead to sustainable, diverse, and ethically sound technological developments, hand in hand with translation quality and productivity gains.

In such a big industry, it is worth stressing that there are a wide range of users that interact with MT systems. Consequently, their tool and workflow expectations, as well as their goals may differ. This may directly impact their MTUX. Some examples of different MT users include academics (Escartín et al. 2017) or legal practitioners (Nurminen 2019), among many others. This may result in different types of MTUX, and MT system developers may even be able to adapt their tools to the different users to improve their specific MTUX (O'Brien and Conlan 2018). For instance, in the language services industry, even if translation may be considered the central activity, there are specific tools for the multiple domains in which translation takes place, like subtitling (Subtitle Workshop, AegiSub), localisation (Passolo) or more general translation (Trados Studio). Therefore, the goal of studying MTUX aims to put every user in the centre of the interaction and devise how they feel when interacting with MT tools, pursuing HCAMT. This will allow for detecting points of friction that can be easily addressed in the MT tool development stage and solve problems that may translate into a better interaction, leading to increased satisfaction and pleasure of the user, more productivity and efficiency, a resulting product of higher quality, and higher adoption of technology.

However, as there are many different MT users and analysing their different MTUXs is out of the scope of this PhD, here we will only focus on professional translators in a very specific type of translator-MT interaction: TPE and IPE tasks. The goal is to establish a methodology for developing HCAMT tools through the measurement of MTUX and the comparison with translator performance data (translation quality and productivity), so that researchers interested in other types of users can replicate this methodology for fostering HCAMT technologies and workflows for their intended users.

5.2. Research questions and hypotheses

This research work is therefore driven by the following overarching research question (RQ):

- RQ. Is IPE a better alternative to TPE in terms of machine translation user experience (MTUX), translation productivity, and translation quality?

The overarching RQ is affected by three factors: MTUX, translation productivity and translation quality. These factors guide the experiments performed and therefore the research questions are separated in this section by these three factors.

5.2.1. Factor one: Machine Translation User Experience (MTUX)

The MTUX factor in this research work will be assessed through different RQs, which are as follows:

- RQ1. Is MTUX statistically significantly impacted by MTPE modality (TPE or IPE) and does this vary with increased experience?

With this twofold research question, we will be able to obtain the following information:

(i) Whether either of the MTPE modalities has a statistically significant effect on MTUX. This would mean that translators using a particular MTPE modality would experience a better level of translator-MT interaction.

(ii) Whether this effect on MTUX evolves when translators have more experience in the MTPE modality. As IPE is a more novel modality and translators are inexperienced in it, but many have experience in TPE, we cannot compare both modalities because it would be an unfair comparison to IPE. Therefore, we will conduct a longitudinal study to assess whether there is a difference in experience in the MTUX results. If point (i) above did not report a statistically significant difference from the beginning, it may be the case that, after translators' experience with IPE increased, initial MTUX results may change.

Section 6.2 in Chapter 6 discusses the methodology used to address RQ1. Section 7.1 in Chapter 7 presents the results that answer RQ1.

5.2.2. Factor two: translation productivity

The translation productivity factor of this research work will be assessed through the following RQ:

- RQ2. Is translation productivity statistically significantly impacted by MTPE modality (TPE or IPE) and does this vary with increased experience?

With this twofold research question, we will be able to obtain the following information:

(i) Whether either of the MTPE modalities has a statistically significant effect on translation productivity. This would mean that translators using a specific MTPE modality would be able to translate faster.

(ii) Whether this effect on translation productivity evolves when translators have more experience in one or other of the MTPE modalities. By conducting a longitudinal study, we will account for this experience difference in MTPE modality and study whether it influences translation productivity.

Section 6.2 (Chapter 6) discusses the methodology to address RQ2. Section 7.2 (Chapter 7) presents the results that answer RQ2.

5.2.3. Factor three: translation quality

The translation quality factor of this research work will be assessed through two RQs, which are as follows:

- RQ3. Is fluency statistically significantly impacted by MTPE modality (TPE or IPE) and does this vary with increased experience?
- RQ4. Is adequacy statistically significantly impacted by MTPE modality (TPE or IPE) and does this vary with increased experience?

With these two RQs, we will be able to obtain the following information:

(i) Whether either of the MTPE modalities has a statistically significant effect on translation quality (measured through fluency and adequacy). This would mean that translators using a specific MTPE modality would be able to produce translations with higher quality (in terms of fluency or adequacy).

(ii) Whether this effect on translation quality (fluency or adequacy) evolves when translators have more experience in one of the MTPE modalities. By conducting a longitudinal study, we will account for this experience difference in MTPE modality and study whether it influences translation quality (in terms of fluency or adequacy).

Section 6.2 discusses the methodology to address RQ3 and RQ4. Section 7.3 presents the results that answer RQ3 and Section 7.4 those of RQ4.

5.2.4. Further exploration of Machine Translation User Experience

In addition, as this is the first research work analysing MTUX, it is interesting to explore all the implications of the data collected via the questionnaires. We will obtain two types of MTUX data: pre-task perceptions of MTPE and MTUX scores. Therefore, we will study whether these two types of MTUX data have any relationship with translation quality (fluency and adequacy) or translation productivity. Thus, a further exploration of MTUX will be assessed through various RQs, which are as follows:

- RQ5. Do pre-task perceptions of MTPE correlate with fluency?
- RQ6. Do pre-task perceptions of MTPE correlate with adequacy?
- RQ7. Do pre-task perceptions of MTPE correlate with translation productivity?

With these research questions, we will be able to obtain the following information:

(i) Whether pre-task perceptions have any correlation with translation performance measures, namely translation quality (fluency or adequacy) or translation productivity. Our hypothesis is that translators with negative pre-task perceptions of MTPE may record lower quality translations and reduced productivity if compared with translators with positive pre-task perceptions of MTPE.

Section 6.2 discusses the methodology to address RQ5, RQ6, and RQ7. Section 7.5 presents the results that answer RQ5, RQ6 and RQ7.

- RQ8. Does MTUX correlate with fluency?
- RQ9. Does MTUX correlate with adequacy?
- RQ10. Does MTUX correlate with translation productivity?

With these research questions, we will be able to obtain the following information:

(i) Whether MTUX scores have any correlation with translation performance measures, namely translation quality (fluency or adequacy) or translation productivity. Our hypothesis is that translators with higher MTUX scores may record higher quality translations and increased productivity if compared with translators with lower MTUX scores.

Section 6.2 discusses the methodology to address RQ8, RQ9, and RQ10. Section 7.6 presents the results that answer RQ8, RQ9 and RQ10.

The next chapter will introduce the methodology used to gather and analyse data in order to respond to each of these questions.

CHAPTER 6. METHODOLOGY

This chapter addresses the methodology used for answering the research questions governing this PhD thesis. As discussed in Chapter 5, a HCAMT tool, system or workflow should consider both subjective and objective elements of the user-MT interaction. Therefore, different but complementary experiments were conducted to assess all these elements. Section 6.1 describes a pilot experiment performed to identify potential methodological issues before the main study. Then, Section 6.2 describes the methodology employed for the main longitudinal study after applying the lessons learned from the pilot experiment. As the experiments involved human participants, application was made to the DCU Faculty Research Ethics committee for approval, and the approval letter is included in Appendix A. Finally, Section 6.3 explains the selection of a mixed-methods approach that will inform the analysis and interpretation of the results of the main longitudinal study.

6.1. Pilot experiment

The aims of the pilot experiment were twofold: (i) to identify the best UX questionnaire for measuring MTUX, and (ii) to test the methodological design of the main longitudinal study. The funding for the pilot experiment was granted through the European Association for Machine Translation (EAMT) Sponsorship of Activities – Students Edition of 2021. The results of this experiment were presented in two venues: a first paper covering the selection of the UX questionnaire for MTUX measurement at the EAMT2023 Conference (see Briva-Iglesias and O'Brien 2023); and a second paper covering the pilot user study in the journal *Translation, Cognition and Behavior* (see Briva-Iglesias, O'Brien, and Cowan 2023).

6.1.1. Participants

Involving professional translators, and not crowd workers, is crucial for obtaining valid and expert feedback from the interaction with MT (Läubli et al. 2020). We therefore hired 15 English-Spanish professional translators on a first-come, first-served basis through ProZ and paid them €20 hourly. The hiring requirements were to have between one and five years of full-time professional translation experience,[9] to have Spanish as an L1 and have professional

[9] We wanted to work with junior translators, who have been suggested to be more inclined to adopt newer technological tools and workflows (Weinberg 2004) and also because this is the generation of translators who will have to deal more with future technological advances in translation.

experience in the legal domain because this was the domain of the content to be translated. We chose legal translation because it is one of the main domains in the language services industry (ELIS 2022) and the legal language has intrinsic linguistic complexities that difficult the translation process (Borja 2000; Briva-Iglesias 2021). We also controlled the experience level of participants to minimise variable levels of translation experience. Hiring one translator with two years of experience and another with 25 years of experience may have an impact in the results. In addition, we hired three senior reviewers with more than five years of industry experience in reviewing by following the same hiring methodology. These reviewers elaborated a set of guidelines for translation quality assessment through different iterations, and one of them evaluated every translation by following the guidelines through an adequacy and fluency assessment.

6.1.2. Content

Complex English legal contracts were the texts chosen for our controlled pilot experiment. Each translator worked with four different texts, two under each condition (two texts in TPE and two texts in IPE), and we randomly divided the assignments, ensuring that the combination of text and modality were counterbalanced across the experiment. Also, to avoid problems associated with text difficulty that occurred in previous studies on IPE (see Sanchis-Trilles et al. 2014), all texts were controlled for length and complexity with the Flesch-Kinkaid index and the type token ratio (TTR).

6.1.3. IPE workbench

The IPE workbench used was Lilt, where the participants were assigned both TPE and IPE tasks. Though Lilt is a proprietary tool, the advantages of using Lilt outnumbered any other possible open-source IMT workbenches for conducting IPE tasks. In the first place, to have valid results in line with today's industry-standard tools, we needed a tool with good quality MT output for the language pair under consideration. Lilt offers high-quality MT output from English into Spanish (the raw MT output of the texts of the pilot study received an average Adequacy score of 3.4/4 and an average Fluency score of 3.65/4, according to the three reviewers), and we could turn on or off the interactive translation completion proposals for recreating the TPE or IPE modalities.

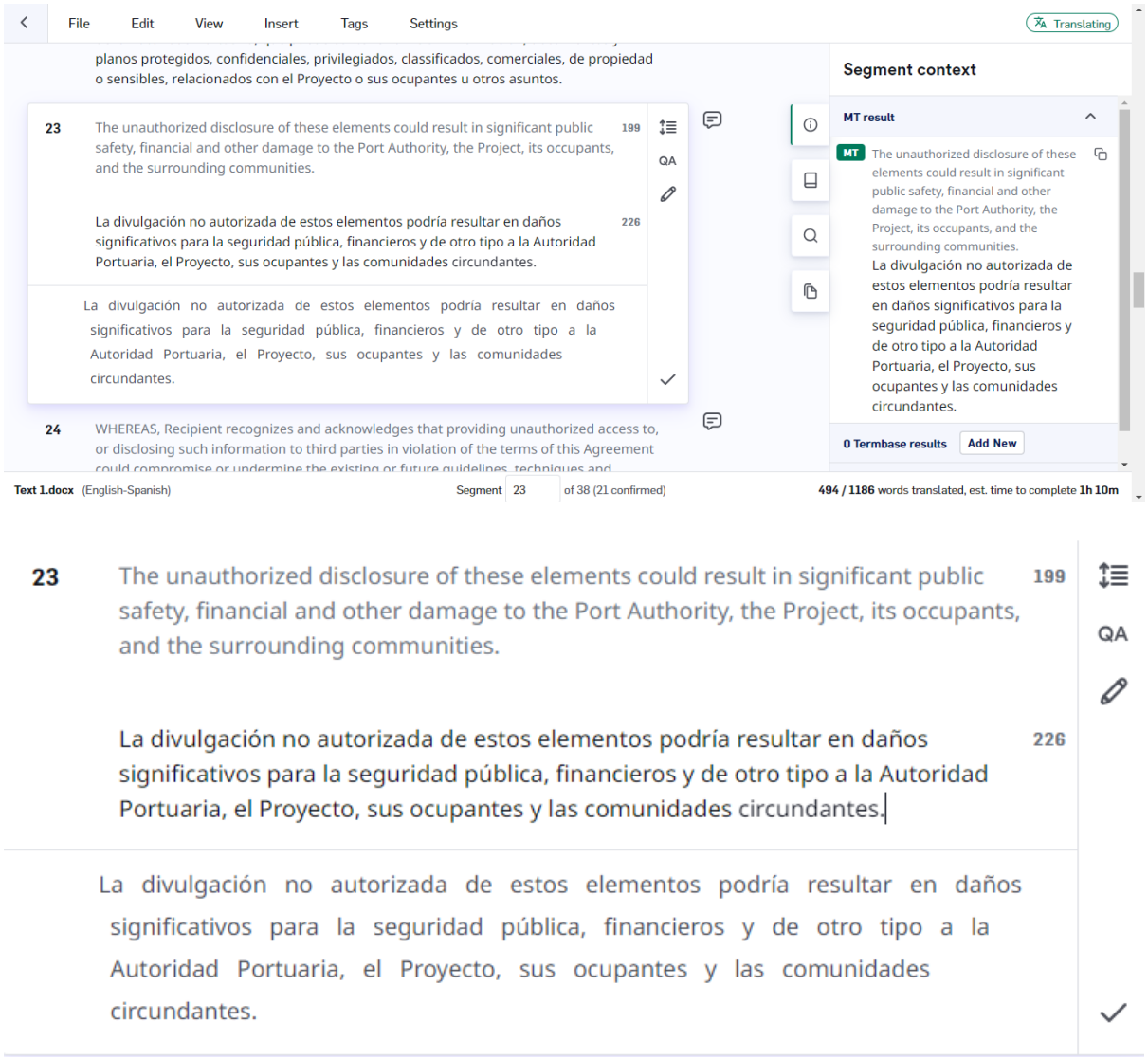

Figure 6.1. Graphic user interface of Lilt in the TPE modality

Figure 6.1 shows the graphic user interface of Lilt in the TPE modality with one of the sample texts used. In the screenshot, the translator was editing segment 23, and the MT completion proposal was already fully propagated in the target segment. Therefore, the translator had to conduct a TPE task by amending the static, adaptive MT output.

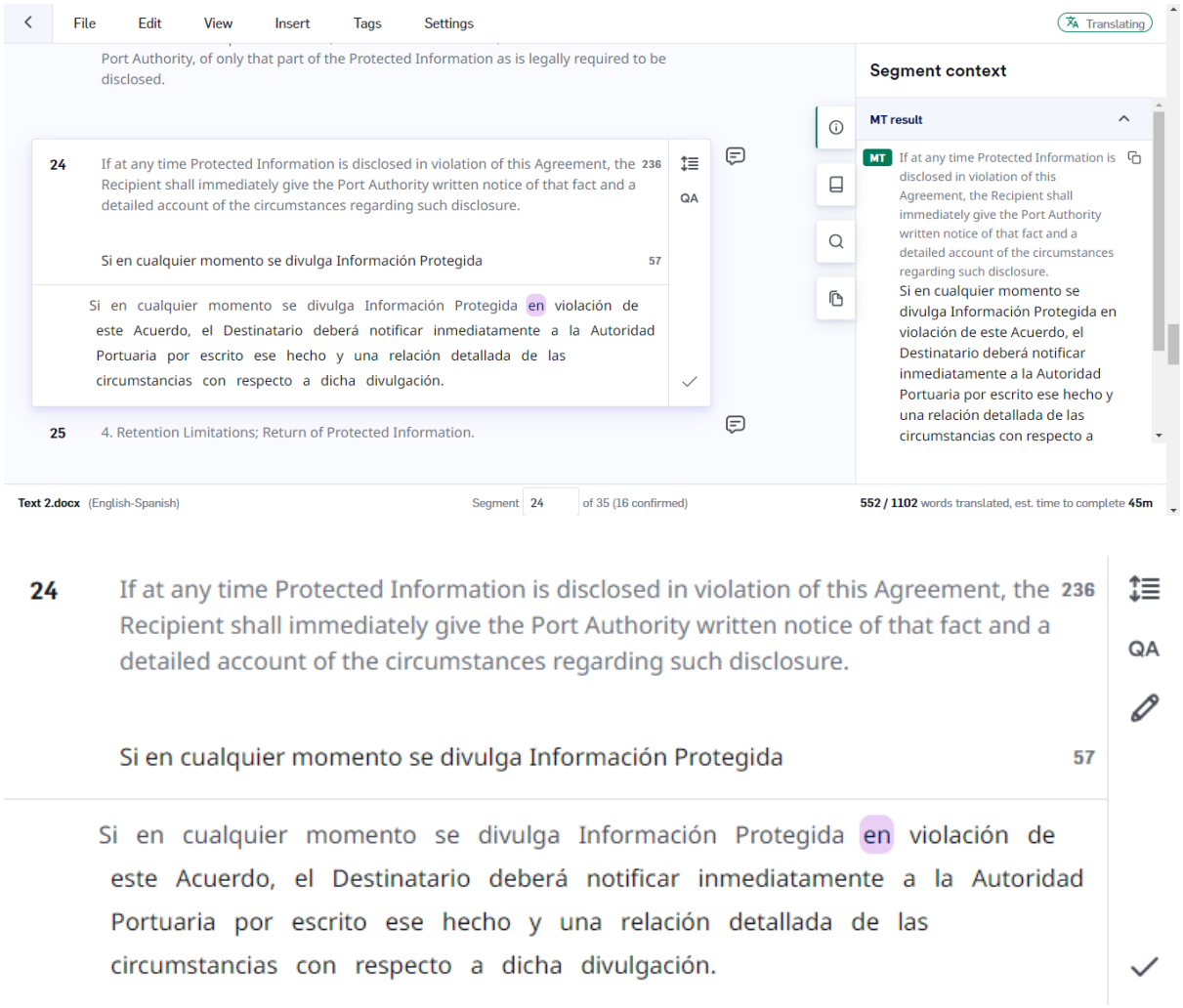

Figure 6.2. Graphic user interface of Lilt in the IPE modality

In contrast, Figure 6.2 shows the graphic user interface of Lilt in the IPE modality with a different sample text. In this screenshot, the translator was editing segment 24 and could see the MT completion proposal, which they could accept partially or completely with specific hotkeys. The word “en” highlighted in purple was the next completion proposal that the system was suggesting at a word-level, which the translator could accept with a hotkey. Should the translator think that this word was not appropriate, and start writing an alternative, the MT completion proposal would change in real time by considering the addition of the translator. Thus, the translator had to conduct an IPE task by amending the interactive, adaptive MT output.

Lilt was contacted to collaborate with the experiments, and they provided an academic license of the platform and the system to conduct the study. We ensured that the workflow

was controlled to eliminate any potential compounding effect when creating the different translation tasks. Before creating any translation task, the following tasks were controlled:

a. The platform MT system was reset. According to the terminology of the platform, we created a new "Custom model" for the English to Spanish combination, so that we could ensure that there was no difference in the MT between the first and the last translator participating in the experiments.
b. A new translation memory was created. According to the terminology of the platform, we created a new "Data source".

This project configuration and strict methodology allowed us to ensure that every translator was shown the same MT proposals, both in the TPE and the IPE methodologies, and that every translator worked under the same conditions. In addition, by using Lilt's academic license, we did not have to train an MT system for the study, reducing costs and emissions (Zhong et al. 2023). Since we did this research from a UX perspective, no CAT tool was identified that was both open-source, interactive, and met the basic requirements for a real-life, professional translation tool, which included an easy-to-use GUI.

6.1.4. Design of the controlled user study

We performed a user study, using a within participants' design. First, translators completed a pre-task questionnaire, where we collected their pre-task perceptions of MTPE and experience with translation technologies.

Then, translators were asked to conduct the post-editing tasks from their home, as if they were doing their day-to-day work as freelance, professional translators. Research shows that real-life scenarios are the most appropriate way to collect reliable data, and that lab interactions may impact the results (Alabau et al. 2013; Ehrensberger-Dow 2014; Ehrensberger-Dow and Heeb 2016). To ensure that these post-editing tasks were conducted as requested and required by the controlled user study, translators were asked to connect to an end-to-end encrypted computer from Dublin City University through AnyDesk, a remote open-source software application. Thus, translators worked in the encrypted Dublin City University computer, even if they were at their home. We screen recorded the interaction of the translators with both post-editing modalities to ensure that they were working the

allocated time, and to measure the productivity (in words per hour). Even though screen recording could have been used to further analyse the keystrokes, the fact that Lilt is a proprietary tool hampered this empirical analysis, unlike open-source tools like PET (Aziz, Castilho, and Specia 2012) and Translog (Carl 2012). In addition, even if it would have been interesting to gather such additional data, it is worth noting that keystroke analysis was beyond the scope of the research questions governing this PhD thesis.

When translators connected to the encrypted computer, they had 20 minutes to test Lilt, both in TPE and IPE conditions. Guidelines explaining Lilt's hotkeys for using the IPE features were provided. It is worth stressing that none of the participants had experience with IPE, but all had experience with TPE. This participant profile represents the reality in the high-tech translation industry at the moment, where many translators have experience with TPE, but very few have thus far experienced IPE. When carrying out our analysis, we were cognisant of the fact that there were deficits in IPE experience and took this into account when interpreting results. Thus, translators completed four tasks with both systems, divided across two consecutive days, to see whether increased experience had any effect on MTUX, translation quality or productivity in either of the different MTPE modalities. The order effect of the TPE and IPE tasks were changed for every translator and every translation session, so that the order had no effect on the final experience. After performing each of the MTPE tasks, translators completed a post-task questionnaire that included two UX questionnaires to measure MTUX: the User Experience Questionnaire (UEQ) (Laugwitz, Held, and Schrepp 2008) and AttrakDiff (Hassenzahl, Burmester, and Koller 2003). To triangulate the results obtained from the subjective MTUX evaluation, translation time was tracked within Lilt and the quality of the resulting texts was assessed by one of the expert reviewers.

6.1.5. Results and lessons learned

When comparing both UX questionnaires, we observed that the UEQ was more suitable for measuring MTUX than AttrakDiff. This is because the UEQ focuses on both Hedonic Quality (HQ) and Pragmatic Quality (PQ) factors, while in AttrakDiff the HQ factor is given more attention (see Section 4.3.2). Thus, it was decided that hereafter, we would use UEQ to measure MTUX as per the conclusions shared in Briva-Iglesias and O'Brien (2023).

In terms of MTUX scores, translators indicated that their MTUX during IPE tasks was higher than during TPE tasks. This difference was statistically significant. Consequently, these results

would suggest that IPE presented a workflow in which the user felt more comfortable, and translators enjoyed the translator-MT interaction more. However, with only two interactions, this difference may have been due to the novelty of the IPE system. This reinforces the need of conducting a longitudinal study to analyse whether this difference continues to be present or if, as translators' experience of IPE increases, their MTUX in this MTPE modality decreases.

In terms of productivity, there was no statistically significant difference between working with TPE or IPE. The average productivity was slightly higher in the IPE modality, but it is important to note that there were only two interactions. It was planned that the longitudinal study would allow us to see, with increased experience in IPE, the effect that MTPE modality has on productivity changes.

Regarding quality, we also did not obtain a statistically significant difference in the results. This implies that the MTPE modality had no effect on the quality of the translations. There was also no statistically significant correlation between MTUX scores and translation performance measures (translation quality or translation productivity).

Thus, in summary, the pilot experiment allowed us to observe that the methodological design was adequate and sound, and it suggested that IPE provided a statistically superior MTUX to TPE. However, with only two interactions, we did not have enough data to fairly compare TPE (all translators had experience in this MTPE modality) with IPE (none of the translators had used IPE before). Hence, the need to conduct a longitudinal study to minimise the impact of varying experience levels in these MTPE modalities, as well as to obtain a larger amount of data to strengthen the statistical analyses conducted.

6.2. Main longitudinal study

In this section, the methodology of the main longitudinal study is explained in detail. The main study adopts a mixed-methods approach (Saldanha and O'Brien 2013), integrating both qualitative and quantitative methodologies to enhance the robustness and depth of our findings through data triangulation (Alves 2003; Mellinger and Hanson 2016) (further details in Section 6.3). As the results of the pilot experiment highlighted the importance of increasing the number of interactions with the MTPE modalities to collect a larger data set to see if the pilot experiment results were replicated over a longer period of time, in the main study we

implement a two-week longitudinal study, offering significant advantages over traditional cross-sectional studies (Diggle et al. 2002). This longitudinal aspect allows us to observe changes in user behaviour and attitudes over time, providing insights into the dynamics of HCI in the context of MT. Such a design not only facilitates a deeper understanding of immediate user responses but also sheds light on the evolution of these interactions, enabling a more detailed and informed analysis of the factors influencing user engagement with MT tools, systems and workflows (Caruana et al. 2015).

6.2.1. Participants

6.2.1.1. Translators

Longitudinal studies have high costs because data must be collected from participants on a recurring basis to gather information on the evolution of the study variables. Taking this into account, we had an available budget to work with 11 translators over two consecutive weeks by paying them an hourly rate of €20. Thus, we first contacted the participants from the pilot experiment and offered them the opportunity to participate in the main longitudinal study. Six of the participants from the pilot experiment confirmed their interest in participating. The only experience these six translators had with IPE were the two interactions from the pilot study, which took place one year and a half before the main study. As a consequence, we consider that their IPE experience was very limited. Then, we hired five additional professional translators on a first-come, first-served basis through ProZ and X (formerly Twitter) (see the recruitment ad for translators of the main longitudinal study in Appendix B). We ensured that every translator:

- Had between one to five years of full-time professional translation experience (for the reasons explained earlier).
- Had Spanish as an L1.
- And had professional experience in the legal domain, as this was the domain of the content to be translated.

The 11 participants were professional translators with experience in TPE tasks. Their professional experience as full-time translators ranged from 12 to 48 months (*N = 11, M = 29; SD = 12*). However, in terms of experience in MTPE tasks, their experience ranged from one to 24 months of full-time professional experience (*N = 11; M = 10; SD = 8*).

6.2.1.2. Reviewers

In addition, like in the pilot experiment, we hired three senior reviewers. We published a recruitment ad on ProZ and X (formerly Twitter) (see the recruitment ad for reviewers of the main longitudinal study in Appendix B) and hired three reviewers with more than five years of experience by following once again the first-come, first-served methodology. Then, we used the annotation guidelines compiled in the pilot experiment as a starting point for homogenizing the quality evaluation criteria with these three reviewers, but only escalated and assessed the bulk of the texts with one expert reviewer (more information below in Section 6.3.5.3) after obtaining a solid inter-annotator agreement (IAA).

6.2.2. Design of the controlled, main longitudinal study

The 11 translators were asked to translate for two consecutive weeks, from Monday to Friday. This involved 10 days of interaction (for a more visual overview of the main longitudinal study, see Figure 6.3).

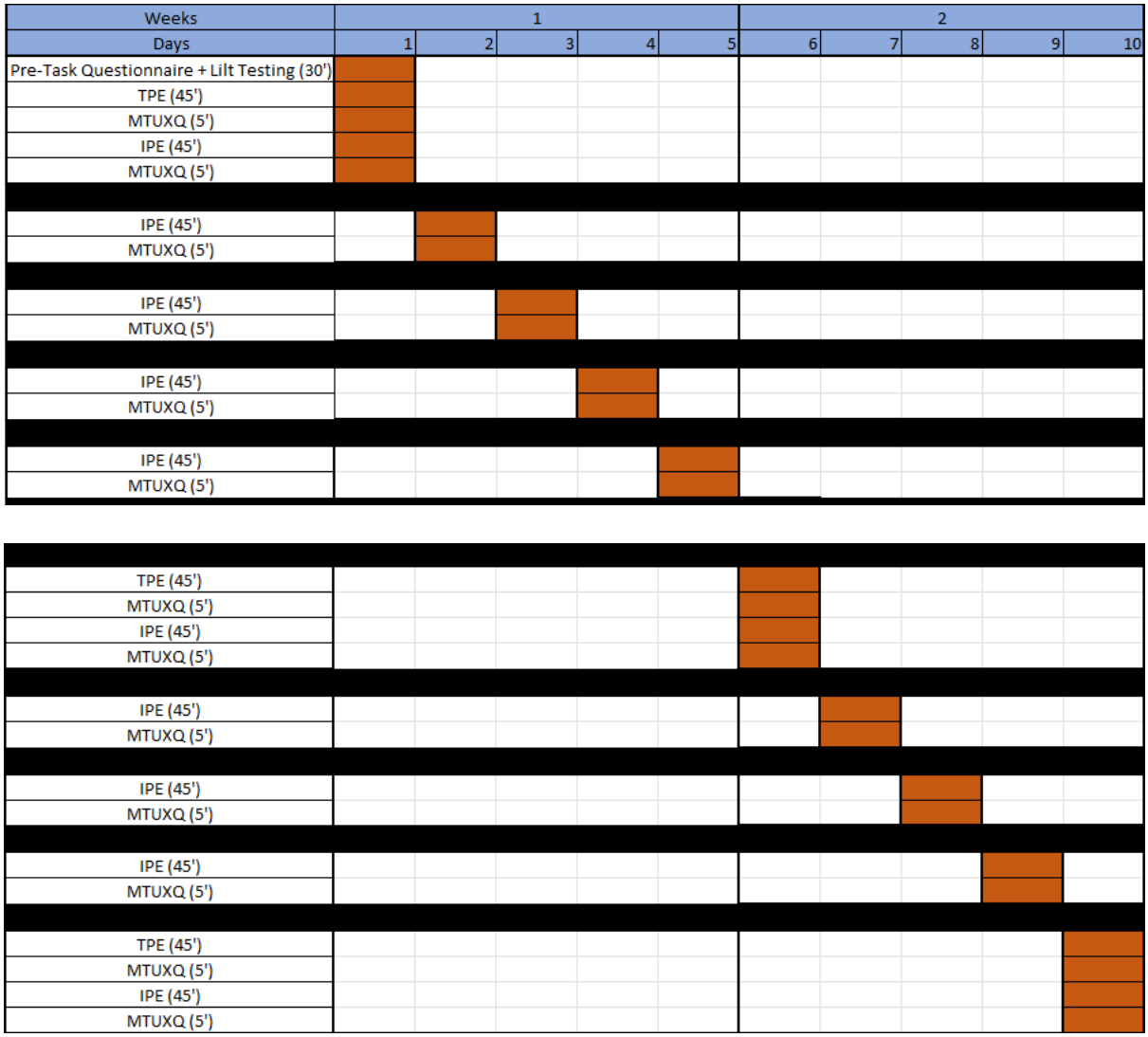

Figure 6.3. Design of the controlled, main longitudinal study

Before starting any translation task, translators had 5 minutes to complete a pre-task questionnaire (see Appendix C). Here, we collected data on their past experiences and pre-task perceptions of MT and MTPE. Then, as in the pilot study, translators connected to an end-to-end encrypted computer from Dublin City University through AnyDesk and had 25 minutes to read instructions on how to use Lilt and to understand the hotkeys for the IPE modality. During this warm-up session of 25 minutes, translators had time to work on a sample project, so they could get acquainted with the tool and the interactive, adaptive features.

After these introductory steps, the two-week longitudinal study started. Every week was structured in the following way:

- On the first day (Monday), translators interacted for 45 minutes with TPE and had 5 minutes to complete a MTUX questionnaire to collect their experiences after the interaction with TPE (see Appendix D). Then, translators interacted for 45 minutes with IPE and had 5 additional minutes to complete another MTUX questionnaire about the IPE interaction. The MTPE modality of the starting task was randomised to avoid any compounding order effect.
- In the remaining days of the week (from Tuesday to Friday), translators only interacted 45 minutes with IPE and completed a MTUX questionnaire after each interaction.
- The only exception to this structure was the last day of the study (on Friday of the second week), where translators also conducted a double translation session consisting of 45 minutes of TPE plus 5 minutes to complete the MTUX questionnaire, and 45 minutes of IPE and 5 other minutes of MTUX questionnaire. Again, the MTPE modality of the starting task was randomised.

Thus, we ended the longitudinal study with 10 interactions of IPE and 3 interactions of TPE for every translator. This longitudinal study design allowed us to directly compare the TPE and IPE interaction sessions in days 1, 6 and 10. Hereafter, these interaction sessions where we have data for TPE and IPE will be named “evaluation sessions”. Then, the 10 interactions with IPE allowed us to analyse the change in IPE over time in finer granularity, as well as the effect of increased IPE experience on the different measures (MTUX, translation quality and productivity). Hereafter, these 10 interaction sessions of IPE will be named “learning sessions”.

The order of the TPE and IPE tasks was randomised for every translator and every translation session, so that it reduced any potential order effect on the final experience. Although the ideal situation would have been to do the same number of TPE and IPE interaction sessions, this was impossible due to the limited budget for the study. Thus, the number of IPE interactions was higher (as this was the MTPE modality in which the translators had the least experience, and we wanted to account for that expertise difference), and different TPE sessions were scheduled at the beginning, middle and end of the study to be used as a baseline.

6.2.3. Texts

As in the pilot experiment, complex English legal contracts were the texts chosen for our controlled study. Each translator worked with 13 different texts, under different conditions (10 in TPE and 3 in IPE), and we randomly divided the assignments, ensuring that the combination of text and modality were counterbalanced across the experiment. Again, all texts were controlled for length and complexity with the Flesch-Kincaid index and TTR. TTR is the total number of unique words (types) divided by the total number of words (tokens) in a given segment of language. This indicates text complexity. The higher the number of unique words, the higher the complexity of the text, and the lower the TTR. The contracts used here were therefore highly complex texts from the legal domain, with an average TTR of 0.29. As a comparison, a normal piece of news from a Spanish newspaper has a TTR of 0.433. Table 6.1 summarises information about the texts used in the main longitudinal study.

	No. of Words	Flesch-Kincaid	TTR
Text 1	1189	21.7	0.29
Text 2	1102	26.5	0.28
text 3	1149	25.1	0.28
text 4	1032	26.9	0.33
text 5	1195	24.3	0.32
text 6	1022	27.9	0.3
text 7	1043	24.3	0.26
text 8	1098	26.3	0.3
text 9	1117	20.1	0.29
text 10	1103	27.2	0.29
text 11	1071	33.1	0.29
text 12	1094	27.4	0.3
text 13	1160	26.1	0.3
avg	1106	26	0.29
total words per translator (total no. of words)	14,375 (158,125)		

Table 6.1. Characteristics of the texts of the main longitudinal study

The final word count if we grouped every text was 14,375 words per translator (158,125 words if we include the 11 translators). Nevertheless, some translators did not finish translating the complete text during the 45-minute period, and therefore we ended up having 120,102 translated words. Every translator worked approximately 9.75 hours (107.25 hours in total).

If compared with all the previous IPE studies reviewed in Chapter 3, our study is one of the biggest samples of TPE and IPE research in terms of translating time over a longitudinal period, the number of words translated, and the number of professional translators hired. TT_{1a}'s study only worked with ten translators (four professionals and six students) for 20 minutes.

TT_{1b}'s study counted with nine translators (it is not stated whether they were professionals) for 36 minutes. The ER4 and ER5 of TT_2 hired six professional translators, who worked for 10 days and translated around 20,000 words (Macklovitch 2006). The evaluation of CAITRA involved 10 students who worked with 5,000 words, but some of them were non-native speakers of the languages they were working with (Koehn 2009a). The longitudinal study of CASMACAT involved five professional translators who produced around 24,000 words over two weeks (Alabau et al. 2016). CASCAMAT's third field trial hired seven translators who worked only two days and produced around 9,000 words (Ibid.). Alves and colleagues (2016) worked with 16 professional translators who worked with only 36 segments. Daems and Macken (2019) hired 4 translators who worked with 20 segments, and Torregrosa-Rivero (2018) hired 8 translators who had to translate a maximum of 300 sentences over two hours.

To facilitate the replicability, or encourage the further analysis of this big dataset, the source and translated texts can be found in Zenodo.[10]

6.2.4. IPE workbench

Again, in the main longitudinal study, the IPE workbench used was Lilt, where the participants were assigned both TPE and IPE tasks. The three professional reviewers assessed the MT quality of the 13 different texts, and obtained an average Adequacy score of 3.48/4 and an average Fluency score of 3.71/4, indicating high-quality raw MT output.

Once again, before creating the different translation tasks of the main longitudinal study, the following tasks were controlled:

a. The platform MT system was reset. According to the terminology of the platform, we created a new "Custom model" for the English to Spanish combination, so that we could ensure that there was no difference in the MT between the first and the last translator participating in the experiments.
b. A new translation memory was created. According to the terminology of the platform, we created a new "Data source".

[10] Link to the source and translated texts: https://xl8.link/fulldataset

This project configuration and methodology allowed us to assure that every participant was shown the same MT proposals, both in the TPE and the IPE modalities, and that every participant worked under the same conditions.

6.2.5. Measures

6.2.5.1. Translators' pre-task perceptions

To collect translators' pre-task perceptions of MTPE, we created an online questionnaire to be completed before starting the post-editing task. This included the following questions.

- *Experience in MTPE tasks:* How long have you engaged with MTPE tasks? Give an approximate time of use with months or years and months (e.g., 1 year and 6 months). [These experiences were then normalized to the number of months].
- *Do you like MTPE?:* On a scale of 1-7, where 1 is "Strongly Dislike" and 7 is "Strongly Like", please rate your perception of doing MTPE tasks in professional translation projects.
- *Do you trust MTPE?:* On a scale of 1-7, where 1 is "Not trustworthy at all" and 7 is "Very trustworthy", please rate if you can trust MTPE to help you successfully deliver a professional translation project.
- *MT as a threat:* Please rate how much you agree or disagree with this statement: "Machine Translation is a threat to the sustainability of the translation profession (Score 1 is "Disagree", Score 7 is "Agree").
- *Is MTPE boring?:* Please rate the following statement: "When I am doing MTPE tasks, I find them [SCORE]". (Score 1 is "Boring", Score 7 is "Engaging").

We correlated translators' pre-task perceptions with final translation quality and productivity to examine if there was any relationship between them.

6.2.5.2. Machine Translation User Experience (MTUX)

Translators completed a self-report UX questionnaire after completing each post-editing task, resulting in ten measures of MTUX for IPE, and three measures of MTUX for TPE. As per the work reported in Briva-Iglesias and O'Brien (2023), we used the User Experience

Questionnaire (UEQ; Laugwitz, Held and Schrepp 2008). UEQ is a validated questionnaire that measures UX, commonly used in the field of HCI (Schrepp, Hinderks, and Thomaschewski 2014; Schrepp, Thomaschewski, and Hinderks 2017). UEQ is a 26-item semantic differential scale that assesses the experiences of users. Each adjective pair is scored on a 7-point scale (e.g., Annoying–Enjoyable, Impractical–Practical, Slow–Fast) with items focusing on six factors:

- Attractiveness (6 items, Cronbach alpha = 0.91): The overall impression of the system. Do users like it?
- Perspicuity (4 items, Cronbach alpha = 0.74): Is the system easy to learn and understand?
- Efficiency (4 items, Cronbach alpha = 0.71): Is the system fast and not effort demanding?
- Dependability (4 items, Cronbach alpha = 0.73): Do users feel in control of the interaction? Is the system predictable and secure?
- Stimulation (4 items, Cronbach alpha = 0.85): Is the system exciting and motivating to use?
- Novelty (4 items, Cronbach alpha = 0.90): Is the system innovative and creative?

The display of the 26 items at each point were randomised with positive and negative poles for each item alternated to avoid any confounding order effects or response acquiescence (see Appendix D).

6.2.5.3. Translation productivity

Translation productivity was measured within Lilt while translators were performing the MTPE tasks by recording the number of words translated per hour (WPH).

6.3.5.4. Translation quality

Section 2.4.1.2 makes clear the difficulty of human evaluation of translation quality. The literature suggests that, whenever possible, it is advisable to use several evaluators to reduce the subjectivity of each evaluator (Guerberof-Arenas 2008; Rossi and Carré 2022). Given the impossibility of hiring several people to review 120,102 words each, we used another methodology that is very common in computer science to minimise the annotators' individual bias and obtain more robust results: a homogenisation of evaluation criteria with different

evaluators through different evaluation steps, and then escalation of the evaluation with one expert evaluator after a high IAA has been achieved (Artstein and Poesio 2008).

In other words, we hired three professional reviewers at an early stage of the project and sent them the annotation guidelines developed in the pilot experiment (Briva-Iglesias, O'Brien and Cowan 2023) to assess 50 translated segments with similar complexity to the texts to be translated. We then calculated the IAA and obtained a result of 0.83 (Artstein 2017). Though this result is already good, we held a meeting over Zoom with the three reviewers to go over the inconsistencies and updated the annotation guidelines. We then sent 50 additional segments to be annotated with the updated guidelines. The resulting IAA from the second evaluation with the three reviewers was 0.95. We then annotated all the translations performed by every translator (120,102 words) plus the MT raw output (14,734 words) with one of the three reviewers, which we consider to be the expert reviewer after homogenizing the annotation criteria. The expert reviewer was able to see the complete texts, so the translation quality evaluation was performed by taking the context into consideration (Castilho 2021). To further validate that the evaluations of the expert reviewer still followed consistent criteria, at 50% of the evaluation performed, 250 random segments were selected and annotated by the two other reviewers. The resulting IAA was 0.88, which indicated a strong and robust quality evaluation thanks to the annotation guidelines. The annotated data can be found in Zenodo.[11]

6.3. The mixed-methods approach

One of the most important elements in the research design process is to decide what methods will inform the research questions and how to interpret the results. After careful consideration, we decided that the mixed-methods research approach was the most appropriate way to conduct our study because it combined elements of both quantitative and qualitative research methodologies, providing a more comprehensive analysis of a research problem. Saldanha and O'Brien (2013) and Moorkens (2012) suggested that mixed-methods research allows for providing a better understanding of the research problem than could be obtained from either method alone. Thus, the mixed-methods approach leverages the

[11] Link to Zenodo: https://xl8.link/fulldataset

strengths of both quantitative and qualitative methods to address more complex research questions (Johnson, Onwuegbuzie, and Turner 2007).

According to Creswell and Clark (2007), mixed-methods research has multiple advantages. One of them is that we can conduct a more comprehensive data analysis because we collect a broader array of data, offering a richer, more nuanced understanding of the research problem. In addition, the triangulation of both quantitative and qualitative data also allows for enhancing the credibility and validity of the research findings by corroborating data from different sources and methods (Alves 2003; Johnson, Onwuegbuzie, and Turner 2007).

Yet, we need to take into account that, when speaking of mixed-methods studies, we can classify them based on whether the qualitative or quantitative component is dominant, or whether both are given equal emphasis. Creswell and Clark (2007) suggest that there are several types of mixed-methods designs: First, the explanatory sequential design, in which quantitative data is collected and analysed first, followed by qualitative data to explain or elaborate on the quantitative findings. Second, the exploratory sequential design, where the study begins with qualitative data collection and analysis, which then informs the subsequent quantitative phase. Third, the convergent parallel design, in which both qualitative and quantitative data are collected simultaneously but analysed separately, with the results compared or combined during the interpretation phase. Fourth, the embedded design, where one type of data (qualitative or quantitative) is nested within a larger, primary research design of the other type.

The analysis and discussions of this PhD thesis are governed by a convergent parallel design, as we concurrently collected and analysed both qualitative and quantitative data to cross-validate and corroborate findings (Creswell and Clark 2007; Johnson, Onwuegbuzie, and Turner 2007). Through a process of triangulation (Alves 2003), we did not give priority to any specific method, and both types of data were analysed separately but concurrently. The results from qualitative and quantitative analyses were then compared and integrated during the interpretation phase to provide a more robust understanding of the research problem.

In other words, in the PhD, the qualitative data on MTUX scores collected through questionnaires from professional translators (see Section 6.2.5.2) was analysed alongside quantitative data on translation productivity (6.2.5.3) and translation quality (6.2.5.3). Using

mixed-methods and triangulation enhanced the reliability and validity of the research findings, offering a holistic view of the research problem that leverages the strengths of both qualitative and quantitative approaches (Johnson, Onwuegbuzie, and Turner 2007).

Otherwise, we could see ourselves in a situation in which a specific post-editing modality reported higher MTUX scores (qualitative data), but lower translation productivity and/or quality (quantitative data). This would not suffice to answer our main research question, as our interest lied in knowing whether a higher MTUX could also go hand in hand with higher productivity and comparable quality. This can only be interpreted after the triangulation of the both subjective (qualitative) and objective (quantitative) data commented above through a convergent parallel design process.

This chapter provides a detailed account of the methodology of the main longitudinal study, which was first tested via a pilot experiment. In the next chapter all results will be presented and discussed

CHAPTER 7. RESULTS OF THE MAIN LONGITUDINAL STUDY

Chapter 7 describes the data analyses performed with the data collected during the main longitudinal study. This PhD is governed by the following overarching RQ: Is IPE a better alternative to TPE in terms of machine translation user experience (MTUX), translation productivity, and translation quality? As discussed in Chapter 5, this overarching RQ is divided into different RQs. Therefore, Chapter 7 is divided into different subsections that cover these RQs. Finally, Section 7.7 presents a discussion of the results of all the RQs.

7.1. RQ1. Is MTUX statistically significantly impacted by MTPE modality (TPE or IPE) and does this vary with increased experience?

The main longitudinal study design allowed us to collect 10 MTUX measures for IPE and 3 MTUX measures for TPE. MTUX scores were collected through a 7-point Likert scale, but results were normalized to scores ranging from −3 (very bad experience) to +3 (very good experience). On days 1, 6 and 10, we have data for both TPE and IPE interactions (hereafter, "evaluation sessions"). In addition, we have data for IPE interactions from day 1 to day 10 (hereafter, "learning sessions").

This longitudinal study design allows us to make a direct comparison of the evaluation sessions on days 1, 6 and 10 by comparing the MTUX scores at these exact times. The evaluation sessions also allow for analysing if there is any change in MTUX with the increased experience in either of the MTPE modalities. Then, the "learning sessions" allow us to visualise the evolution of the MTUX scores during the 10 IPE sessions with more granularity, as this is the MTPE modality in which the translators had no experience. Even if the MTUX scores could range from -3 to +3, we are only visualising the 0 to +3 range because there were no negative values.

Thus, we will perform different 2x3 repeated-measures ANOVAs to analyse the effect of the MTPE modality (Levels: TPE and IPE) and interaction session (Levels: Interaction 1, Interaction 6, Interaction 10) on MTUX scores by considering the data of the evaluation sessions. Then, the data of the learning sessions will also be visualised. Section 7.1.1 describes the statistical analyses conducted to analyse the effect on average MTUX scores, while Section 7.1.2 describes the MTUX scores per factor analysed in the UX questionnaire.

7.1.1. Average MTUX scores

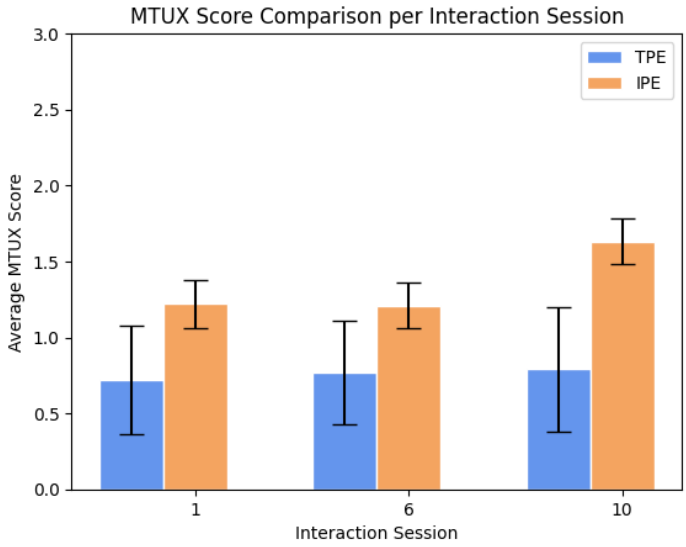

Figure 7.1. MTUX Score Comparison in the Evaluation Sessions (with SD bars)

Figure 7.1 shows the visualization of the average MTUX scores in the different MTPE modalities across the different evaluation sessions analysed, including the error bars displaying the standard deviation. The repeated-measures ANOVA results indicated that there was a statistically significant main effect of MTPE modality on MTUX ($F(1, 10) = 9.91$, $p = 0.01$), with a statistically significantly higher MTUX score when post-editing using an IPE workflow ($M = 1.24$, $SD = 0.16$) compared to when using a TPE workflow ($M = 0.72$, $SD = 0.36$). This means that translators perceived their interactions with MT through the IPE modality improved their user experience, which was substantially better if compared with TPE.

The repeated measures ANOVA also suggests that there was a statistically significant difference in MTUX scores between the different interaction sessions across the MTPE conditions ($F(2, 20) = 4.29$, $p = 0.04$). This suggests that translators reported higher MTUX scores with increased experience with the tool (Interaction 1: TPE $M = 0.72$, $SD = 0.36$; IPE $M = 1.22$, $SD = 0.16$/ Interaction 6: TPE $M = 0.77$, $SD = 0.34$; IPE $M = 1.21$, $SD = 0.15$/ Interaction 10: TPE $M = 0.79$, $SD = 0.15$; IPE $M = 1.63$, $SD = 0.15$). However, the ANOVA results show that there was no statistically significant interaction effect between the interaction session and MTPE modality ($F12, 120 = 2.78$, $p = 0.09$). This means that the effect of interaction session does not vary across the MTPE modalities.

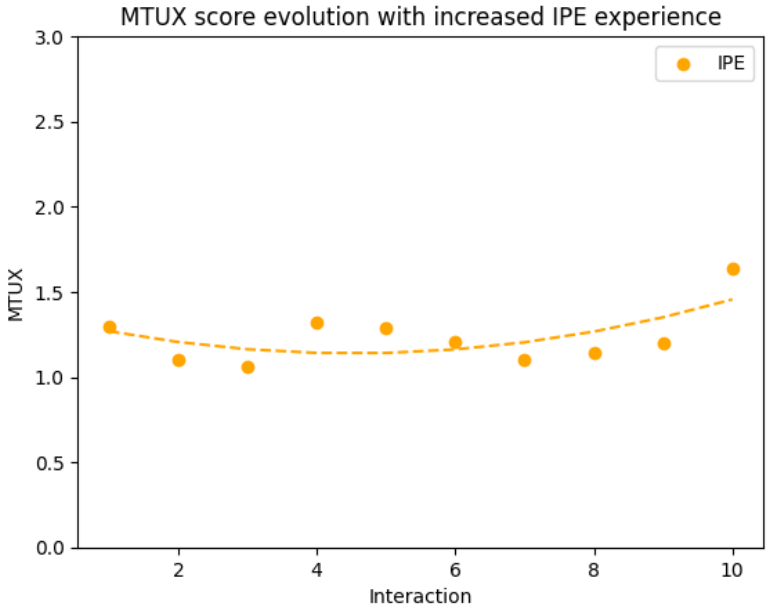

Figure 7.2. MTUX score evolution during the learning sessions

If we change the focus, Figure 7.2 displays the evolution of the MTUX over the 10 interactions of participants with IPE (the learning sessions). MTUX scores initially averaged at 1.3, reflecting a moderate but positive level of user satisfaction with the MTPE modality. Over subsequent interactions, these scores exhibited a general upward trajectory, indicating an enhancement in MTUX as users became more familiar with the IPE features. Notably, a minor decrease in MTUX scores during the 7th interaction could suggest a temporary plateau in the learning curve or point to specific aspects of the tool that might benefit from further optimization. Overall, the positive slope of the trend line underscores the potential of IPE tools to improve participant satisfaction, provided users are afforded adequate time to acclimate to this new MTPE modality.

7.1.2. MTUX scores per factor

This section presents the different 2x3 repeated-measures ANOVAs conducted to analyse the effects of the MTPE modality and the interaction session on MTUX factors.

Attractiveness

In our MTUX questionnaire, Attractiveness was measured through the following opposite adjective pairs: annoying-enjoyable, bad-good, unlikable-pleasing, unpleasant-pleasant, unattractive-attractive, and unfriendly-friendly.

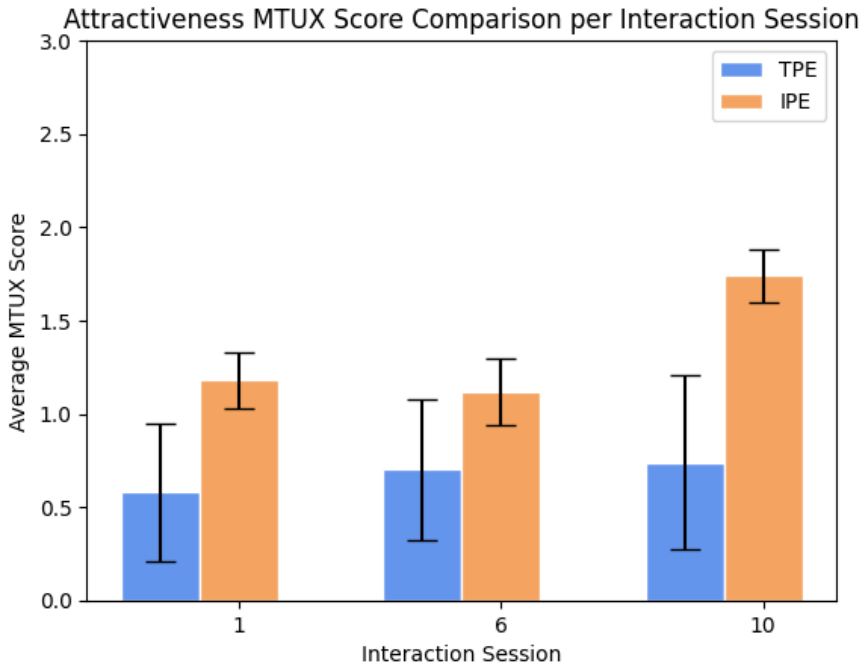

Figure 7.3. Attractiveness MTUX Score Comparison in the Evaluation Sessions (with SD bars)

In our analysis of the impact of TPE and IPE on the Attractiveness factor of MTUX, a 2x3 repeated measures ANOVA was conducted. The main effect of interaction session on Attractiveness revealed a statistically significant effect ($F(2, 20) = 4.12$, $p = .03$). From a practical standpoint, this suggests that the perceived attractiveness does not remain constant; instead, it is contingent upon the specific session, implying that users' perceptions of attractiveness may evolve or change as they become more familiar with the modalities over time.

The main effect of the modality on Attractiveness was also significant ($F(1, 10) = 6.73$, $p = .03$), pointing to inherent differences in attractiveness between the MTPE modalities themselves. In lay terms, there is an MTPE modality that is intrinsically more appealing to users than the other, regardless of the session or interaction frequency. Figure 7.3 displays the Attractiveness scores comparison in the evaluation sessions, and we can therefore see

that IPE (M = 1.17; SD = 0.16) was the MTPE modality best valued by the translators, offering statistically significantly higher attractiveness than TPE (M = 0.67; SD = 0.39).

Conversely, the interaction between interaction session and MTPE modality did not yield a statistically significant effect (F (2, 20) = 2.83, p = .08), indicating that there was no combination of interaction session with MTPE modality that had a unique influence on Attractiveness. This implies that, while overall sessions and modality types do affect Attractiveness, no single session-modality pairing stood out as having a distinct impact.

The results underscore the importance of considering both the type of MTPE modality and the cumulative experience of users across sessions when evaluating Attractiveness. These findings have implications for the iterative design and refinement of such systems, highlighting the dynamic nature of user experience in the context of MTPE modalities.

Perspicuity

In our MTUX questionnaire, Perspicuity was measured through the following opposite adjective pairs: not understandable-understandable, difficult to learn-easy to learn, complicated-easy, confusing-clear.

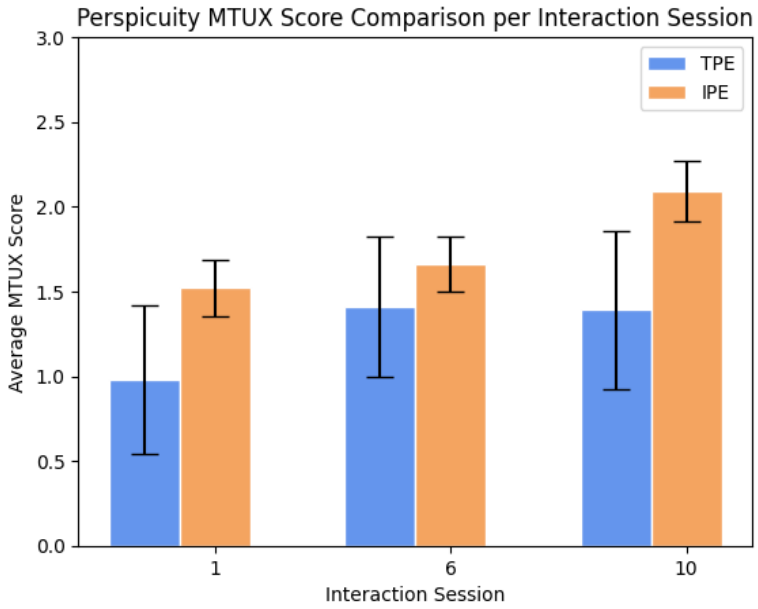

Figure 7.4. Perspicuity MTUX Score Comparison in the Evaluation Sessions (with SD bars)

The statistical analysis of Perspicuity in terms of MTUX, with respect to different MTPE modalities and interaction sessions, was evaluated using a 2x3 repeated measures ANOVA. The results indicated a statistically significant main effect of interaction sessions on Perspicuity ($F(2, 20) = 4.26$, $p = .029$), which suggests that participants' perceptions of the system's ease of use varied across different sessions (Interaction 1: TPE $M = 0.98$, $SD = 0.44$; IPE $M = 1.52$, $SD = 0.17$/ Interaction 6: TPE $M = 1.41$, $SD = 0.41$; IPE $M = 1.66$, $SD = 0.16$/ Interaction 10: TPE $M = 1.39$, $SD = 0.47$; IPE $M = 2.09$, $SD = 0.18$). This implies that as participants engage with the system over time, their understanding and ease of use of the system could change, potentially improving as they become more accustomed to it.

Moreover, the main effect of MTPE modality on Perspicuity was found to be significant as well ($F(1, 10) = 5.34$, $p = .044$), demonstrating that there was one MTPE modality perceived as inherently easier than the other, irrespective of the session. This means that certain modalities are more intuitive or user-friendly, leading to a better immediate understanding and use of the system by participants. Figure 7.4 shows that participants perceived that IPE ($M = 1.59$; $SD = 0.18$) was easier to use than TPE ($M = 1.26$; $SD = 0.41$). This result is interesting because we could expect participants to perceive TPE as an easier workflow since they already had full-time professional experience in this MTPE modality. However, it may be due to the fact that the IPE system highlights the word to be inserted in the translation completion proposals that it has a statistically significantly higher ease of use, according to the participants.

The interaction effect between interaction sessions and modality was not significant ($F(2, 20) = 1.80$, $p = .1908$), indicating that there was no specific session-modality combination that stood out in affecting the system's perspicuity. This finding suggests that while individual sessions and modalities each have an effect on ease of use, there is no compounded effect when they are combined that significantly enhances or detracts from the system's perspicuity.

These findings provide valuable insights into the design and development of MTPE systems, emphasizing the need for modalities that support perspicuity from the first interaction and that can maintain or improve ease of use across multiple sessions. The significance of session effects points to the potential benefits of providing participants with sufficient time and experience with a system to achieve optimal ease of use.

Efficiency

In our MTUX questionnaire, Efficiency was measured through the following opposite adjective pairs: slow-fast, inefficient-efficient, impractical-practical, cluttered-organized.

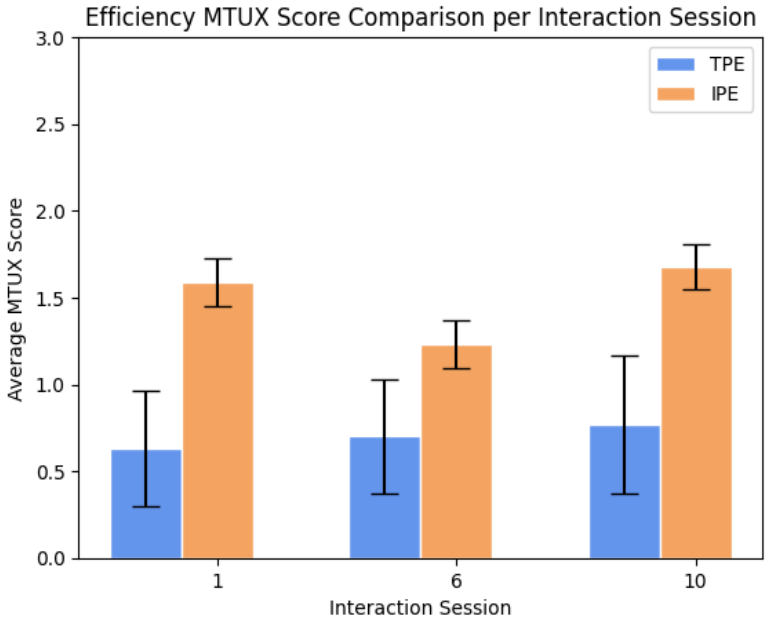

Figure 7.5. Efficiency MTUX Score Comparison in the Evaluation Sessions Sessions (with SD bars)

Our study's 2x3 repeated measures ANOVA analysed the effects of interaction sessions and MTPE modality on the perceived efficiency, which reflects the system's speed and the effort required from users. Statistically, the main effect of interaction sessions was not significant ($F(2, 20) = 1.90$, $p = .18$), indicating that there were no substantial changes in perceived efficiency across different sessions. This suggests that, from a user's perspective, the system's efficiency does not markedly improve or deteriorate with repeated use over time.

However, the main effect of modality on perceived efficiency was statistically significant ($F(1, 10) = 8.94$, $p = .0136$), revealing that one modality was inherently perceived as faster and less effort-demanding than the other. Figure 7.5 compares the Efficiency scores during the evaluation sessions, and we can observe that participants' perceived efficiency of IPE ($M = 1.26$; $SD = 0.14$) is statistically significantly higher than perceived efficiency of TPE ($M = 0.7$; $SD = 0.33$). In practical terms, this means that the design characteristics of IPE makes it more

efficient from the users' standpoint, highlighting the importance of modality design in user satisfaction.

The interaction between interaction sessions and modality was not statistically significant (F (2, 20) = 1.46, p = .2554), suggesting that no particular combination of session and MTPE modality uniquely impacts the perceived efficiency of the system. Essentially, while each factor has its own influence, they do not synergistically affect how users perceive the system's efficiency.

In conclusion, this ANOVA highlights the significance of modality selection for enhancing the efficiency of MTPE systems. The findings emphasize the importance of the MTPE modality features in determining user efficiency perceptions, which has direct implications for the development and optimization of user interfaces in translation tools.

Dependability

In our MTUX questionnaire, Dependability was measured through the following opposite adjective pairs: unpredictable-predictable, obstructive-supportive, not secure-secure, does not meet expectations-meet expectations.

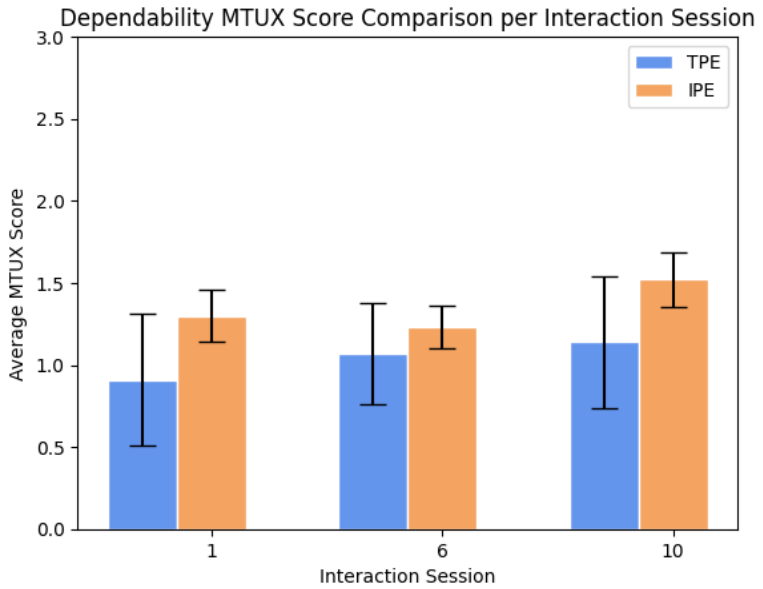

Figure 7.6. Dependability MTUX Score Comparison in the Evaluation Sessions Sessions (with SD bars)

Dependability refers to the users' perception of control over the interaction and the predictability of the system. The repeated measures ANOVA was conducted to evaluate the influence of interaction sessions and MTPE modality on the perceived Dependability. From a statistical viewpoint, the main effect of interaction sessions on dependability was not significant ($F(2, 20) = 1.51$, $p = .34$), suggesting that users' sense of control and predictability did not significantly vary over multiple sessions. This implies that the familiarity gained through repeated use does not appear to alter the perception of dependability in a statistically significant way.

The main effect of MTPE modality was also non-statistically significant ($F(1, 10) = 2.62$, $p = .14$). From a practical perspective, the results displayed in Figure 7.6 point to a potential preference for IPE ($M = 1.29$; $SD = 0.16$) over TPE ($M = 1.04$; $SD = 0.34$) regarding how in control users feel and the predictability of the system, though this preference is not strong enough to be conclusively supported by the data. This may be caused because IPE is a novel methodology, the translation completion proposal suggestions are a new feature for the participants, and getting used to them may be difficult with only 10 interactions.

The interaction effect between interaction sessions and MTPE modality on dependability was also not statistically significant ($F(2, 20) = 0.98$, $p = .39$), which indicates that no specific session-modality combination had a significant effect on users' perceptions of dependability. This means that the variations in how dependable the system felt were not dependent on the particular combinations of session and modality.

Overall, the analysis suggests that while there are some indications of differences in perceived dependability associated with the MTPE modality, these are not substantial enough to be considered statistically significant within the bounds of this study. The perceived dependability of MTPE systems appears to be relatively stable across different interaction sessions, and the interaction between sessions and modality does not significantly affect users' sense of control and predictability. Therefore, the design and selection of modality should cautiously consider these trends towards dependability, even if current results do not show a strong statistical backing. Future research could explore these dynamics further, potentially with larger sample sizes to detect subtler effects, to enhance the development of dependable MTPE systems and interfaces.

Stimulation

In our MTUX questionnaire, Stimulation was measured through the following opposite adjective pairs: inferior-valuable, boring-exciting, not interesting-interesting, demotivating-motivating.

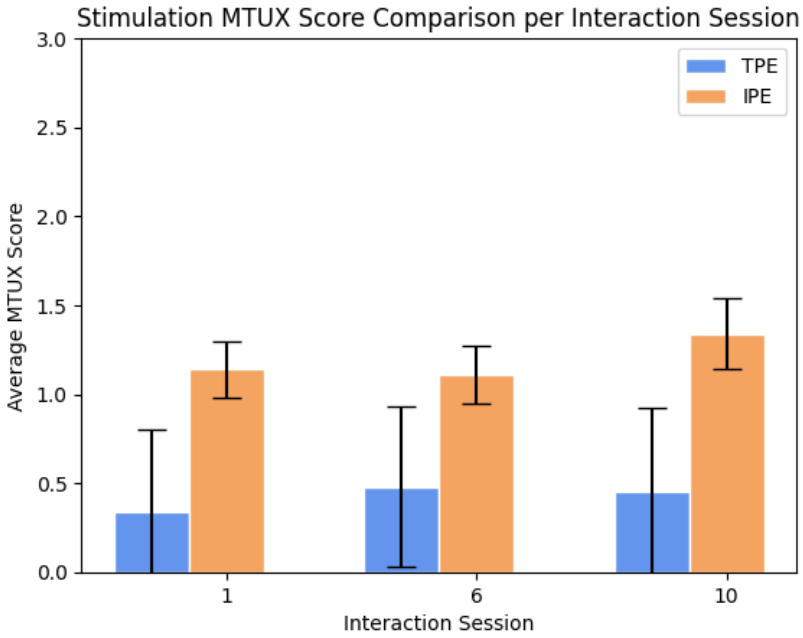

Figure 7.7. Stimulation MTUX Score Comparison in the Evaluation Sessions (with SD bars)

In exploring the effects of interaction sessions and MTPE modality on the perceived Stimulation, our study employed a 2x3 repeated measures ANOVA. The interaction sessions did not have a significant main effect on stimulation ($F(2, 20) = 0.47$, $p = .63$), suggesting that the excitement and motivation provided by the system did not vary significantly across different sessions. This result indicates that users' engagement levels, in terms of stimulation, are consistent over time when interacting with the system.

The main effect of MTPE modality on stimulation was statistically significant ($F(1, 10) = 10.92$, $p = .0079$), demonstrating that IPE ($M = 1.11$; $SD = 0.17$) is perceived as more stimulating than TPE ($M = 0.42$; $SD = 0.43$) (see Figure 7.7). This significant difference in the excitement and motivational appeal of the modalities suggests that some designs inherently elicit a more positive and engaging experience for users.

However, the interaction effect between interaction sessions and modality was not statistically significant (F (2, 20) = 0.35, p = .71), indicating that no specific session-modality combination had a significant impact on how stimulating the system was perceived to be. In essence, while modalities differ in their ability to engage users, these differences do not depend on the number of times a user has interacted with the system.

These findings are particularly relevant to the design of MTPE systems, tools and workflows, underscoring the importance of the MTPE modality in fostering an engaging and motivating user experience. The distinct impact of modality on stimulation highlights the potential for enhancing user engagement by carefully designing the interface and interaction processes of MTPE systems. This could lead to increased user satisfaction and potentially improve the overall performance and wellbeing of people using such systems.

Novelty

In our MTUX questionnaire, Novelty was measured through the following opposite adjective pairs: dull-creative, conventional-inventive, usual-leading edge, and conservative-innovative.

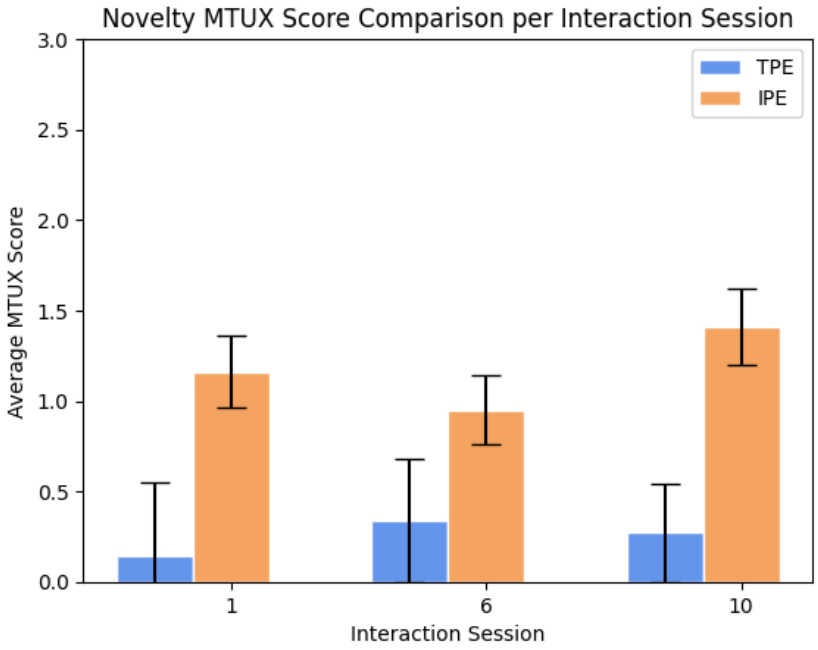

Figure 7.8. Novelty MTUX Score Comparison in the Evaluation Sessions Sessions (with SD bars)

Again, we measured the effects of interaction sessions and different MTPE modalities on the perceived novelty of the system with a 2x3 repeated-measures ANOVA. From a statistical perspective, the interaction sessions did not have a significant main effect on novelty (F (2,

20) = 1.97, p = .16), indicating that users' perceptions of the system's novelty did not significantly change across the different sessions. This suggests that the users' initial impressions of the system's novelty likely remained stable over time.

In terms of MTPE modalities, the main effect was statistically significant (F (1, 10) = 11.44, p = .006), which means that one modality was perceived differently in terms of its innovation and creativity. By looking at Figure 7.8, this result implies that IPE (M = 1.02; SD = 0.19) is seen as more novel than TPE (M = 0.25; SD = 0.44), which could have substantial implications for user engagement and satisfaction. This result was to be expected because of the participants' different experience levels in each of these MTPE modalities.

The interaction between interaction sessions and modality was not statistically significant (F (2, 20) = 1.29, p = .29), suggesting that there wasn't a specific session-modality combination that stood out in influencing the perception of novelty. Essentially, while MTPE modalities themselves can be distinguished by their perceived novelty, this perception is not significantly affected by repeated use or by specific combinations of use over time.

These findings highlight the importance of modality design in MTPE systems to foster a sense of innovation and creativity. The significant main effect of modality suggests that investing in the development of novel features can be crucial for enhancing user experiences and maintaining user interest. This could encourage ongoing user engagement and potentially contribute to the long-term success of MTPE tools.

7.2. RQ2. Is translation productivity statistically significantly impacted by MTPE modality (TPE or IPE) and does this vary with increased experience?

With each TPE or IPE interaction, we tracked translation time with Lilt, and we measured the number of words translated per hour (WPH) in each of the translation sessions, both for TPE and IPE. Again, a 2x3 repeated-measures ANOVA was conducted to assess whether there was a statistically significant effect of MTPE modality (Levels: TPE and IPE) and interaction session (Levels: Interaction 1, Interaction 6 and Interaction 10) on average translation productivity by considering the evaluation sessions.

We found that the main effect of interaction session on translation productivity was not statistically significant (F (2, 20) = 3.38, p = 0.05). This suggests that when we only look at the number of interaction sessions, it does not seem to make a real difference in terms of productivity. This makes sense since participants already have professional experience in TPE, and therefore they may have already attained their maximum productivity speed in this MTPE modality (this is further supported by Figure 7.9, where we can see that productivity in TPE is flat).

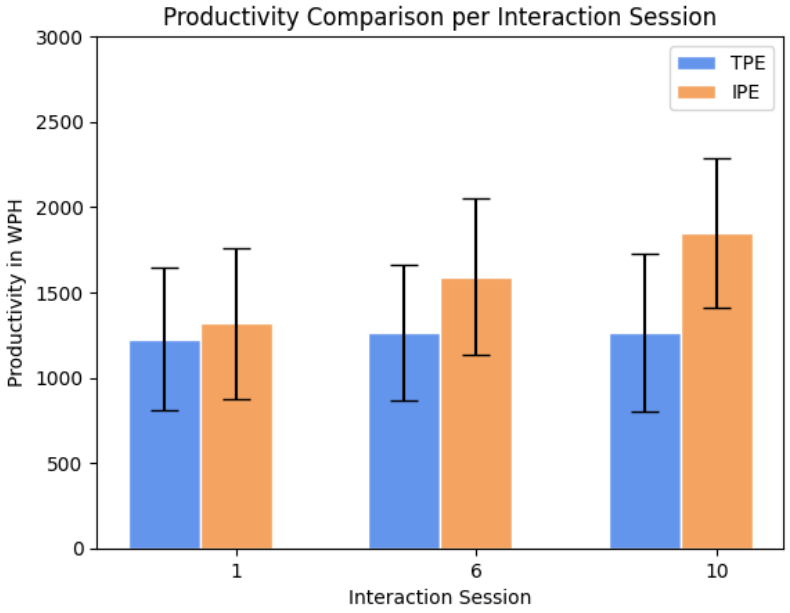

Figure 7.9. Productivity Comparison in the Evaluation Sessions Sessions (with SD bars)

In addition, we found a statistically significant effect of MTPE modality on productivity (F (1, 10) = 19.63, p = 0.001). This means that the MTPE modality used has a clear impact on productivity. By looking at Figure 7.9, we can see that participants worked statistically significantly faster in the IPE modality (M = 1578.85; SD = 935) than in the TPE modality (M = 1359.28; SD = 1004.74).

Lastly, ANOVA results also suggest that there was a statistically significant interaction effect between the two independent variables (F (2, 20) = 16.56, p = 0.0001). This interaction effect suggests that the effect of MTPE varies with experience with improvements in productivity within the IPE condition over time and the TPE performance staying relatively stable across the sessions.

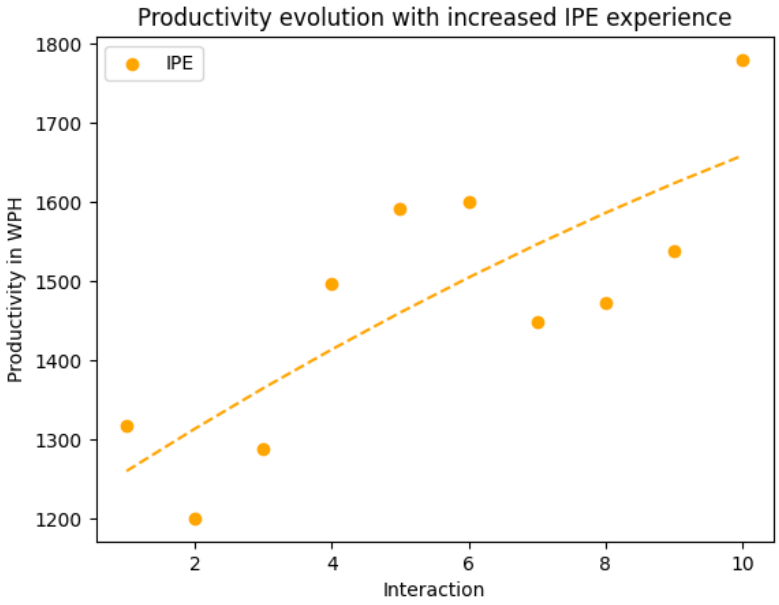

Figure 7.10. Productivity evolution during the learning sessions

When observing the learning sessions (see Figure 7.10 above), over a span of 10 interactions, the participants demonstrated a consistent increase in productivity, as measured in WPH. The graph illustrates this ascent from an initial average productivity of 1317 WPH to nearly 1800 WPH. This pattern underscores a significant enhancement in translation productivity, correlating positively with the participants' growing familiarity with the IPE functionalities. These results highlight the effectiveness of IPE tools in optimizing the translation workflow, indicating that such technologies can substantially elevate productivity levels when users are given the opportunity to adapt to this new MTPE modality.

7.3. RQ3. Is fluency statistically significantly impacted by MTPE modality (TPE or IPE) and does this vary with increased experience?

Again, a two-way repeated measures ANOVA was conducted to assess the impact of interaction session and MTPE modality on the fluency scores of participants. The results indicated a non-statistically significant effect of the interaction session on fluency scores ($F(2, 20) = 1.79$, $p = 0.20$). This suggests that there was no substantial variance in fluency scores that could be attributed to the interaction session. That is, when participants got acquainted with the tool, their fluency score did not improve (see Figure 7.11).

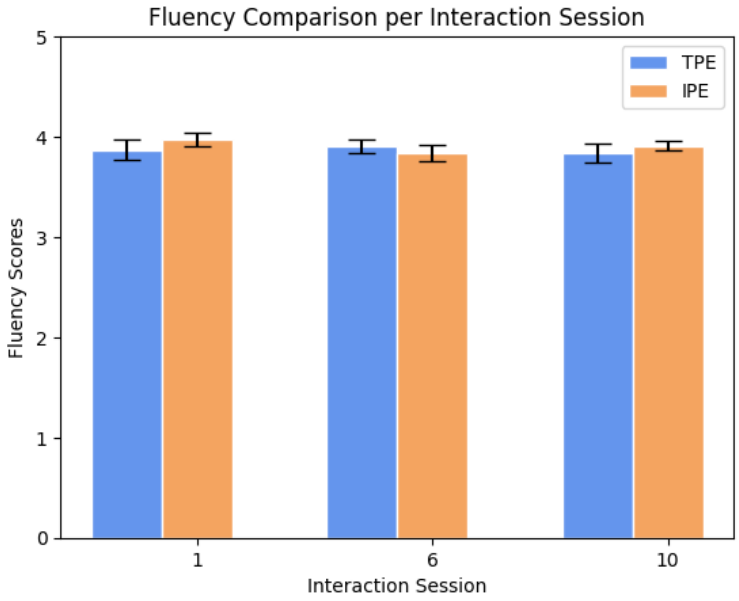

Figure 7.11. Fluency Comparison in the Evaluation Sessions Sessions (with SD bars)

By contrast, the main effect of MTPE modality on fluency was statistically significant ($F(1, 10) = 9.80$, $p = 0.01$), indicating that MTPE modality alone did statistically significantly influence fluency outcomes. By observing the averages in both MTPE modalities, we can see that Fluency scores in IPE ($M = 3.89$; $SD = 0.09$) were statistically significantly higher than Fluency scores in TPE ($M = 3.87$; $SD = 0.09$).

The interaction effect between interaction session and MTPE modality was also statistically significant ($F(2, 20) = 10.19$, $p = 0.001$), which implies that the effect of interaction session on fluency scores is modulated by the MTPE modality. In essence, while increased experience alone does not change fluency scores, the MTPE modality has a statistically significant effect on fluency scores. In addition, the two factors together (interaction session and MTPE modality) also have a significant impact on fluency. That is, that the effects of experience vary for each MTPE modality.

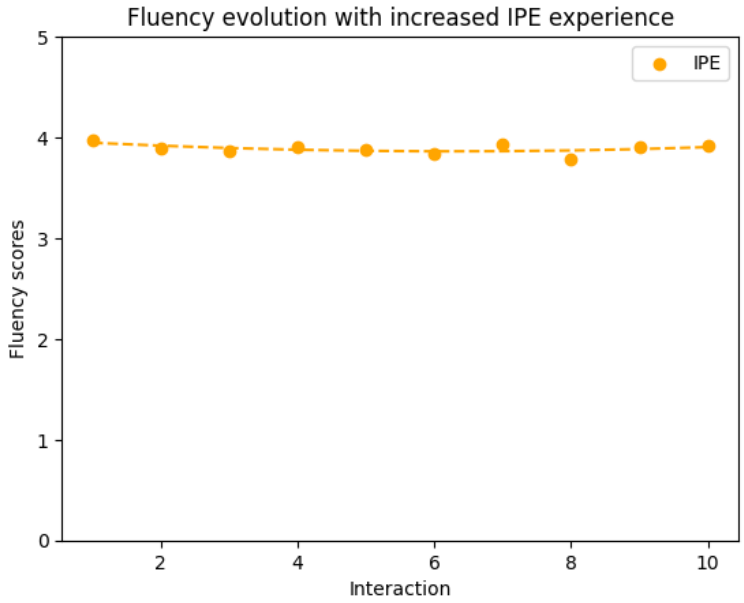

Figure 7.12. Fluency evolution during the learning sessions

Figure 7.12 illustrates that fluency scores, anchored firmly around the score of 4, remained remarkably consistent throughout the 10 learning sessions. This sustained level of fluency suggests that fluency scores did not change with increased experience with IPE. It is also worth stressing that fluency scores were close to the maximum fluency score of 4, affirming the viability of IPE for professional translation tasks where fluent and natural-reading translations are paramount.

7.4. RQ4. Is adequacy statistically significantly impacted by MTPE modality (TPE or IPE) and does this vary with increased experience?

The statistical analysis of the influence of the interaction session and the MTPE modality on the adequacy scores of the translations produced by the different participants was also assessed through a 2x3 repeated measures ANOVA.

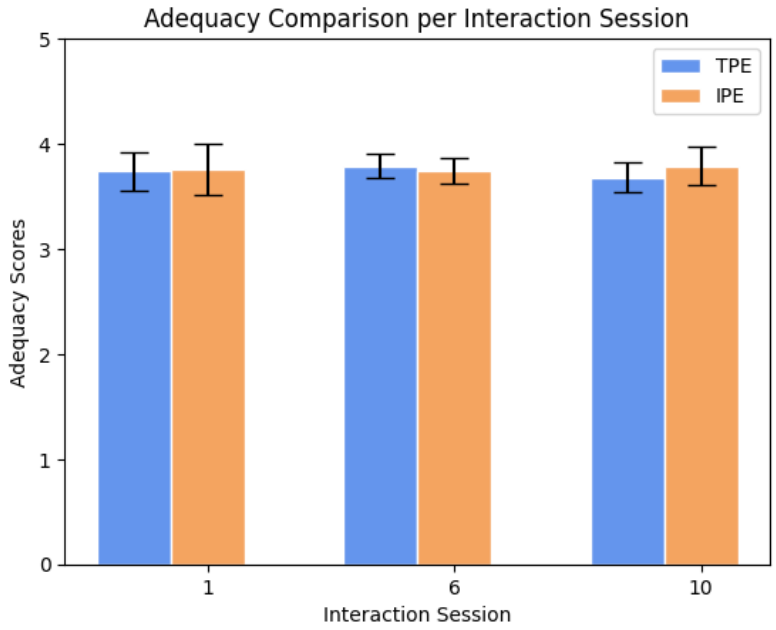

Figure 7.13. Adequacy Comparison in the Evaluation Sessions

The analysis did not indicate a statistically significant main effect of interaction session on adequacy scores ($F(2, 20) = 0.31$, $p = 0.71$), suggesting that increasing interaction session alone did not substantially enhance the adequacy of translations. Similarly, the main effect of MTPE modality on translation adequacy was not statistically significant ($F(1, 10) = 0.77$, $p = 0.40$), indicating that the choice of MTPE modality by itself did not statistically significantly influence the adequacy of the translations. Even if the adequacy scores of the IPE modality ($M = 3.76$; $SD = 0.19$) were higher than the adequacy scores of the TPE modality ($M = 3.74$; $SD = 0.15$), no statistically significant difference could be observed, as observed in Figure 7.13.

The interaction between interaction session and MTPE modality was also not statistically significant ($F(2, 20) = 1.85$, $p = 0.19$). This implies that the potential effect of interaction session on translation adequacy did not vary with the different MTPE modalities. Collectively, these results suggest that neither the increased experience with a system nor the MTPE modality significantly impacts the adequacy of the translations.

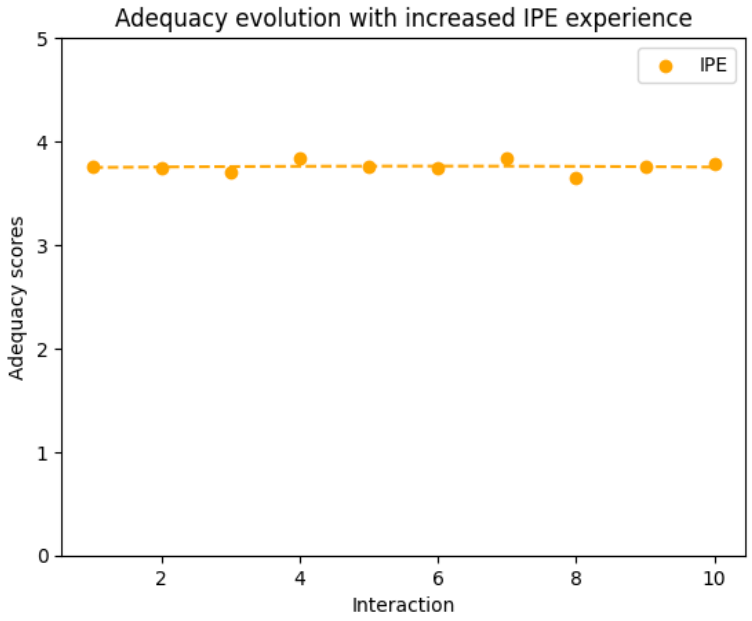

Figure 7.14. Adequacy evolution during the learning sessions

Figure 7.14 reflects the adequacy evolution across the 10 learning sessions. Adequacy scores consistently hovered around the maximum adequacy score (4), indicating a stable performance in maintaining the integrity of the source content throughout the interactions. It is significant to note, however, that these scores were modestly lower than those obtained for fluency in the section above. This applies both to TPE and IPE: even if IPE scores in terms of adequacy and fluency are generally higher than TPE scores, both TPE and IPE adequacy scores report slightly lower global scores than fluency. This discrepancy emphasizes a critical nuance in translator-MT interactions: while fluency may be more readily achieved with the aid of state-of-the-art NMT systems, ensuring the translation's adequacy—a measure of how well the source message is preserved—may pose a greater challenge, and may depend more on the translator.

7.5. RQ5 to RQ7. Do pre-task perceptions of MTPE correlate with fluency, adequacy, or productivity?

As discussed in Chapter 5, there were some additional elements related to participants' perceptions that we wanted to assess in this longitudinal study. As this PhD thesis can be considered the first study analysing MTUX in translator-MT interactions, exploring all the potential implications of MTUX (in terms of pre-task perceptions or post-task perceptions) may give us indications for developing more HCAMT tools, systems, or workflows.

Consequently, to answer RQ5 to RQ7, we conducted a series of statistical tests to see whether there was any correlation between participants' pre-task perceptions and fluency, adequacy, or productivity (Briva-Iglesias and O'Brien 2024).

First, we ran some descriptive statistics tests to see the distribution of the data, so we could use more appropriate inferential statistical tests. In addition, we plotted every variable in histograms to see whether the variables were normally distributed, and to strengthen our methodology we also performed the Shapiro-Wilk's test. As data violated the assumptions of normal distribution ($p>.05$), we conducted a Kendall's T correlation test for all the variables so as to explore the relationships between the measures collected, by following recent recommendations to use Kendall's T over Spearman's correlation for non-parametric data (Mellinger and Hanson 2016). Due to the number of correlations performed, increasing the likelihood of type I error, we recommend interpreting correlations at the .05 level with caution. In addition, it is worth stressing that the strength of the correlation coefficients vary according to the statistical test conducted. Therefore, by following Schober, Boer, and Schwarte's (2018) advice, we interpret Kendall Tau's correlation coefficient strength in the following form: Weak (0.06-0.25), Moderate (0.26 to 0.49), Strong (0.50 to 0.71), and Very strong (0.71 to 1). Below, different heatmaps display the correlation coefficients of every pre-task perception variable in relationship to adequacy, fluency, and productivity. Variables containing an asterisk "*" show a statistically significant correlation.

Fluency scores

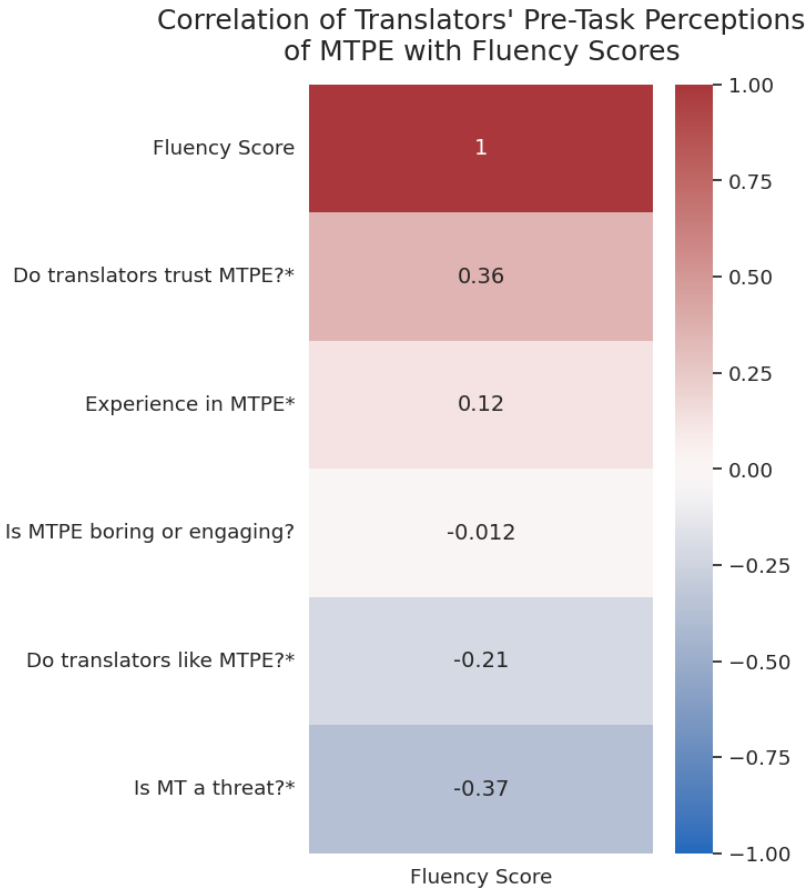

Figure 7.15. Correlation of translators' pre-task perceptions of MTPE with fluency scores

Figure 7.15 shows that participants' feeling of boredom or engagement when performing MTPE assignments in a professional environment (r (10) = -.012, p = .85) showed no statistically significantly correlation with Fluency scores.

However, all the other pre-task perceptions variables showed a statistically significant correlation with Fluency. On the one hand, the level of experience that participants had in performing MTPE tasks (r (10) = .12, p = .002) and participants' attitude towards liking or disliking post-editing tasks (r (10) = -.21, p = .0007) showed weak statistically significant correlations. On the other hand, participants' pre-task perceptions of MT being a threat to the translation profession (r (10) = -.37, p = .001) and the level of trust they had on MTPE (r (10) = .36, p = .02) showed statistically significant moderate correlations. This means that participants who had higher levels of trust in MTPE tasks as an aid in their professional translation projects tended to report higher fluency scores, as the moderate correlation

shows. In a similar way, those participants who thought that MT was a threat for their profession tended to produce less fluent translations.

Adequacy scores

The correlation heatmap of Figure 7.16 provides a visual summary of the statistical analysis conducted on participants' pre-task perceptions of MTPE and their translation quality results, specifically focusing on translation adequacy. Notably, the heatmap reveals a spectrum of correlations, from strongly positive to strongly negative.

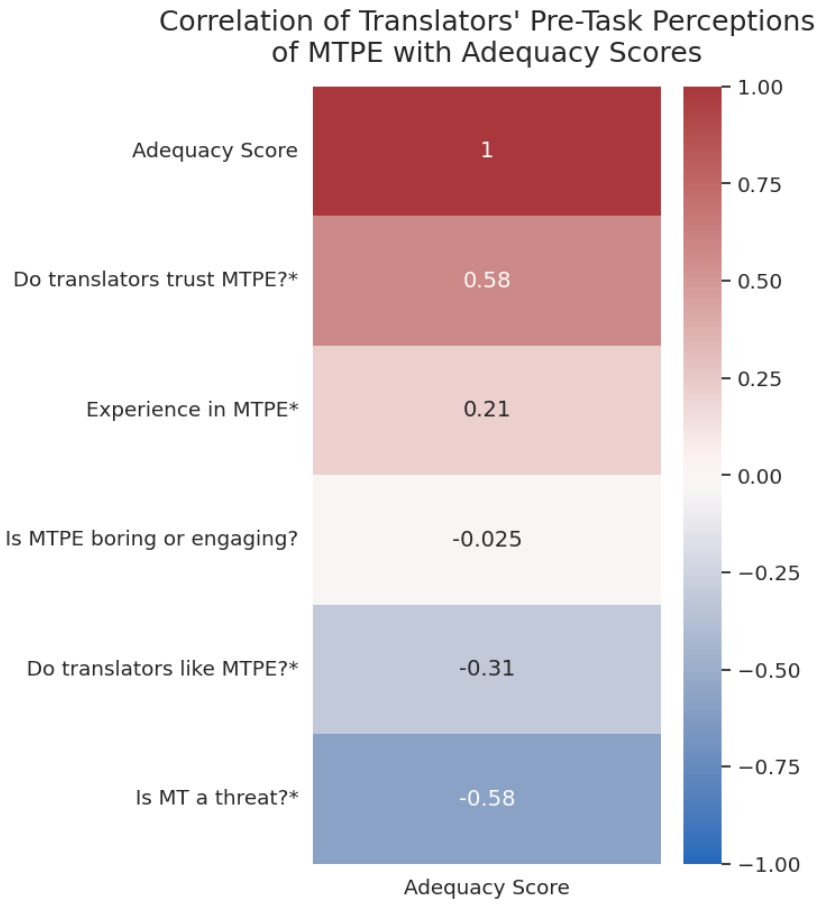

Figure 7.16. Correlation of translators' pre-task perceptions of MTPE with adequacy scores

One of the most notable results within our participants group is that we observed a strong statistically significant positive correlation ($r(10) = .58$, $p = 0.0001$) between participants' trust in MTPE and the adequacy scores, implying that higher trust in the system is linked to higher performance levels in producing adequate translations. The other remarkable result is that participants' view of MT as a threat yielded a strong statistically significant negative

correlation (r (10) = -.58, p = 0.0001), suggesting that apprehensions about the technology's impact on the profession may undermine translation adequacy.

Factors such as the enjoyment of conducting MTPE tasks (r (10) = -.31, p = 0.0001), and the overall experience in MTPE (r (10) = .21, p = 0.0004), showed less pronounced yet statistically significant correlations, indicating that these perceptions might not be as critical in influencing the adequacy of translation outcomes.

These insights contribute to the ongoing discourse on the human factors influencing contemporary translator-computer interactions, underscoring the complex interplay between subjective perceptions and objective translation performance metrics. These results show that participants' lack of trust in MTPE tasks and the consideration of MT as a threat to their profession may have an effect on translation quality, even before the task has already started.

Productivity scores

Figure 7.17 provides a quantitative depiction of the correlations between participants' pre-task perceptions of MTPE and their measured productivity in WPH. The results from every pre-task perception variable were statistically significant.

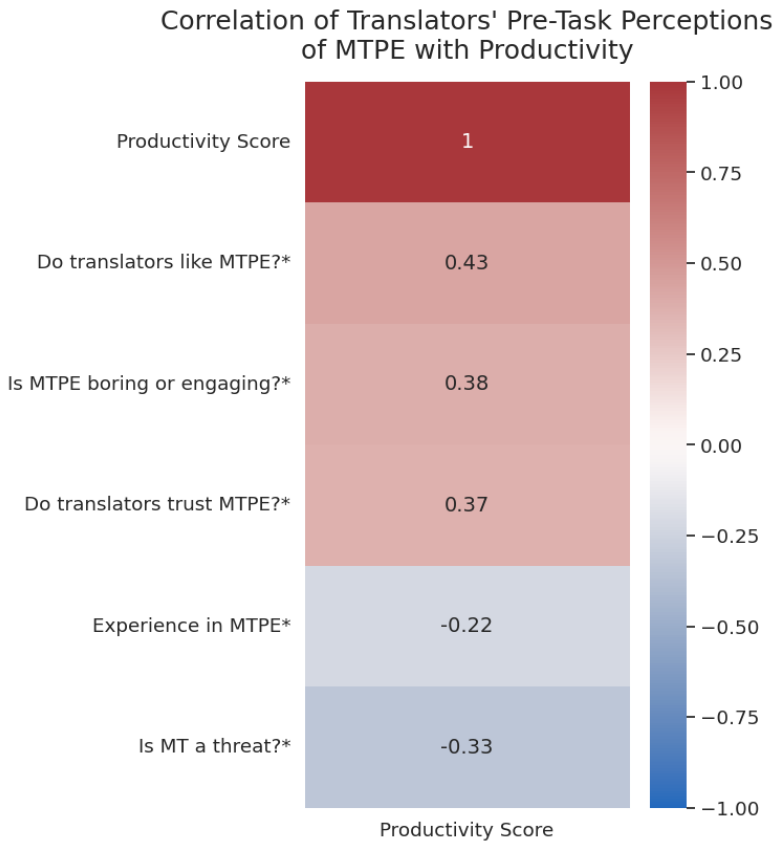

Figure 7.17. Correlation of translators' pre-task perceptions of MTPE with productivity

The data indicates that positive sentiments towards MTPE, such as liking MTPE tasks as a professional aid (r (10) = .43, p = 0.0001), finding them engaging (r (10) = .38, p = 0.0001), or the level of trust in MT (r (10) = .37, p = 0.0001) are moderately correlated with higher productivity scores.

Conversely, the negative moderate correlation with the perception of MT as a professional threat (r (10) = -.33, p = 0.0001) highlight potential areas of concern. These findings may reflect a complexity in MTPE's perceived impact on the translation industry, which could influence translator productivity. There is also a weak statistically significant negative correlation with MTPE experience (r (10) = -.22, p = 0.0001).

The results of this section reveal the psychological impact of participants' pre-task perceptions on translation performance measures, affirming the importance of considering MTUX in the design of MTPE systems, tools and workflows. These results also have implications in translator training. If negative translators' pre-task perceptions are correlated with lower productivity and lower quality, teaching translators the potential benefits of MT

(if applied in an ethical way) may translate into translators being more productive and offering higher-quality translations.

7.6. RQ8 to R10. Does MTUX correlate with fluency, adequacy, or productivity?

We also wanted to analyse whether MTUX scores had any statistically significant relationship with translation quality or productivity. Thus, different variables were checked for correlation, namely translator's overall MTUX scores in the TPE and IPE condition with measures of translation adequacy, fluency, and productivity. First, we checked data normality with Shapiro-Wilk's test. MTUX scores both in the TPE and the IPE modalities followed a normal distribution (p-value > .05). Adequacy scores in IPE were also normally distributed. All the other variables (i.e., productivity scores in TPE and IPE, fluency scores in TPE and IPE, and adequacy scores in IPE) did not follow a normal distribution (e.g., violated the assumption for normality (p-value < .05)). We therefore performed a Kendall's T correlation test for every non-parametric variable, except when we correlated the normally distributed variables, where we conducted a Pearson's correlation test. As a conclusion, we found there were no statistically significant correlations between MTUX scores and final translation quality. MTUX scores in the TPE condition [Adequacy: r (32) = -0.04, p = 0.83; Fluency: r (32) = 0.05, p = 0.67] and in the IPE condition [Adequacy: r (109) = 0.11, p = 0.08; Fluency: r (109) = 0.08, p = 0.19] did not significantly correlate with final translation quality. Regarding participants' productivity, MTUX scores in the TPE condition [r (32) = 0.08, p = 0.295] or in the IPE condition [r (109) = −0.09, p = 0.14], did not significantly correlate with measures of translator productivity.

7.7. Discussion of the results

This section discusses the findings of the main longitudinal study reported in Chapter 7. As discussed in the rationale for conducting a longitudinal study in Chapter 5, the overarching research question is affected by different factors, namely MTUX, translation productivity, and translation quality. As a consequence, the results are also discussed considering these factors, first individually, and then as a whole. This section starts with a discussion of the MTUX scores, first on an average level, and then at a subfactor level (Section 7.7.1), followed by the

translation productivity factor in Section 7.7.2, the translation quality in Section 7.7.3, and the correlations that MTUX have on translation quality and productivity (Section 7.7.4). The Chapter ends with a final discussion on the implications of the results (Section 7.7.5). The research questions are revisited to find whether they are answered by the findings presented.

7.7.1. IPE produces a statistically significantly higher MTUX

The findings of this PhD research underscore the superiority of the IPE modality over TPE in terms of average MTUX scores. It is evident from the statistical analyses that participants perceive a more HCAMT workflow when engaging with MT through IPE tasks. This observation is supported by the statistically significantly higher MTUX scores in the IPE modality, indicating a substantial enhancement in the overall user experience if compared with MTUX scores in the TPE modality. The observation also allows us to answer affirmatively the first part of *RQ1. Is MTUX statistically significantly impacted by MTPE modality (TPE or IPE) and does this vary with increased experience?* IPE produces a statistically significantly higher MTUX than TPE.

These results support the findings of previous IPE studies that reported that this MTPE yielded higher translation satisfaction (Sánchez-Torrón 2017). Furthermore, the results also suggest that, as participants become more familiar with the IPE features, their proficiency and comfort with this type of translator-computer interaction improve, leading to a more positive perception of the MTUX, answering affirmatively the second part of RQ1. This highlights the importance of adequate training and adaptation time for translators to fully acclimatize to the IPE modality and leverage its benefits in their workflows, as suggested by previous research on IPE (Sánchez Torrón 2017; Sanchis-Trilles et al. 2014). If the perspective from which MTUX scores are analysed is changed, and the different MTUX subfactors are scrutinised for finer granularity, we can also extract additional interesting lessons.

The main longitudinal study reveals that IPE is perceived as a more pleasant and friendly MTPE modality by our participants. The inherent appeal of the IPE modality, as evidenced by its higher Attractiveness scores, suggests a more engaging and user-friendly interface that resonates well with users. In addition, IPE's ease of use and understandability are notably superior to TPE (as measured by Perspicuity scores). Even though participants already had professional experience with TPE tasks, IPE's design, highlighting translation completion proposals and providing different complementary aids, has been deemed more intuitive and straightforward.

Participants also perceived that IPE tasks were faster than TPE tasks (by considering Efficiency scores). However, it is important to triangulate these subjective assessments with empirical productivity data for a more comprehensive understanding of IPE's efficiency. As reported by Guerberof-Arenas (2008), translators sometimes perceived that they were faster without MT, but empirical data showed the opposite, indicating a higher productivity when post-editing. In section 7.7.2 below, we describe how participants' productivity data shows that IPE tasks allowed them to translate statistically significantly faster than with TPE. Consequently, perceptions need to be examined alongside productivity.

The findings of the main longitudinal study also indicate that IPE is a more motivating and exciting MTPE modality for participants in a substantial way (Stimulation factor). This could be attributed to the interactive nature of IPE, which may stimulate more cognitive engagement and interest in the translation task. In addition, IPE tasks are perceived as more creative, possibly due to the novel features and the dynamic interaction they offer, challenging participants to engage more actively with the translation process (if considering Novelty scores). This dynamic interaction may also influence translation fluency. In TPE, translators need to adapt static MT output and most of the text may remain unchanged. In IPE, on the other hand, translators start typing and the MT output adapts on the fly to the proposed translation of the translator. As a consequence, the resulting text may be more fluent in IPE, though this should be analysed with empirical data. Section 7.7.3 below covers the effects on translation quality, in terms of fluency and adequacy. Once again, empirical data back participants' perceptions, and they produced more fluent translations in IPE than in TPE.

The non-statistically significant difference in Dependability scores between IPE and TPE suggests that participants are still adapting to the IPE modality. Familiarity with TPE might contribute to a sense of reliability, whereas IPE, being a newer MTPE modality, may require more time for participants to fully trust and adapt to its features, as suggested previously by Macklovitch (2006). However, it is worth stressing that average Dependability scores were higher for IPE in every interaction session, even if this difference was not statistically significant.

7.7.2. IPE allows for working statistically significantly faster after some acclimatisation

The statistical analysis conducted to answer *RQ2. Is translation productivity statistically significantly impacted by MTPE modality (TPE or IPE) and does this vary with increased experience?* revealed that the MTPE modality played a crucial role in translation productivity scores. The data indicates a statistically significant effect of the MTPE modality on productivity, with IPE outperforming TPE in terms of translation productivity in WPH. This is a key finding, underscoring the effectiveness and efficiency of interactive tools in the translation process.

Interestingly, initial interactions did not show a substantial difference between IPE and TPE. This can be attributed to the initial learning phase and adaptation to the IPE interface and functionality (Alabau et al. 2013;Alabau et al. 2013; Sanchis-Trilles et al. 2014). As participants progressed in their use of IPE, a clear increase in productivity was observed, as suggested previously by Casacuberta et al. (2009). On average, over the 10 interaction sessions, participants increased their productivity in IPE by 64%. By contrast, in TPE, this figure only increased by 5%. For example, the translator 1 (T01) had a TPE productivity of 704 WPH at the start of the experiment and ended up with a TPE productivity of 701 WPH at the end of the experiment (a -0.4% productivity change, suggesting that their productivity in TPE stayed flat). Conversely, T01 had an IPE productivity of 632 WPH at the start of the experiment and ended up with an IPE productivity of 1216 WPH (a 92% increase after acclimatisation to the tools and the 10 interaction sessions).

The longitudinal approach of this study played a crucial role in capturing the evolving proficiency of participants with the IPE system, considering their varying experience levels with each MTPE modality. This approach provided a nuanced understanding of how familiarity and expertise with IPE tools influence translation productivity over time.

This research distinguishes itself as the first in the field to harness empirical data in demonstrating the impact of the IPE modality on productivity. Previous studies speculated on this effect based on regressions and hypothetical calculations (e.g. the work on CASMACAT reported in Alabau et al. (2014) or the early TransType prototypes (Esteban et al. 2004; Macklovitch 2006)), but our work substantiates these theories with direct data analysis. This leap from theoretical to empirical validation marks a significant advancement in the study of translator-computer interactions and situates IPE as a feasible and viable workflow for today's

contemporary language services industry. IPE does not only grant a higher MTUX to participants, but also allows them to translate faster.

It is worth stressing that these results may also be partially attributed to recent technological advancements. With the adoption of NMT as a new state-of-the-art MT paradigm (Castilho, Moorkens, Gaspari, Sennrich, et al. 2017), plus improvements in computational power, these advancements have likely contributed to the faster generation of translation completion proposals and higher quality MT output, streamlining the translation process. This technological evolution, coupled with refined IPE tools, and state-of-the-art, adaptive NMT systems has paved the way for greater translation productivity in translation, setting a new benchmark in the field.

7.7.3. IPE statistically significantly impacts fluency, but not adequacy

Regarding translation quality, this PhD research tried to answer, through RQ3 and RQ4, whether IPE produced a statistically significantly higher fluency and adequacy scores than TPE. The findings indicate that participants produced more fluent translations when working with the IPE modality. This fluency difference with their TPE translations was statistically significant. This suggests that the intrinsic dynamic interaction of the IPE modality, where translation proposals are adapted on the fly, enhances the fluency of the output, as suggested above (see Section 7.4). The ability of IPE to adjust translation suggestions in real-time in response to translator inputs appears to be a key factor in achieving higher fluency levels (a step-by-step example of how participants produce more fluent translations with IPE if compared with TPE can be seen in Appendix E). Recent research has identified the phenomenon of “post-editese” (Daems, De Clercq, and Macken 2017; Toral 2019; Castilho and Resende 2022), a distinctive feature set in post-edited translations. Post-edited translations have been found to be simpler and to have a higher degree of interference from the source language than human translations. An intriguing future research question emerges from our findings: does the interactive nature of IPE reduce the occurrence of "post-editese"? This question opens new avenues for exploring how different MTPE modalities influence the linguistic characteristics of translated texts and poses IPE as a potential MTPE modality that helps produce more natural translations, as if written directly in the target language.

In addition, it is worth noting that the data reveals consistently high fluency scores for both IPE and TPE modalities across all interaction sessions. This uniformity in fluency can likely be attributed to the advancements in NMT, known for generating highly fluent outputs (Bentivogli, Bisazza, et al. 2016).

In terms of translation adequacy, the study did not find a statistically significant difference between TPE and IPE over time. However, it is noteworthy that the average adequacy score was slightly higher for IPE than for TPE. This suggests that while IPE may offer a marginal benefit in maintaining the integrity of the source message, the difference is not substantial enough to be statistically significant. From these results, we can also extract that translation adequacy is translator-dependent, i.e., a good translator will deliver high translation quality in both MTPE modalities, regardless of experience level; just as a less skilled translator will deliver poorer quality translations, regardless of the MTPE modality or the experience.

To have a more comprehensive view of the translation quality produced by the participants of the study and the relevance for the language services industry, we have compared the fluency and adequacy scores of the raw MT output with the final adequacy and fluency scores of the TPE and IPE modalities. The raw MT output obtained an average Adequacy score of 3.48/4 and an average Fluency score of 3.71/4. The TPE output obtained an average Adequacy score of 3.74/4 and an average Fluency score of 3.86; and the IPE output and average Adequacy score of 3.76/4 and an average Fluency score of 3.89/4. Thus, we can consider that participants of the main longitudinal study produced translations that are valid and meet the industry requirements. This finding is significant for the language services industry, as it underscores the practical applicability of both MTPE modalities in professional settings.

7.7.4. Perceptions of MT influence quality and productivity

RQs 5 to RQ10 attempted to investigate whether participants' MTUX perceptions, analysed before, during, and after interacting with MT, had any relevant relationship with translation quality and productivity. Even if the study results indicate that during- and post-task MTUX perceptions collected via the UX questionnaire did not demonstrate a direct correlation with translation performance (neither at the quality nor productivity level), we deem them critical in forming translators' pre-task perceptions. These pre-task perceptions, in turn, have shown

a moderate to strong statistically significant correlation with translation quality and productivity. Consequently, we consider that it is essential to study translators' pre-task and post-task perceptions of MTUX collectively, as they together influence the overall experience and expectations of translators. The reasons that justify why we think that pre- and post-task MTUX perceptions should be analysed together are described below in Section 8.1, after explaining the correlations between translators' pre-task perceptions and translation quality and productivity in this section.

One of the study's salient findings is the statistically significant correlation between participants' pre-task perceptions of MT and translation quality and productivity. The most notable correlation coefficients were observed in two specific variables: participants' level of trust in MTPE and the perception of MT being a threat to the translation profession.

The level of trust participants have in MT showed a strong correlation with Adequacy and a moderate correlation with Fluency, supporting once again the results discussed above, which suggested that Adequacy was more translator-dependent, while Fluency was more influenced by NMT quality (Castilho, Moorkens, Gaspari, Sennrich, et al. 2017). These correlations indicate that participants who trust MT systems to help them work in their daily tasks offered higher final translation quality than those who did not trust MT systems. These results are in line with previous research in cognitive science (Albarracín 2021), which indicated that prior negative perceptions are an important and crucial determinant for future attitudes and behaviours. In our case, participants' pre-task perceptions of MTPE tasks had a strong negative correlation with the quality of the translation. This may be because participants who do not trust MT do not enjoy this interaction or do not give their best when interacting with MT in their regular workflows. This backs up the results of the second pre-task perception variable with a strong negative correlation in our study, that is, whether participants consider MT as a threat to the translation profession. The perception of MT being a threat to the profession showed a statistically significant moderate to strong association with Fluency and Adequacy, respectively. What this correlation implies is that participants who, even before starting a post-editing task, think that MT is a threat and harmful for the translation profession are more likely to produce lower translation quality. If we combine this negative correlation with the positive correlation of trust in MTPE tasks, we can say that translators' pre-task perceptions have a strong relationship with the final quality of the translation, and that the

mindset or attitude with which translators start a post-editing task is key to their ability to deliver high-quality translations.

Furthermore, the results of the main longitudinal study also suggest that translation quality is not the only variable affected by participants' pre-task perceptions. Whether participants liked or disliked doing MTPE tasks in their professional workflow was also positively correlated with translation productivity. This meant that participants who liked post-editing recorded higher productivity scores than those who did not like MTPE tasks.

The study's findings on the relationship between participants' perceptions and translation quality and productivity have profound implications, offering novel insights into the dynamics of modern translator-computer interactions. It is evident that the approach translators adopt towards a task plays a critical role in determining the final outcome, with varying degrees of influence on different aspects of translation performance (in terms of quality or productivity). This highlights the vital importance of MT literacy (Bowker and Ciro 2019), so that translators working with MT know what they can or cannot do with MT, so their views or perspectives change even before this specific type of user-MT interaction takes place. The chapter below concludes with a reflection on these implications.

CHAPTER 8. CONCLUSIONS, STRENGTHS, LIMITATIONS AND FUTURE WORK

The aim of this research was to answer the following overarching research question:

- RQ. Is IPE a better alternative to TPE in terms of MTUX, translation productivity, and translation quality?

The PhD thesis demonstrates that IPE is a better alternative to TPE at different levels in the use case analysed. In terms of MTUX, participants reported a statistically significantly higher MTUX when interacting with MT in the IPE modality. This means that participants perceived the IPE task to be more pleasant, friendly, practical, and efficient, among other experiential elements. The objective results also show that, after some acclimatisation to the novel features of IPE, participants were able to translate faster than with TPE at a statistically significant level. In terms of translation quality, on the one hand, participants produced more fluent translations with IPE than with TPE, probably due to the advantages of the interactive human-MT interaction. On the other hand, although the adequacy evaluation revealed no significant statistical difference between the translations produced via IPE or TPE, IPE consistently yielded slightly higher adequacy scores. This underscores the significance of IPE in enhancing HCAMT in contemporary translation workflows because IPE does not only produce a more pleasurable human-machine interaction, but also augments translators and allows them to work faster and produce more fluent translations with comparable adequacy, at least within the scope of our study.

It is worth stressing that these results may be applicable to the professional translation of specialized, complex legal texts from English into Spanish. However, we must be aware that this is a language combination between two major languages with high-quality MT. Consequently, even if we cannot claim generalisability of the results, we can anticipate that there may be transferability in similar scenarios (Saldanha and O'Brien 2013). In other words, the results may be applicable in other translations directions between major languages where there is high-quality MT available, and the translation takes places in the legal domain. It remains to be investigated whether these results will also be the same in other specialised fields, such as medical translation, or in more generic fields of translation.

Yet, it is important to stress that, with the current situation and the excellent quality that NMT systems offer now, we hypothesised it was possible that TPE would allow translators to be more productive than IPE in language combinations with MT output of high-quality because not many changes had to be implemented through post-editing. Results show that, even in this case, translators are statistically significantly faster in IPE. In the case of lower quality MT language combinations, IPE may be even a more viable option because the adaptivity features may improve the low-quality MT output by considering what the translator has already started to type. This is another research question that needs to be further explored.

8.1. Breaking the vicious circle: designing for pleasure rather than for absence of pain

An additional goal of this PhD thesis was to establish a methodology for developing HCAMT tools, systems, and workflows through the measurement of MTUX and the comparison with objective performance data (e.g., translation quality and productivity in the case of the language services industry), so that researchers interested in other types of users, language pairs, and domains, could replicate this methodology. Consequently, this study marks one of the early ventures in UX- and HCI-informed research on MT. To date, technology adoption in the language services industry has been done through human adaptation (Winner 2007; Vallor 2024), where the technology is developed in the first place, and humans are then trained to adapt to the technology.

We consider that this is not the appropriate way forward. Hence, this research makes a theoretical contribution by fostering the adoption of technology by following the opposite direction: understanding users first and, then, developing technology that meets users' needs. This should be the goal of HCAMT tools, systems, and workflows. This theoretical contribution has been built by utilising Roto's (2016) components of UX (considering the user, the system, and the context where the interaction takes place), which has proven to be a fruitful basis for the research design. At the very least, it moved beyond previous research where user experience, background, context, task and system interaction were often neglected, as discussed in Chapter 4. We therefore emphasise the pivotal role of MTUX in human-MT interactions, spotlighting its critical role in shaping the dynamics between users

and MT. Understanding users' perceptions before, during, and after interaction with MT can (and should) inform the development of more HCAMT tools, systems, and workflows.

The foundational rationale behind this PhD research stemmed from the growing dissatisfaction among translators towards technology (as discussed in Chapter 5). Historically, the focus of MT development by companies has been predominantly on enhancing productivity and quality, neglecting the user's experience and needs in translator-computer interactions (Briva-Iglesias, O'Brien and Cowan 2023). This oversight has led to generally negative user experiences in TPE tasks, fuelling a vicious cycle of translator dissatisfaction, rejection of technology or non-adoption of MTPE modalities (Torres-Hostench et al. 2016; Macías 2020; Cadwell, O'Brien, and Teixeira 2018; Fırat 2021; ELIS Research 2023; Moorkens 2020; 2023). This technology rejection can be seen as a double-edged sword. By acknowledging that technology should always be developed and adopted in an ethical and sustainable way, the rejection of technology may hinder translators from leveraging the advancements in MT, potentially leading to reduced augmentation, increased cognitive effort and the need to do more repetitive and time-consuming tasks. This can contribute to job dissatisfaction and burnout due to the continued reliance on more traditional translation methods. On a larger scale, technology rejection could also have a societal impact and slow the overall advancement and adoption of efficient translation solutions, impacting global communication and information exchange, especially in multilingual contexts where rapid and accurate translation is crucial (e.g. in the legal, medical, or academic domains). On the other side, technology rejection can also have positive outcomes because it can drive the development of more human-centered tools, systems, and workflows, encouraging a shift from a purely efficiency-focused approach to one that better addresses the needs and experiences of users. Today's increasing technology rejection in the language services industry is what triggered and motivated this PhD thesis.

Our results highlight a vicious circle that starts with today's non-human-centered technology development and adoption. If MT users are not satisfied with their interactions with MT, they will have negative perceptions of MT, viewing it less as an aid and more as a hindrance. The next time translators will interact with MT, these negative pre-task perceptions will result in a user-MT interaction where translators will not give their best, as suggested by our results. Then, this will translate into a non-pleasurable and demotivating interaction, as well as to

lower translation productivity. This vicious cycle is perpetuated by the continuous interaction of translators with systems that do not align with their expectations or needs.

Thus, how can we break this vicious circle? Although post-task MTUX scores did not show a direct correlation with translation performance measures, they are identified as the key element in breaking this cycle. The development and adoption of HCAMT tools, systems, and workflows aims to elevate MTUX scores. By enhancing MTUX during- and post-task perceptions, translators will be in a position where their pre-task perceptions of MTPE are higher before engaging in new translator-MT interactions. As a consequence, translators are expected to engage more positively with MT in future interactions. This positive engagement is not just about feeling respected and in control, but also about being comfortable in their interaction with MT. Consequently, such an enhanced MTUX is anticipated to lead not only to higher satisfaction and motivation, but also improved translation quality and increased productivity, breaking the vicious circle that the use and adoption of non-human-centered technology originated.

As a conclusion and reusing Blackwell's (2015) quote on the goal of HCI, we conducted a "questioning, provocative, and disruptive" research project on contemporary human-MT interactions. This PhD advocates for a paradigm shift in the development of MT, aiming to design systems, tools and workflows that prioritize user pleasure and satisfaction over mere functionality and efficiency through better user-MT interactions. Technology development and adoption should contribute to our quality of life by designing for pleasure rather than for absence of pain. By focusing on enhancing MTUX, the research suggests a pathway to more pleasurable and efficient human-computer interactions, urging developers to implement and measure these improvements. We have demonstrated that taking care of users' needs and experiences is not mutually exclusive from today's productivity and efficiency focus. Indeed, these two approaches should be considered together.

8.2. Strengths

As per the notable contributions this PhD thesis makes to the Translation Studies and the MT communities, we must stress that this research project was conducted applying mixed methods, utilizing both quantitative and qualitative data to triangulate findings. This

enhanced the robustness and reliability of the results (Alves 2003). In addition, the innovative use of a standardized HCI UX questionnaire in a previously neglected area of MTUX marks a significant advancement, providing validated insights into user satisfaction and interaction with MT systems. This proposed methodology will allow researchers to explore the experiences and MT needs of different MT users in multiple use cases. The execution of a two-week longitudinal study with professional translators is exceptionally rare in the domain of Translation Studies and MT due to its logistical challenges and costs. This approach has provided invaluable longitudinal data, capturing the evolving user experience and productivity over time, a methodological strength that offers deeper insights than cross-sectional studies. In addition, the data is shared in an open repository to invite researchers to further explore the TPE and IPE processes, as well as to allow for the replication of this study.

Working with complex legal texts deviates from the norm of using simpler texts (normally news) in MT research, thereby challenging state-of-the-art MT systems and providing insights into their performance on more demanding content. The engagement of professional translators ensures the validity and reliability of the data collected, recognizing the importance of expertise in translation tasks. This choice underscores the study's commitment to capturing authentic interaction experiences and outcomes.

The meticulous approach to reducing reviewer subjectivity in quality assessment, involving multiple reviewers and robust translation annotation guidelines after achieving a high IAA, demonstrates a rigorous and methodologically sound approach to evaluating translation quality. This aspect of the study addresses a common critique in Translation Studies regarding the subjectivity of human assessments of translation quality.

Collectively, these methodological choices and the study's execution showcase a rigorous, innovative, and comprehensive approach to understanding the nuances of human interaction with MT systems, particularly in professional translation contexts. This research not only contributes valuable empirical evidence to the field but also sets a new standard for future studies in MT and HCI within the domain of Translation Studies.

8.3. Limitations

Even if this research represents one of the few longitudinal studies in Translation Studies and MT, providing valuable insights into the interaction between professional translators and MT over time, it is important to acknowledge the challenges and limitations inherent in conducting such a study. Ideally, evaluating ten TPE interactions, akin to our approach with IPE, would have offered a more comprehensive view, allowing for a direct comparison of TPE and IPE in every translation session. Unfortunately, due to budget constraints, this was not feasible and we could only collect data from three TPE interactions.

Another limitation of our study is the lack of analysis concerning the potential benefits of IPE for translators who are not touch typists. The impact of typing speed on the productivity of IPE versus TPE remains an unexplored dimension in our research. Typing skills could potentially influence a translator's interaction with IPE because slow translators may give additional time to the system to rerun its algorithm and offer new translation completion proposals. This, together with the fact that Lilt is a proprietary tool and that we had no application to measure the keystrokes, could have provided additional insights. Future research could beneficially explore these aspects to provide a more holistic understanding of how different translators interact with IPE and TPE systems.

In addition, at the late stages of the PhD, a new MT architecture based on large language models (LLMs) appeared (Jiao et al. 2023). Although the MT capability of these systems has started to be researched (e.g. Lyu, Xu, and Wang (2023), Castilho et al. (2023), Briva-Iglesias, Camargo, and Dogru (2024)), the user studies of the PhD used NMT and LLMs have not been considered.

8.4. Future work

Future research should investigate the applicability of these results to different language combinations, especially those with lower-quality MT outputs. It would be valuable to determine if IPE offers greater advantages in such scenarios, potentially offering insights into MT's adaptability across diverse linguistic contexts. In addition, exploring the effectiveness of IPE in various translation domains remains an open question. The PhD focused on the legal translation field, but additional translation domains like technical, medical, or literary translation may benefit differently from IPE. This could provide a broader understanding of

its utility across the translation spectrum. For instance, Guerberof-Arenas and Toral (2022) report that translators are less creative with TPE and this impacts on their UX. Would IPE impact similarly on the product from a creativity perspective and on their perceptions? It may be the case that translators could leverage the interactivity of IPE and produce more creative translations. This is a question that needs to be further explored.

Another intriguing avenue is examining how IPE advantages may vary between novice and experienced translators. This could reveal insights into the learning curves associated with IPE and how different levels of translation expertise interact with new MT modalities. Then, exploring the study beyond the 10 days of interaction may also be really interesting to see when the productivity in IPE flattens, and compare the maximum human productivity in both MTPE modalities.

In addition, there is potential in exploring whether IPE can aid in language learning or in developing diverse translation strategies. This aspect could be particularly beneficial in educational settings, offering a dual advantage of language acquisition and translation skill development. To date, there has been limited work on the use of MT for language acquisition (Deng and Yu 2022), and IPE may be a more appropriate way of interacting with MT than TPE to this specific end.

But, most importantly, this PhD opens a vast new research world because of the generalisability of MTUX. The significance of MTUX extends beyond professional translators. According to Nurminen (2019), a vast majority of MT users, approximately 99.5%, are non-professionals. To date, focus on MT research has been on professional translators, accounting for approximately only the 0.05% of the MT users. This diverse user base includes subtitlers, academics, medical professionals, and others, each with unique needs and expectations from MT interactions. This latter area is where we need to put the attention focus on future research. For instance, a patient whose main language is not English living in Ireland may benefit from the use of MT to communicate with their doctor if there is no interpreting help available. What are the needs of this MT user? These needs will also differ from what the doctor expects of MT. Due to the increased globalisation worldwide and the limited resources of contemporary health systems, MT may form a core component in future health communications (Ugas, Giuliani, and Papadakos 2024). Another example is that of subtitlers. Preparing the subtitles of a TV series requires specific translation tasks related with

segmentation and/or a character limits (García-Escribano and Díaz-Cintas 2023). As a consequence, subtitlers' needs of MT will differ from those of legal translators. Why are these MT needs neglected and not considered when developing new technologies for this use case? One last example may be non-English native academics, who need to publish their research in English because it is the lingua franca of research (Bennett 2014). Some researchers have already started analysing the potential benefits of MT in scholarly communication (Steigerwald et al. 2022). IPE may be a very viable MT modality for specialists in one specific field that know what they want to write, who will be able to receive additional MT help while writing. The MT needs of these academics whose L1 is not English will also be different to the needs of the other use-cases analysed. Consequently, understanding users' specific MT requirements is crucial for personalizing tools, systems, and workflows, thereby enhancing their MTUX, aiming to optimize MT for every user category. There is a whole world of MT users to explore, so that we can develop HCAMT systems, tools and workflows for them.

This PhD underscores the relevance of HCAMT through the study of MTUX and paves the way for making Translation Studies scholars and the translation technology and MT developer communities work together for fostering more human-centered ways of interacting with MT for a wide variety of users and use-cases. The dialogue between these different stakeholders has been largely absent to date, and this research is a first step towards that shift from a perspective that only focuses on productivity and quality, towards a newer perspective that also includes MTUX and users' needs in software development and newer technology adoption. HCAMT through the analysis of MTUX is the way forward in the AI age.

REFERENCES

Alabau, V, J Gonzalez-Rubio, L A Leiva, D Ortiz-Martınez, G Sanchis-Trilles, F Casacuberta, B Mesa-Lao, R Bonk, M Carl, and M Garcıa-Martınez. 2013. 'User Evaluation of Advanced Interaction Features for a Computer-Assisted Translation Workbench', 8.

Alabau, Vicent, Ragnar Bonkb, Christian Buck, Michael Carlb, Francisco Casacuberta Nolla, Mercedes García-Martínez, Jesus Gonzalez Rubio, et al. 2013. 'CASMACAT: An open source workbench for advanced computer aided translation'. In *Prague Bulletin of Mathematical Linguistics*, 100:101–12. De Gruyter Open. https://doi.org/10.2478/pralin-2013-0016.

Alabau, Vicent, Michael Carl, Mercedes García-Martínez, and Jesús González-Rubio. 2016. 'Learning Advanced Post-Editing'. In *New Directions in Empirical Translation Process Research*.

Albarracín, Dolores, ed. 2021. 'The Impact of Past Experience and Past Behavior on Attitudes and Behavior'. In *Action and Inaction in a Social World: Predicting and Changing Attitudes and Behavior*, 129–57. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108878357.006.

Alben, Lauralee. 1996. 'Quality of Experience: Defining the Criteria for Effective Interaction Design'. *Interactions* 3 (3): 11–15. https://doi.org/10.1145/235008.235010.

Albert, Bill, and Tom Tullis. 2022. *Measuring the User Experience: Collecting, Analyzing, and Presenting UX Metrics*. Morgan Kaufmann.

Alicea, Bradly. 2018. 'An Integrative Introduction to Human Augmentation Science'. arXiv. https://doi.org/10.48550/arXiv.1804.10521.

ALPAC. 1966. 'Language and Machines', 138.

Alves, Fabio. 2003. *Triangulating Translation*. *Btl.45*. John Benjamins Publishing Company. https://benjamins.com/catalog/btl.45.

Alves, Fabio, and Arnt Lykke Jakobsen. 2020. 'Grounding Cognitive Translation Studies: Goals, Commitments and Challenges'. In *The Routledge Handbook of Translation and Cognition*, 545–54. Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315178127-35/grounding-cognitive-translation-studies-goals-commitments-challenges-fabio-alves-arnt-lykke-jakobsen.

Alves, Fabio, Karina Sarto Szpak, José Luiz Gonçalves, Kyoko Sekino, Marceli Aquino, Rodrigo Araújo e Castro, Arlene Koglin, Norma B. de Lima Fonseca, and Bartolomé Mesa-Lao. 2016. 'Investigating Cognitive Effort in Post-Editing: A Relevance-Theoretical Approach'. In *Eyetracking and Applied Linguistics*, Silvia Hansen-Schirra&Sambor Grucza, 109–42. Language Science Press. http://langsci-press.org/catalog/book/108.

Arthern, Peter J. 1979. 'MACHINE TRANSLATION AND COMPUTERIZED TERMINOLOGY SYSTEMS A TRANSLATOR'S VIEWPOINT.', 32.

Artstein, Ron. 2017. 'Inter-Annotator Agreement'. In *Handbook of Linguistic Annotation*, edited by Nancy Ide and James Pustejovsky, 297–313. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-024-0881-2_11.

Artstein, Ron, and Massimo Poesio. 2008. 'Inter-Coder Agreement for Computational Linguistics'. *Computational Linguistics* 34 (4): 555–96. https://doi.org/10.1162/coli.07-034-R2.

Association for Computing Machinery. 2020. ‘Computing Curricula 2020’. 2020. https://www.acm.org/binaries/content/assets/education/curricula-recommendations/cc2020.pdf.

Aziz, Wilker, Sheila Castilho, and Lucia Specia. 2012. ‘PET: A Tool for Post-Editing and Assessing Machine Translation.’ In *LREC*, 3982–87. https://aclanthology.org/www.mt-archive.info/10/LREC-2012-Aziz.pdf.

Bargas-Avila, Javier A., and Kasper Hornbæk. 2011. ‘Old Wine in New Bottles or Novel Challenges: A Critical Analysis of Empirical Studies of User Experience’. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2689–98. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1978942.1979336.

Bar-Hillel, Yehoshua. 1959. ‘Report on the State of Machine Translation in the United States and Great Britain’.

Barrachina, Sergio, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, et al. 2009. ‘Statistical Approaches to Computer-Assisted Translation’. *Computational Linguistics* 35 (1): 3–28. https://doi.org/10.1162/coli.2008.07-055-R2-06-29.

Bender, Oliver, Sas˘a Hasan, David Vilar, Richard Zens, and Hermann Ney. 2005. ‘Comparison of Generation Strategies for Interactive Machine Translation’. *Conference Proceedings*, 8.

Bennett, Karen. 2014. ‘English as a Lingua Franca in Academia: Combating Epistemicide through Translator Training’. In *English as a Lingua Franca*. Routledge.

Bentivogli, Luisa, Nicola Bertoldi, Mauro Cettolo, Marcello Federico, Matteo Negri, and Marco Turchi. 2016. ‘On the Evaluation of Adaptive Machine Translation for Human Post-Editing’. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24 (2): 388–99. https://doi.org/10.1109/TASLP.2015.2509241.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. ‘Neural versus Phrase-Based Machine Translation Quality: A Case Study’. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 257–67. Austin, Texas: Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1025.

Bevan, Nigel, Zhengjie Liu, Cathy Barnes, Marc Hassenzahl, and Weijie Wei. 2016. ‘Comparison of Kansei Engineering and AttrakDiff to Evaluate Kitchen Products’. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2999–3005. CHI EA ’16. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2851581.2892407.

Blackwell, Alan F. 2015. ‘HCI as an Inter-Discipline’. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 503–16. Seoul Republic of Korea: ACM. https://doi.org/10.1145/2702613.2732505.

Blanchon, Herve. 1994. ‘Perspectives of DBMT for Monolingual Authors on the Basis of LIDIA-1, an Implemented Mock-Up’. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*. https://www.aclweb.org/anthology/C94-1017.

Blythe, Mark, and Andrew Monk, eds. 2018. *Funology 2: From Usability to Enjoyment*. 2nd ed. Human–Computer Interaction Series. Springer International Publishing. https://doi.org/10.1007/978-3-319-68213-6.

Bojar, Ondřej, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. 'Findings of the 2018 Conference on Machine Translation (WMT18)'. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 272–303. Belgium, Brussels: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6401.

Borja, Anabel. 2000. *El texto jurídico inglés y su traducción al español*. Grupo Planeta (GBS).

———. 2013. 'A Genre Analysis Approach to the Study of the Translation of Court Documents'. *Linguistica Antverpiensia, New Series – Themes in Translation Studies* 0 (12). https://lans-tts.uantwerpen.be/index.php/LANS-TTS/article/view/235.

Bowker, Lynne. 2015. 'Translatability and User eXperience: Compatible or in Conflict?' *Localisation Focus* 14 (January):13–27.

Bowker, Lynne, and Jairo Buitrago Ciro. 2018. 'Localizing Websites Using Machine Translation: Exploring Connections between User eXperience and Translatability'. In *The Human Factor in Machine Translation*. Routledge.

———. 2019. *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Emerald Publishing Limited. https://www.emerald.com/insight/content/doi/10.1108/978-1-78756-721-420191009/full/html.

Bowker, Lynne, and Des Fisher. 2010. 'Computer-Aided Translation'. In *Handbook of Translation Studies*, 1:60–66.

Braun, Virginia, and Victoria Clarke. 2006. 'Using Thematic Analysis in Psychology'. *Qualitative Research in Psychology* 3 (2): 77–101. https://doi.org/10.1191/1478088706qp063oa.

Briva-Iglesias, Vicent. 2021. 'Traducción humana vs. traducción automática: análisis contrastivo e implicaciones para la aplicación de la traducción automática en traducción jurídica'. *Mutatis Mutandis. Revista Latinoamericana de Traducción* 14 (2): 571–600. https://doi.org/10.17533/udea.mut.v14n2a14.

———. 2023. 'Translation Technologies Advancements: From Inception to the Automation Age'. In *La Família Humana: Perspectives Multidisciplinàries de La Investigació En Ciències Humanes i Socials*, Lucía Bellés-Calvera; María Pallarés-Renau, 137–52. Emergents 3. Publicacions de la Universitat Jaume I. Servei de Comunicació i Publicacions.

Briva-Iglesias, Vicent, Joao Lucas Cavalheiro Camargo, and Gokhan Dogru. 2024. 'Large Language Models "Ad Referendum": How Good Are They at Machine Translation in the Legal Domain?' *MonTI. Monografías de Traducción e Interpretación* 14 (February). https://doi.org/10.48550/arXiv.2402.07681.

Briva-Iglesias, Vicent, and Sharon O'Brien. 2022. 'The Language Engineer: A Transversal, Emerging Role for the Automation Age'. *Quaderns de Filologia - Estudis Lingüístics* 27 (0): 17–48. https://doi.org/10.7203/qf.0.24622.

———. 2023. 'Measuring Machine Translation User Experience: A Comparison between AttrakDiff and User Experience Questionnaire'. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 335–44.

———. 2024. 'Pre-Task Perceptions of MT Influence Quality and Productivity: The Importance of Better Translator-Computer Interactions and Implications for Training'. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation*. https://www.researchgate.net/publication/381375611_Pre-task_perceptions_of_MT_influence_quality_and_productivity_the_importance_of_better_translator-computer_interactions_and_implications_for_training.

Briva-Iglesias, Vicent, Sharon O'Brien, and Benjamin R. Cowan. 2023. 'The Impact of Traditional and Interactive Post-Editing on Machine Translation User Experience, Quality, and Productivity': *Translation, Cognition & Behavior* 6 (1). https://doi.org/10.1075/tcb.00077.bri.

Brousseau, Julie, Caroline Drouin, George Foster, Pierre Isabelle, Roland Kuhn, Yves Normandin, and Pierre Plamondon. 1995. 'French Speech Recognition in an Automatic Dictation System for Translators: The TransTalk Project', 10.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. 'A Statistical Approach to Machine Translation'. *Computational Linguistics* 16 (2): 79–85.

Brown, Ralf D., and Sergei Nirenburg. 1990. 'Human-Computer Interaction for Semantic Disambiguation'. In *COLING 1990 Volume 3: Papers Presented to the 13th International Conference on Computational Linguistics*. https://www.aclweb.org/anthology/C90-3008.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. 'Language Models Are Few-Shot Learners'. arXiv. https://doi.org/10.48550/arXiv.2005.14165.

Bulté, Bram, and Arda Tezcan. 2019. 'Neural Fuzzy Repair : Integrating Fuzzy Matches into Neural Machine Translation'. In *57TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL 2019)*, 1800–1809. https://doi.org/10.18653/v1/P19-1175.

Bundgaard, Kristine. 2017. 'Translator Attitudes towards Translator-Computer Interaction– Findings from a Workplace Study'. *Hermes*, no. 56, 125–44.

Burchardt, Aljoscha, and Arle Lommel. 2014. 'Practical Guidelines for the Use of MQM in Scientific Research on Translation Quality'. 2014. http://www.qt21.eu/downloads/MQM-usage-guidelines.pdf.

Cadwell, Patrick, Sheila Castilho, Sharon O'Brien, and Linda Mitchell. 2016. 'Human Factors in Machine Translation and Post-Editing among Institutional Translators'. *Translation Spaces* 5 (2): 222–43. https://doi.org/10.1075/ts.5.2.04cad.

Cadwell, Patrick, Sharon O'Brien, and Carlos S. C. Teixeira. 2018. 'Resistance and Accommodation: Factors for the (Non-) Adoption of Machine Translation among Professional Translators'. *Perspectives* 26 (3): 301–21. https://doi.org/10.1080/0907676X.2017.1337210.

Cambridge Dictionary. 2024. 'Automation'. 29 May 2024. https://dictionary.cambridge.org/dictionary/english/automation.

Carl, Michael. 2012. 'Translog-II: A Program for Recording User Activity Data for Empirical Translation Process Research'. https://research.cbs.dk/files/58900336/Michael_Carl_2012.pdf.

Carl, Michael, Srinivas Bangalore, and Moritz Schaeffer, eds. 2016. *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*. New Frontiers in Translation Studies. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-20358-4.

Carl, Michael, Barbara Dragsted, Jakob Elming, Daniel Hardt, and Arnt Lykke Jakobsen. 2011. 'The Process of Post-Editing: A Pilot Study'. *Copenhagen Studies in Language* 41 (1): 131–42.

Carroll, John M., and Mary Beth Rosson. 1992. 'Getting around the Task-Artifact Cycle: How to Make Claims and Design by Scenario'. *ACM Transactions on Information Systems* 10 (2): 181–212. https://doi.org/10.1145/146802.146834.

Caruana, Edward Joseph, Marius Roman, Jules Hernández-Sánchez, and Piergiorgio Solli. 2015. 'Longitudinal Studies'. *Journal of Thoracic Disease* 7 (11): E537–40. https://doi.org/10.3978/j.issn.2072-1439.2015.10.63.

Casacuberta, Francisco, Jorge Civera, Elsa Cubel, Antonio L. Lagarda, Guy Lapalme, Elliott Macklovitch, and Enrique Vidal. 2009. 'Human Interaction for High-Quality Machine Translation'. *Communications of the ACM* 52 (10): 135–38. https://doi.org/10.1145/1562764.1562798.

Castaño, M Asunción, Francisco Casacuberta, and Enrique Vidal. 1997. 'Machine Translation Using Neural Networks and Finite-State Models', 8.

Castilho, Sheila. 2016. 'Measuring Acceptability of Machine Translated Enterprise Content'. *Castilho, Sheila ORCID: 0000-0002-8416-6555 <https://Orcid.Org/0000-0002-8416-6555> (2016) Measuring Acceptability of Machine Translated Enterprise Content. PhD Thesis, Dublin City University.* Doctoral, Dublin City University. Faculty of Humanities and Social Science. http://doras.dcu.ie/21342/.

———. 2021. 'Towards Document-Level Human MT Evaluation: On the Issues of Annotator Agreement, Effort and Misevaluation'. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, 34–45. Online: Association for Computational Linguistics. https://aclanthology.org/2021.humeval-1.4.

Castilho, Sheila, Clodagh Mallon, Rahel Meister, and Shengya Yue. 2023. 'Do Online Machine Translation Systems Care for Context? What about a GPT Model?' In . Tampere, Finland. https://events.tuni.fi/eamt23/.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. 'Is Neural Machine Translation the New State of the Art?' *The Prague Bulletin of Mathematical Linguistics* 108 (1): 109–20. https://doi.org/10.1515/pralin-2017-0013.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, and Andy Way. 2017. 'A Comparative Quality Evaluation of PBSMT and NMT Using Professional Translators', 16.

Castilho, Sheila, and Sharon O'Brien. 2017. 'Acceptability of Machine-Translated Content: A Multi-Language Evaluation by Translators and End-Users'. *Linguistica Antverpiensia, New Series: Themes in Translation Studies* 16:120–36.

Castilho, Sheila, Sharon O'Brien, Fabio Alves, and Morgan O'Brien. 2014. *Does Post-Editing Increase Usability? A Study with Brazilian Portuguese as Target Language*.

Castilho, Sheila, and Natália Resende. 2022. 'Post-Editese in Literary Translations'. *Information* 13 (2): 66. https://doi.org/10.3390/info13020066.

Cettolo, Mauro, Jan Niehues, Sebastian Stuker, Luisa Bentivogli, and Marcello Federico. 2013. 'Report on the 10th IWSLT Evaluation Campaign', 18.

Chan, Sin-Wai. 2014. *Routledge Encyclopedia of Translation Technology*. 1st ed. Routledge. https://doi.org/10.4324/9781315749129.

Church, Kenneth W, and Eduard H Hovy. 1993. 'Good Applications for Crummy Machine Translation', 20.

Clark, Leigh, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, et al. 2019. 'The State of Speech in HCI: Trends, Themes and Challenges'. *Interacting with Computers* 31 (4): 349–71. https://doi.org/10.1093/iwc/iwz016.

Cockton, Gilbert. 2002. 'From Doing to Being: Bringing Emotion into Interaction'. *Interacting with Computers*. Oxford University Press Oxford, UK. https://academic.oup.com/iwc/article-abstract/14/2/89/758850.

Cowan, Benjamin Richard. 2011. 'Causal Effects of Wiki Site Design on Anxiety and Usability'.

Creswell, John W., and Vicki L. Plano Clark. 2007. *Designing and Conducting Mixed Methods Research*. Designing and Conducting Mixed Methods Research. Thousand Oaks, CA, US: Sage Publications, Inc.

Daems, Joke, Orphée De Clercq, and Lieve Macken. 2017. 'Translationese and Post-Editese : How Comparable Is Comparable Quality?' *LINGUISTICA ANTVERPIENSIA NEW SERIES-THEMES IN TRANSLATION STUDIES* 16:89–103.

Daems, Joke, and Lieve Macken. 2019. 'Interactive Adaptive SMT versus Interactive Adaptive NMT: A User Experience Evaluation'. *Machine Translation* 33 (1): 117–34. https://doi.org/10.1007/s10590-019-09230-z.

Daems, Joke, Sonia Vandepitte, Robert J. Hartsuiker, and Lieve Macken. 2017. 'Identifying the Machine Translation Error Types with the Greatest Impact on Post-Editing Effort'. *Frontiers in Psychology* 8:1282. https://doi.org/10.3389/fpsyg.2017.01282.

Deng, Xinjie, and Zhonggen Yu. 2022. 'A Systematic Review of Machine-Translation-Assisted Language Learning for Sustainable Education'. *Sustainability* 14 (13): 7598. https://doi.org/10.3390/su14137598.

Denkowski, Michael, Alon Lavie, Isabel Lacruz, and Chris Dyer. 2014. 'Real Time Adaptive Machine Translation for Post-Editing with Cdec and TransCenter'. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-Assisted Translation*, 72–77. Gothenburg, Sweden: Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-0311.

Desmet, Pieter. 2002. *Designing Emotions*.

Diener, Ed, Eunkook M. Suh, Richard E. Lucas, and Heidi L. Smith. 1999. 'Subjective Well-Being: Three Decades of Progress.' *Psychological Bulletin* 125 (2): 276.

Diggle, Peter, Department of Mathematics and Statistics Peter J. Diggle, Peter J. Diggle, Patrick Heagerty, Kung-Yee Liang, Patrick J. Heagerty, Scott Zeger, and Both at Biostatistics Department Scott Zeger. 2002. *Analysis of Longitudinal Data*. OUP Oxford.

Dillinger, Mike, and Arle Lommel. 2004. *Implementing Machine Translation*. Localization Industry Standards Association. https://scholar.google.com/scholar?cluster=8850355502229876167&hl=en&oi=scholarr.

Dix, Alan. 2003. *Human-Computer Interaction*. Pearson Education.

———. 2010. 'Human–Computer Interaction: A Stable Discipline, a Nascent Science, and the Growth of the Long Tail'. *Interacting with Computers* 22 (1): 13–27. https://doi.org/10.1016/j.intcom.2009.11.007.

Doherty, Neil, and Malcolm King. 2005. 'From Technical to Socio-Technical Change: Tackling the Human and Organizational Aspects of Systems Development Projects'. *European Journal of Information Systems* 14 (March):1–5. https://doi.org/10.1057/palgrave.ejis.3000517.

Doherty, Stephen, and Sharon O'Brien. 2014. 'Assessing the Usability of Raw Machine Translated Output: A User-Centered Study Using Eye Tracking'. *International Journal of Human-Computer Interaction* 30 (1): 40–51. https://doi.org/10.1080/10447318.2013.802199.

Ehrensberger-Dow, Maureen. 2014. ‘Challenges of Translation Process Research at the Workplace’. *MonTI. Monografías de Traducción e Interpretación*, 355–83. https://doi.org/10.6035/MonTI.2014.ne1.12.

———. 2017. ‘An Ergonomic Perspective of Translation’. In *The Handbook of Translation and Cognition*, 332–49. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119241485.ch18.

———. 2020. ‘Translation, Ergonomics and Cognition’. In *The Routledge Handbook of Translation and Cognition*. Routledge.

Ehrensberger-Dow, Maureen, and Andrea Hunziker Heeb. 2016. ‘Investigating the Ergonomics of a Technologized Translation Workplace’. *Reembedding Translation Process Research*. https://www.torrossa.com/gs/resourceProxy?an=5015995&publisher=FZ4850#page=76.

Ehrensberger-Dow, Maureen, Andrea Hunziker Heeb, Gary Massey, Ursula Meidert, Silke Neumann, and Heidrun Becker. 2016. ‘An International Survey of the Ergonomics of Professional Translation’. *ILCEA*, no. 27 (November). https://doi.org/10.4000/ilcea.4004.

Ehrensberger-Dow, Maureen, and Gary Massey. 2014. ‘Cognitive Ergonomic Issues in Professional Translation’. *The Development of Translation Competence : Theories and Methodologies from Psycholinguistics and Cognitive Science*, 58–86.

Ehrensberger-Dow, Maureen, and Sharon O’Brien. 2015. ‘Ergonomics of the Translation Workplace’. *Translation Spaces* 4 (1): 98–118. https://doi.org/10.1075/ts.4.1.05ehr.

ELIS. 2022. ‘EUROPEAN LANGUAGE INDUSTRY SURVEY 2022’, 44.

ELIS ELIS Research. 2023. ‘EUROPEAN LANGUAGE INDUSTRY SURVEY 2023’.

Elliston, John S. G. 1978. ‘Computer Aided Translation: A Business Viewpoint’. https://doi.org/10.7551/mitpress/5779.003.0031.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. ‘GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models’. arXiv. https://doi.org/10.48550/arXiv.2303.10130.

Engelbart, Douglas C. 1962. *Augmenting Human Intellect: A Conceptual Framework*. http://archive.org/details/1962-engelbart-AHI-framework.

Epp, Clayton, Michael Lippold, and Regan L. Mandryk. 2011. ‘Identifying Emotional States Using Keystroke Dynamics’. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 715–24. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1978942.1979046.

Escartín, Carla Parra, Sharon O’Brien, Marie-Josée Goulet, and Michel Simard. 2017. ‘Machine Translation as an Academic Writing Aid for Medical Practitioners’, 14.

Esteban, José, José Lorenzo, Antonio S. Valderrábanos, and Guy Lapalme. 2004. ‘TransType2: An Innovative Computer-Assisted Translation System’. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions* -, 1-es. Barcelona, Spain: Association for Computational Linguistics. https://doi.org/10.3115/1219044.1219045.

Etchegoyhen, Thierry, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard van Loenhout, Arantza del Pozo, Mirjam Sepesy Maučec, Anja Turner, and Martin Volk. 2014. ‘Machine Translation for Subtitling: A Large-Scale Evaluation’. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, 46–53. Reykjavik, Iceland: European Language Resources

Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/463_Paper.pdf.

Etchegoyhen, Thierry, Anna Fernández Torné, Andoni Azpeitia, Eva Martínez Garcia, and Anna Matamala. 2018. 'Evaluating Domain Adaptation for Machine Translation Across Scenarios'. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L18-1002.

European Union. 2024. *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

Faulkner, Xris. 2000. *Usability Engineering*. Macmillan Education UK.

Federico, Marcello, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, et al. 2014. 'The MateCat Tool'. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, 129–32. Dublin, Ireland: Dublin City University and Association for Computational Linguistics. https://www.aclweb.org/anthology/C14-2028.

Fields, Bob, Paola Amaldi, William Wong, and Satinder Gill. 2007. 'In Use, In Situ: Extending Field Research Methods'. *International Journal of Human-Computer Interaction* 22 (1–2): 1–6. https://doi.org/10.1080/10447310709336952.

———. 2008. 'Introduction: In-Use, In-Situ: Extending Field Research Methods—Part 2'. *International Journal of Human–Computer Interaction* 24 (4): 359–60. https://doi.org/10.1080/10447310801991921.

Fincher, Sally, and Marian Petre. 2004. *Computer Science Education Research*. https://www.routledge.com/Computer-Science-Education-Research/Fincher-Petre/p/book/9780367604530.

Fırat, Gökhan. 2021. 'Uberization of Translation: Impacts on Working Conditions'. *The Journal of Internationalization and Localization* 8 (1): 48–75. https://doi.org/10.1075/jial.20006.fir.

Forcada, Mikel L. 2017. 'Making Sense of Neural Machine Translation'. *Translation Spaces* 6 (2): 291–309. https://doi.org/10.1075/ts.6.2.06for.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. 'Apertium: A Free/Open-Source Platform for Rule-Based Machine Translation'. *Machine Translation* 25 (2): 127–44. https://doi.org/10.1007/s10590-011-9090-0.

Forcada, Mikel L., and Ramón P. Ñeco. 1997. 'Recursive Hetero-Associative Memories for Translation'.

Forlizzi, Jodi, and Katja Battarbee. 2004. 'Understanding Experience in Interactive Systems'. In *Proceedings of the 5th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, 261–68. DIS '04. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1013115.1013152.

Foster, George, Pierre Isabelle, and Pierre Plamondon. 1997. 'Target-Text Mediated Interactive Machine Translation', 20.

Freitag, Markus, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. 'Experts, Errors, and Context: A Large-Scale Study of Human

Evaluation for Machine Translation'. *arXiv:2104.14478 [Cs]*, April. http://arxiv.org/abs/2104.14478.

Galison, Peter. 1997. *Image and Logic*. https://press.uchicago.edu/ucp/books/book/chicago/I/bo3710110.html.

Ganglbauer, Eva, Johann Schrammel, Stephanie Deutsch, and Manfred Tscheligi. 2009. 'Applying Psychophysiological Methods for Measuring User Experience'. In *Possibilities, Challenges, and Feasibility. Proc. User Experience Evaluation Methods in Product Development Workshop*.

Garcia, Ignacio. 2010. 'Is Machine Translation Ready Yet?' *Target. International Journal of Translation Studies* 22 (1): 7–21. https://doi.org/10.1075/target.22.1.02gar.

———. 2014. 'Computer-Aided Translation: Systems'. In *Routledge Encyclopedia of Translation Technology*.

Garcia, Ignacio, and Vivian Stevenson. 2005. 'TRADOS and the Evolution of Language Tools: The Rise of the De Facto TM Standard - And Its Future with SDL'. *Multilingual Computing and Technology* 16 (7).

García-Escribano, Alejandro Bolaños, and Jorge Díaz-Cintas. 2023. 'Integrating Post-Editing into the Subtitling Classroom: What Do Subtitlers-to-Be Think?' *Linguistica Antverpiensia, New Series–Themes in Translation Studies* 22. https://artojs01.uantwerpen.be/index.php/LANS-TTS/article/view/777.

Gaspari, Federico, Antonio Toral, Sudip Kumar Naskar, Declan Groves, and Andy Way. 2014. 'Perception vs Reality: Measuring Machine Translation Post-Editing Productivity'. In *Third Workshop on Post-Editing Technology and Practice*. Vol. 60.

Gaver, Bill, and Heather Martin. 2000. 'Alternatives: Exploring Information Appliances through Conceptual Design Proposals'. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 209–16. The Hague The Netherlands: ACM. https://doi.org/10.1145/332040.332433.

Ginestí, Mireia, and Mikel L. Forcada. 2009. 'LA TRADUCCIÓ AUTOMÀTICA EN LA PRÀCTICA: APLICACIONS, DIFICULTATS I ESTRATÈGIES DE DESENVOLUPAMENT', 18.

Görög, Attila. 2014. 'Dynamic Quality Framework: Quantifying and Benchmarking Quality'. *Tradumàtica Tecnologies de La Traducció*, no. 12 (December), 443–54. https://doi.org/10.5565/rev/tradumatica.66.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. 'Continuous Measurement Scales in Human Evaluation of Machine Translation', 9.

Graham, Yvette, Barry Haddow, and Philipp Koehn. 2020. 'Statistical Power and Translationese in Machine Translation Evaluation'. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, edited by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, 72–81. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.6.

Green, Spence, Jason Chuang, Jeffrey Heer, and Christopher D. Manning. 2014. 'Predictive Translation Memory: A Mixed-Initiative System for Human Language Translation'. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, 177–87. Honolulu Hawaii USA: ACM. https://doi.org/10.1145/2642918.2647408.

Green, Spence, Jeffrey Heer, and Christopher D. Manning. 2013. 'The Efficacy of Human Post-Editing for Language Translation'. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 439–48. CHI '13. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2470654.2470718.

Green, Spence, Sida I. Wang, Jason Chuang, Jeffrey Heer, Sebastian Schuster, and Christopher D. Manning. 2014. 'Human Effort and Machine Learnability in Computer Aided Translation'. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1225–36. Doha, Qatar: Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1130.

Guerberof, Ana. 2013. 'What Do Professional Translators Think about Post-Editing?' 2013. https://www.jostrans.org/issue19/art_guerberof.pdf.

Guerberof Arenas, Ana. 2008. 'Productivity and Quality in the Post-Editing of Outputs from Translation Memories and Machine Translation'. *Localisation Focus The International Journal of Localisation* 7 (1): 11–21.

Guerberof Arenas, Ana, Joss Moorkens, and Sharon O'Brien. 2021. 'The Impact of Translation Modality on User Experience: An Eye-Tracking Study of the Microsoft Word User Interface'. *Machine Translation* 35 (2): 205–37. https://doi.org/10.1007/s10590-021-09267-z.

Guerberof-Arenas, Ana, and Antonio Toral. 2022. 'Creativity in Translation: Machine Translation as a Constraint for Literary Texts'. *Translation Spaces* 11 (2): 184–212. https://doi.org/10.1075/ts.21025.gue.

Hacker, Philipp, Andreas Engel, and Marco Mauer. 2023. 'Regulating ChatGPT and Other Large Generative AI Models'. arXiv. https://doi.org/10.48550/arXiv.2302.02337.

Hancock, Peter A., Aaron A. Pepe, and Lauren L. Murphy. 2005. 'Hedonomics: The Power of Positive and Pleasurable Ergonomics'. *Ergonomics in Design* 13 (1): 8–14. https://doi.org/10.1177/106480460501300104.

Hassenzahl, Marc. 2018. 'The Thing and I: Understanding the Relationship Between User and Product'. In *Funology 2*, edited by Mark Blythe and Andrew Monk, 301–13. Human–Computer Interaction Series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-68213-6_19.

Hassenzahl, Marc, Andreas Beu, and Michael Burmester. 2001. 'Engineering Joy'. *IEEE SOFTWARE*, 7.

Hassenzahl, Marc, Michael Burmester, and Franz Koller. 2003. 'AttrakDiff: Ein Fragebogen Zur Messung Wahrgenommener Hedonischer Und Pragmatischer Qualität'. *Mensch & Computer 2003: Interaktion in Bewegung*, 187–96.

Hassenzahl, Marc, and Noam Tractinsky. 2006. 'User Experience - a Research Agenda'. *Behaviour & Information Technology* 25 (2): 91–97. https://doi.org/10.1080/01449290500330331.

Hazlett, Richard L. 2006. 'Measuring Emotional Valence during Interactive Experiences: Boys at Video Game Play'. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1023–26. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1124772.1124925.

Hearne, Mary, and Andy Way. 2011. 'Statistical Machine Translation: A Guide for Linguists and Translators'. *Language and Linguistics Compass* 5 (5): 205–26. https://doi.org/10.1111/j.1749-818X.2011.00274.x.

Hewett, Thomas T., Ronald Baecker, Stuart Card, Tom Carey, Jean Gasen, Marilyn Mantei, Gary Perlman, Gary Strong, and William Verplank. 1992. 'ACM SIGCHI Curricula for Human-Computer Interaction'. Technical Report. New York, NY, USA: Association for Computing Machinery.

Hickey, Sarah. 2023. ‘Language Services Verticals: Market Sizing in 2023’. *Nimdzi* (blog). 15 March 2023. https://www.nimdzi.com/language-industry-verticals-market-size-by-segment-leaders/.

Horvitz, Eric. 1999. ‘Principles of Mixed-Initiative User Interfaces’. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 159–66. CHI ’99. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/302979.303030.

ISO. 2018. ‘ISO 9241-11:2018(En), Ergonomics of Human-System Interaction — Part 11: Usability: Definitions and Concepts’. 2018. https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en.

Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. ‘Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine’. arXiv. https://doi.org/10.48550/arXiv.2301.08745.

Johnson, R. Burke, Anthony J. Onwuegbuzie, and Lisa A. Turner. 2007. ‘Toward a Definition of Mixed Methods Research’. *Journal of Mixed Methods Research* 1 (2): 112–33. https://doi.org/10.1177/1558689806298224.

Jordan, Patrick W. 2003. *How to Make Brilliant Stuff That People Love... And Make Big Money out of It*. John Wiley & Sons. https://books.google.com/books?hl=en&lr=&id=dNNLhDPoB2EC&oi=fnd&pg=PR5&dq=info:nPMk1aF-aRMJ:scholar.google.com&ots=2xUBrgTCe0&sig=ERRNGH8cmheNKZcKnodRAhafVGs.

Kankainen, Anu, and Jane Suri. 2001. ‘Supporting Users’ Creativity: Design to Induce Pleasurable Experiences’. *Proceedings of the International Conference on Affective Human Factors Design*, January, 387–94.

Karakanta, Alina, Luisa Bentivogli, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2022. ‘Post-Editing in Automatic Subtitling: A Subtitlers’ Perspective’. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, 261–70. Ghent, Belgium: European Association for Machine Translation. https://aclanthology.org/2022.eamt-1.29.

Kasneci, Enkelejda, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, et al. 2023. ‘ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education’. *Learning and Individual Differences* 103 (April):102274. https://doi.org/10.1016/j.lindif.2023.102274.

Kay, Martin. 1970. ‘The MIND System’. *Computer Science*.

Kaye, Joseph ‘Jofish’. 2007. ‘Evaluating Experience-Focused HCI’. In *CHI ’07 Extended Abstracts on Human Factors in Computing Systems*, 1661–64. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1240866.1240877.

Kenny, Dorothy. 2022. ‘Human and Machine Translation’. *Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence* 18:23.

Kim, Young Jin, and Hoon Sik Yoo. 2021. ‘Analysis of User Preference of AR Head-Up Display Using Attrakdiff’. In *Intelligent Human Computer Interaction*, edited by Madhusudan Singh, Dae-Ki Kang, Jong-Ha Lee, Uma Shanker Tiwary, Dhananjay Singh, and Wan-Young Chung, 335–45. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-68452-5_35.

Koby, Geoffrey S., Paul Fields, Daryl R. Hague, Arle Lommel, and Alan Melby. 2014. 'Defining Translation Quality'. *Tradumàtica Tecnologies de La Traducció*, no. 12 (December), 413–20. https://doi.org/10.5565/rev/tradumatica.76.

Kocmi, Tom, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, et al. 2022. 'Findings of the 2022 Conference on Machine Translation (WMT22)'. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, edited by Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, et al., 1–45. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics. https://aclanthology.org/2022.wmt-1.1.

Kocmi, Tom, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. 'To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation'. In *Proceedings of the Sixth Conference on Machine Translation*, 478–94. Online: Association for Computational Linguistics. https://aclanthology.org/2021.wmt-1.57.

Koehn, Philipp. 2009a. 'A Process Study of Computer-Aided Translation'. *Machine Translation* 23 (4): 241–63. https://doi.org/10.1007/s10590-010-9076-3.

———. 2009b. 'A Web-Based Interactive Computer Aided Translation Tool'. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, 17–20. Suntec, Singapore: Association for Computational Linguistics. https://www.aclweb.org/anthology/P09-4005.

———. 2010. 'Statistical Machine Translation', 447.

———. 2017. 'Neural Machine Translation'. *arXiv:1709.07809 [Cs]*, September. http://arxiv.org/abs/1709.07809.

Koehn, Philipp, and Ulrich Germann. 2014. 'The Impact of Machine Translation Quality on Human Post-Editing'. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-Assisted Translation*, 38–46. Gothenburg, Sweden: Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-0307.

Koehn, Philipp, and Barry Haddow. 2009. 'Interactive Assistance to Human Translators Using Statistical Machine Translation Methods', 8.

Koehn, Philipp, and Christof Monz. 2006. 'Manual and Automatic Evaluation of Machine Translation between European Languages'. In *Proceedings of the Workshop on Statistical Machine Translation - StatMT '06*, 102. New York City, New York: Association for Computational Linguistics. https://doi.org/10.3115/1654650.1654666.

Koehn, Philipp, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst, Hieu Hoang, et al. 2007. 'Moses: Open Source Toolkit for Statistical Machine Translation'. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions - ACL '07*, 177. Prague, Czech Republic: Association for Computational Linguistics. https://doi.org/10.3115/1557769.1557821.

Koponen, Maarit. 2016. 'Is Machine Translation Post-Editing Worth the Effort? A Survey of Research into Post-Editing and Effort'. *The Journal of Specialised Translation* 25 (2). https://www.jostrans.org/issue25/art_koponen.pdf.

Koponen, Maarit, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020. 'MT for Subtitling: Investigating Professional Translators' User Experience and Feedback'. In *Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation*, edited by John E. Ortega, Marcello Federico, Constantin Orasan, and Maja Popovic, 79–92.

Virtual: Association for Machine Translation in the Americas. https://aclanthology.org/2020.amta-pemdt.6.

Kosmaczewska, Kasia, and Matt Train. 2019. 'Application of Post-Edited Machine Translation in Fashion eCommerce'. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, edited by Mikel Forcada, Andy Way, John Tinsley, Dimitar Shterionov, Celia Rico, and Federico Gaspari, 167–73. Dublin, Ireland: European Association for Machine Translation. https://aclanthology.org/W19-6730.

Kovacs, Geza. 2020. 'Predictive Translation Memory in the Wild: A Study of Interactive Machine Translation Use on Lilt', 65.

Krings, Hans P. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Kent State University Press.

Kujala, Sari, Virpi Roto, Kaisa Väänänen-Vainio-Mattila, Evangelos Karapanos, and Arto Sinnelä. 2011. 'UX Curve: A Method for Evaluating Long-Term User Experience'. *Interacting with Computers* 23 (5): 473–83. https://doi.org/10.1016/j.intcom.2011.06.005.

Lagoudaki, Elina. 2008. 'The Value of Machine Translation for the Professional Translator'. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas: Student Research Workshop*, 262–69. Waikiki, USA: Association for Machine Translation in the Americas. https://aclanthology.org/2008.amta-srw.4.

Langlais, Philippe, and George Foster. 2000. 'Using Context-Dependent Interpolation to Combine Statistical Language and Translation Models for Interactive Machine Translation', 13.

Langlais, Philippe, George Foster, and Guy Lapalme. 2000. 'TransType: A Computer-Aided Translation Typing System'. In *ANLP-NAACL 2000 Workshop: Embedded Machine Translation Systems*. https://www.aclweb.org/anthology/W00-0507.

Langlais, Philippe, Guy Lapalme, and Marie Loranger. 2002. 'TransType: Development-Evaluation Cycles to Boost Translator's Productivity'. *Machine Translation* 17 (2): 77–98.

Langlais, Philippe, Marie Loranger, and Guy Lapalme. 2002. 'Translators at Work with TRANSTYPE: Resource and Evaluation', 8.

Langlais, Philippe, Sebastien Sauve, George Foster, Elliott Macklovitch, and Guy Lapalme. 2000. 'Evaluation of TRANSTYPE, a Computer-Aided Translation Typing System: A Comparison of a Theoretical- and a User- Oriented Evaluation Procedures', 8.

Läubli, Samuel, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. 'A Set of Recommendations for Assessing Human–Machine Parity in Language Translation'. *Journal of Artificial Intelligence Research* 67 (March):653–72. https://doi.org/10.1613/jair.1.11371.

Läubli, Samuel, Rico Sennrich, and Martin Volk. 2018. 'Has Machine Translation Achieved Human Parity? A Case for Document-Level Evaluation'. arXiv. http://arxiv.org/abs/1808.07048.

Laugwitz, Bettina, Theo Held, and Martin Schrepp. 2008. 'Construction and Evaluation of a User Experience Questionnaire'. *International Journal of Interactive Multimedia and Artificial Intelligence* 4 (4): 76. https://doi.org/10.1007/978-3-540-89350-9_6.

Law, Effie Lai-Chong, Virpi Roto, Marc Hassenzahl, Arnold P.O.S. Vermeeren, and Joke Kort. 2009. 'Understanding, Scoping and Defining User Experience: A Survey Approach'. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 719–

28. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1518701.1518813.

Law, Effie Lai-Chong, Paul van Schaik, and Virpi Roto. 2014. 'Attitudes towards User Experience (UX) Measurement'. *International Journal of Human-Computer Studies*, Interplay between User Experience Evaluation and System Development, 72 (6): 526–41. https://doi.org/10.1016/j.ijhcs.2013.09.006.

Lewis, James R. 1995. 'IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use'. *International Journal of Human-Computer Interaction* 7 (1): 57–78. https://doi.org/10.1080/10447319509526110.

Licklider, J. C. R. 1960. 'Man-Computer Symbiosis'. 1960. https://groups.csail.mit.edu/medg/people/psz/Licklider.html.

Lin, Chin-Yew, and Franz Josef Och. 2004. 'ORANGE: A Method for Evaluating Automatic Evaluation Metrics for Machine Translation'. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 501–7. Geneva, Switzerland: COLING. https://www.aclweb.org/anthology/C04-1072.

Lin, Dekang. 1996. 'On the Structural Complexity of Natural Language Sentences'. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*. https://www.aclweb.org/anthology/C96-2123.

Ling, Wang, Isabel Trancoso, Chris Dyer, and Alan W. Black. 2015. 'Character-Based Neural Machine Translation'. *arXiv:1511.04586 [Cs]*, November. http://arxiv.org/abs/1511.04586.

Liu, Yong, Jorge Goncalves, Denzil Ferreira, Bei Xiao, Simo Hosio, and Vassilis Kostakos. 2014. 'CHI 1994-2013: Mapping Two Decades of Intellectual Progress through Co-Word Analysis'. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3553–62. Toronto Ontario Canada: ACM. https://doi.org/10.1145/2556288.2556969.

Lloyd, G. E. R. 2009. *Disciplines in the Making: Cross-Cultural Perspectives on Elites, Learning, and Innovation*. OUP Oxford.

Loewenstein, George, and Jennifer S. Lerner. 2003. 'The Role of Affect in Decision Making'. *Handbook of Affective Science* 619 (642): 3.

Lommel, Arle, and Alan Melby. 2014. 'Multidimensional Quality Metrics Definition'. 2014. http://www.qt21.eu/mqm-definition/issues-list-2014-08-19.html.

Lyu, Chenyang, Jitao Xu, and Longyue Wang. 2023. 'New Trends in Machine Translation Using Large Language Models: Case Examples with ChatGPT'. arXiv. https://doi.org/10.48550/arXiv.2305.01181.

Macías, Lorena Pérez. 2020. 'What Do Translators Think About Post-Editing? : A Mixed-Methods Study of Translators' Fears, Worries and Preferences on Machine Translation Post-Editing'. *Revista Tradumàtica: Traducció i Tecnologies de La Informació i La Comunicació*, no. 18, 11–32.

Macklovitch, Elliott. 2006. 'TransType2: The Last Word'. In *Proceedings of the 5th Edition of the International Conference on Language Resources and Evaluation*, 6.

Mäkel, A., and J. Fulton Suri. 2001. 'Supporting Users' Creativity: Design to Induce Pleasurable Experiences'. http://www.mendeley.com/catalog/supporting-users-creativity-design-induce-pleasurable-experiences/.

Mandryk, Regan L., Kori M. Inkpen, and Thomas W. Calvert. 2006. 'Using Psychophysiological Techniques to Measure User Experience with Entertainment Technologies'. *Behaviour*

& Information Technology 25 (2): 141–58. https://doi.org/10.1080/01449290500331156.

Mao, Ji-Ye, Karel Vredenburg, Paul W. Smith, and Tom Carey. 2005. 'The State of User-Centered Design Practice'. *Communications of the ACM* 48 (3): 105–9. https://doi.org/10.1145/1047671.1047677.

Marie, Benjamin, Atsushi Fujita, and Raphael Rubino. 2021. 'Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers'. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, edited by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, 7297–7306. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.566.

Martín-Mor, Adrià. 2017. 'MTradumàtica: Statistical Machine Translation Customisation for Translators', 15.

Martín-Mor, Adrià, Pilar Sánchez-Gijón, and Ramon Piqué. 2016. *Tradumàtica: Tecnologies de La Traducció*.

Maslow, A. H. 1958. 'A Dynamic Theory of Human Motivation'. In *Understanding Human Motivation*, 26–47. Cleveland, OH, US: Howard Allen Publishers. https://doi.org/10.1037/11305-004.

Mathur, Nitika, Timothy Baldwin, and Trevor Cohn. 2020. 'Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics'. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 4984–97. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.448.

Mathur, Nitika, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. 'Results of the WMT20 Metrics Shared Task'. In *Proceedings of the Fifth Conference on Machine Translation*, edited by Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, et al., 688–725. Online: Association for Computational Linguistics. https://aclanthology.org/2020.wmt-1.77.

Matusov, Evgeny, Patrick Wilken, and Yota Georgakopoulou. 2019. 'Customizing Neural Machine Translation for Subtitling'. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, 82–93. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5209.

Melby, Alan K. 1978. 'Design and Implementation of a Computer-Assisted Translation System'. *Équivalences* 9 (2): 37–48. https://doi.org/10.3406/equiv.1978.1014.

Mellinger, Christopher, and Thomas Hanson. 2016. *Quantitative Research Methods in Translation and Interpreting Studies*. London: Routledge. https://doi.org/10.4324/9781315647845.

Mishra, Abhijit, Pushpak Bhattacharyya, and Michael Carl. 2013. 'Automatically Predicting Sentence Translation Difficulty'. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 346–51. Sofia, Bulgaria: Association for Computational Linguistics. https://www.aclweb.org/anthology/P13-2062.

Moorkens, Joss. 2012. 'Measuring Consistency in Translation Memories: A Mixed-Methods Case Study'. Doctoral, Dublin City University. https://doras.dcu.ie/17332/.

———. 2017. ‘Under Pressure: Translation in Times of Austerity’. *Perspectives* 25 (3): 464–77. https://doi.org/10.1080/0907676X.2017.1285331.
———. 2020. ‘“A Tiny Cog in a Large Machine”: Digital Taylorism in the Translation Industry’. *Translation Spaces* 9 (1): 12–34. https://doi.org/10.1075/ts.00019.moo.
———. 2022. ‘Ethics and Machine Translation’. In . https://scholar.google.es/citations?view_op=view_citation&hl=ca&user=OHVJ26MAAAAJ&sortby=pubdate&citation_for_view=OHVJ26MAAAAJ:mvPsJ3kp5DgC.
———. 2023. ‘“I Am Not a Number”: On Quantification and Algorithmic Norms in Translation’. *Perspectives*, November. https://doi.org/10.1080/0907676X.2023.2278536.
Moorkens, Joss, Sheila Castilho, Federico Gaspari, and Stephen Doherty, eds. 2018. *Translation Quality Assessment: From Principles to Practice*. Vol. 1. Machine Translation: Technologies and Applications. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7.
Moorkens, Joss, and Sharon O’Brien. 2015. ‘Post-Editing Evaluations: Trade-Offs between Novice and Professional Participants’. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, 75–81. Antalya, Turkey. https://www.aclweb.org/anthology/W15-4910.
Moorkens, Joss, and Sharon O’Brien. 2017. ‘Assessing User Interface Needs of Post-Editors of Machine Translation’. *Human Issues in Translation Technology*, 109–30.
Moorkens, Joss, and Andy Way. 2016. ‘Comparing Translator Acceptability of TM and SMT Outputs’, 11.
Mumm, Jonathan, and Bilge Mutlu. 2011. ‘Designing Motivational Agents: The Role of Praise, Social Comparison, and Embodiment in Computer Feedback’. *Computers in Human Behavior* 27 (5): 1643–50. https://doi.org/10.1016/j.chb.2011.02.002.
Muñoz Martín, Ricardo. 2012. ‘Just a Matter of Scope’. *Translation Spaces* 1 (August):169–88. https://doi.org/10.1075/ts.1.08mun.
Nielsen, Jakob. 1994. *Usability Engineering*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
———. 2010. *Usability Engineering*. Nachdr. Amsterdam: Kaufmann.
Nitzke, Jean. 2019. *Problem Solving Activities in Post-Editing and Translation from Scratch*. Zenodo. https://doi.org/10.5281/ZENODO.2546446.
Nitzke, Jean, Anke Tardel, and Silvia Hansen-Schirra. 2019. ‘Training the Modern Translator – the Acquisition of Digital Competencies through Blended Learning’. *The Interpreter and Translator Trainer* 13 (3): 292–306. https://doi.org/10.1080/1750399X.2019.1656410.
Norman, Don. 2007. *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic books. https://scholar.google.com/scholar?cluster=6178161917476102556&hl=en&oi=scholarr.
Nurminen, Mary. 2019. ‘Decision-Making, Risk, and Gist Machine Translation in the Work of Patent Professionals’. In *Proceedings of the 8th Workshop on Patent and Scientific Literature Translation*, 32–42. Dublin, Ireland: European Association for Machine Translation. https://aclanthology.org/W19-7204.
O’Brien, Sharon. 2006. ‘Pauses as Indicators of Cognitive Effort in Post-Editing Machine Translation Output’. *Across Languages and Cultures* 7 (1): 1–21. https://doi.org/10.1556/Acr.7.2006.1.1.
O’Brien, Sharon. 2011. ‘Towards Predicting Post-Editing Productivity’. *Machine Translation* 25 (3): 197–215. https://doi.org/10.1007/s10590-011-9096-7.

O'Brien, Sharon. 2012a. 'Towards a Dynamic Quality Evaluation Model for Translation'.
———. 2012b. 'Translation as Human–Computer Interaction'. *Translation Spaces* 1 (1): 101–22. https://doi.org/10.1075/ts.1.05obr.
———. 2022. 'How to Deal with Errors in Machine Translation: Post-Editing'. In *Machine Translation for Everyone*, 105–20. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.6759982.
O'Brien, Sharon. 2023. 'Human-Centered Augmented Translation: Against Antagonistic Dualisms'. *Perspectives*, August, 1–16. https://doi.org/10.1080/0907676X.2023.2247423.
O'Brien, Sharon, and Owen Conlan. 2018. 'Moving towards Personalising Translation Technology'. In *Moving Boundaries in Translation Studies*, 81–97. Routledge.
O'Brien, Sharon, Maureen Ehrensberger-Dow, Marcel Hasler, and Megan Connolly. 2017. 'Irritating CAT Tool Features That Matter to Translators'. *Hermes: Journal of Language and Communication in Business* 56 (October):145–62.
O'Brien, Sharon, and Joss Moorkens. 2014. 'Towards Intelligent Post-Editing Interfaces'. https://doras.dcu.ie/20136/1/Towards_Intelligent_PE_OBrienMoorkens.pdf.
Obrist, Marianna, Virpi Roto, and Kaisa Väänänen-Vainio-Mattila. 2009. 'User Experience Evaluation: Do You Know Which Method to Use?' In *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, 2763–66. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1520340.1520401.
O'Broin, Ultan. 2011. 'Takeaway: Language, Translation and User Experience | MultiLingual'. 2011. https://multilingual.com/articles/takeaway-language-translation-and-user-experience/.
———. 2012a. 'Is Our Industry Still Cold to User Experience? | MultiLingual'. 2012. https://multilingual.com/articles/is-our-industry-still-cold-to-user-experience/.
———. 2012b. 'Who's Afraid of User Experience Now?' *Multilingual Computing and Technology*, 2012.
Olohan, Maeve. 2011. 'Translators and Translation Technology: The Dance of Agency'. *Translation Studies* 4 (3): 342–57. https://doi.org/10.1080/14781700.2011.589656.
———. 2017. 'Knowing in Translation Practice: A Practice-Theoretical Perspective'. *Translation Spaces* 6 (1): 159–80. https://doi.org/10.1075/ts.6.1.08olo.
'OpenAI's GPT-3 Language Model: A Technical Overview'. 2020. 3 June 2020. https://lambdalabs.com/blog/demystifying-gpt-3.
Oron-Gilad, Tal, and Peter A. Hancock. 2017. 'Chapter 7 - From Ergonomics to Hedonomics: Trends in Human Factors and Technology—The Role of Hedonomics Revisited', 10.
Ortiz-Martínez, Daniel, and Francisco Casacuberta. 2014. 'The New Thot Toolkit for Fully-Automatic and Interactive Statistical Machine Translation'. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 45–48. Gothenburg, Sweden: Association for Computational Linguistics. https://doi.org/10.3115/v1/E14-2012.
Ortiz-Martínez, Daniel, Luis A. Leiva, Vicent Alabau, Ismael García-Varea, and Francisco Casacuberta. 2011. 'An Interactive Machine Translation System with Online Learning'. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, 68–73. Portland, Oregon: Association for Computational Linguistics. https://www.aclweb.org/anthology/P11-4012.
Overbeeke, Kees, Tom Djajadiningrat, Caroline Hummels, and Stephan Wensveen. 2002. 'BEAUTY IN USABILITY: FORGET ABOUT EASE OF USE!', 10.

Oviedo-Trespalacios, Oscar, Amy E. Peden, Thomas Cole-Hunter, Arianna Costantini, Milad Haghani, J. E. Rod., Sage Kelly, et al. 2023. 'The Risks of Using ChatGPT to Obtain Common Safety-Related Information and Advice'. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4346827.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. 'BLEU: A Method for Automatic Evaluation of Machine Translation'. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311. Philadelphia, Pennsylvania: Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135.

Pedhazur, Elazar J., and Liora Pedhazur Schmelkin. 2013. *Measurement, Design, and Analysis: An Integrated Approach*. psychology press. https://books.google.com/books?hl=en&lr=&id=61a2V4zv9JsC&oi=fnd&pg=PR2&dq=info:Zn0kTFbOSwIJ:scholar.google.com&ots=AVG4gG3PSN&sig=Ik1EsEo3hURNHDvQJhIUpKDHGQ8.

Pérez-Ortiz, Juan Antonio, Mikel L. Forcada, and Felipe Sánchez-Martínez. 2022. 'How Neural Machine Translation Works'. *Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence* 18:141.

Peris, Álvaro, and Francisco Casacuberta. 2019. 'Online Learning for Effort Reduction in Interactive Neural Machine Translation'. *arXiv:1802.03594 [Cs]*, April. http://arxiv.org/abs/1802.03594.

Plitt, Mirko, and François Masselot. 2010. 'A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context'. In *The Prague Bulletin of Mathematical Linguistics*. Vol. 93. https://doi.org/10.2478/v10108-010-0010-x.

Popović, Maja. 2015. 'chrF: Character n-Gram F-Score for Automatic MT Evaluation'. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392–95. Lisbon, Portugal: Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3049.

Popović, Maja, Mihael Arcan, and Arle Lommel. 2016. 'Potential and Limits of Using Post-Edits as Reference Translations for MT Evaluation'. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, 218–29. https://aclanthology.org/W16-3410.pdf.

Popović, Maja, Alberto Poncelas, Marija Brkic, and Andy Way. 2021. 'On Machine Translation of User Reviews'. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, edited by Ruslan Mitkov and Galia Angelova, 1109–18. Held Online: INCOMA Ltd. https://aclanthology.org/2021.ranlp-1.124.

Raisamo, Roope, Ismo Rakkolainen, Päivi Majaranta, Katri Salminen, Jussi Rantala, and Ahmed Farooq. 2019. 'Human Augmentation: Past, Present and Future'. *International Journal of Human-Computer Studies*, 50 years of the International Journal of Human-Computer Studies. Reflections on the past, present and future of human-centred technologies, 131 (November):131–43. https://doi.org/10.1016/j.ijhcs.2019.05.008.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. 'COMET: A Neural Framework for MT Evaluation'. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213.

Risku, Hanna. 2014. ‘Translation Process Research as Interaction Research: From Mental to Socio-Cognitive Processes’. *MonTI. Monografías de Traducción e Interpretación*, 331–53. https://doi.org/10.6035/MonTI.2014.ne1.11.

Risku, Hanna, and Regina Rogl. 2020. ‘Translation and Situated, Embodied, Distributed, Embedded and Extended Cognition’. In *The Routledge Handbook of Translation and Cognition*. Routledge.

Rossi, Caroline, and Alice Carré. 2022. ‘How to Choose a Suitable NMT Solution?: Evaluation of MT Quality’, June. https://doi.org/10.5281/ZENODO.6759978.

Rothwell, Andrew, Joss Moorkens, María Fernández-Parra, Joanna Drugan, and Frank Austermuehl. 2023. *Translation Tools and Technologies*. 1st ed. London: Routledge. https://doi.org/10.4324/9781003160793.

Roto, Virpi. 2006. ‘User Experience Building Blocks’, January.

Roto, Virpi, Marianna Obrist, and Kaisa Väänänen-vainio-mattila. 2009. ‘User Experience Evaluation Methods in Academic and Industrial Contexts’.

Roturier, Johann. 2006. ‘An Investigation into the Impact of Controlled English Rules on the Comprehensibility, Usefulness and Acceptability of Machine-Translated Technical Documentation for French and German Users’. PhD Thesis, Dublin City University. https://doras.dcu.ie/18190/.

Saldanha, Gabriela, and Sharon O’Brien. 2013. *Research Methodologies in Translation Studies*. Manchester, UK: St. Jerome Publishing.

Salvendy, Gavriel, and Waldemar Karwowski. 2021. *Handbook of Human Factors and Ergonomics*. John Wiley & Sons. https://books.google.com/books?hl=en&lr=&id=JnNEEAAAQBAJ&oi=fnd&pg=PR9&dq=info:IArPGvtilCcJ:scholar.google.com&ots=qEWzI5wnrs&sig=YluTXRg46X8VHneKPJbOHLi-uDM.

Sánchez Torrón, Marina. 2017. ‘Productivity in Post-Editing and in Neural Interactive Translation Prediction: A Study of English-to-Spanish Professional Translators’.

Sánchez-Gijón, Pilar. 2014. ‘La investigación en traducción y calidad, cosa de dos’. . . *ISSN*, 6.

Sánchez-Gijón, Pilar, Joss Moorkens, and Andy Way. 2019. ‘Post-Editing Neural Machine Translation versus Translation Memory Segments’. *Machine Translation* 33 (1–2): 31–59. https://doi.org/10.1007/s10590-019-09232-x.

Sanchis-Trilles, Germán, Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Ulrich Germann, et al. 2014. ‘Interactive Translation Prediction versus Conventional Post-Editing in Practice: A Study with the CasMaCat Workbench’. *Machine Translation* 28 (3): 217–35. https://doi.org/10.1007/s10590-014-9157-9.

Schilit, Bill, Norman Adams, and Roy Want. 1994. ‘Context-Aware Computing Applications’. In *1994 First Workshop on Mobile Computing Systems and Applications*, 85–90. IEEE. https://ieeexplore.ieee.org/abstract/document/4624429/.

Schmidtke, Dag, and Declan Groves. 2019. ‘Automatic Translation for Software with Safe Velocity’. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, 159–66. Dublin, Ireland: European Association for Machine Translation. https://aclanthology.org/W19-6729.

Schrepp, Martin, Andreas Hinderks, and Jörg Thomaschewski. 2014. *Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios*. https://doi.org/10.1007/978-3-319-07668-3_37.

Schrepp, Martin, Jörg Thomaschewski, and Andreas Hinderks. 2017. 'Construction of a Benchmark for the User Experience Questionnaire (UEQ)', June. https://doi.org/10.9781/ijimai.2017.445.

Sebastian, Glorin. 2023. 'Do ChatGPT and Other AI Chatbots Pose a Cybersecurity Risk? - An Exploratory Study'. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4363843.

Secară, Alina. 2005. 'Translation Evaluation: A State of the Art Survey'. In *Proceedings of the eCoLoRe/MeLLANGE Workshop*, 39–44. Citeseer. https://scholar.google.com/scholar?cluster=8610952488864606122&hl=en&oi=scholarr.

Shackel, Brian. 2009. 'Human–Computer Interaction – Whence and Whither?' *Interacting with Computers* 21 (5–6): 353–66. https://doi.org/10.1016/j.intcom.2009.04.004.

Shneiderman, Ben. 2020a. 'Design Lessons from AI's Two Grand Goals: Human Emulation and Useful Applications'. *IEEE Transactions on Technology and Society* 1 (2): 73–82.

———. 2020b. 'Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy'. *International Journal of Human–Computer Interaction* 36 (6): 495–504. https://doi.org/10.1080/10447318.2020.1741118.

———. 2020c. 'Human-Centered Artificial Intelligence: Three Fresh Ideas'. *AIS Transactions on Human-Computer Interaction* 12 (3): 109–24. https://doi.org/10.17705/1thci.00131.

———. 2022a. *Human-Centered AI*. Oxford University Press. https://books.google.com/books?hl=en&lr=&id=mSRXEAAAQBAJ&oi=fnd&pg=PP1&dq=info:R2ABLndGsMAJ:scholar.google.com&ots=n1ci4eiM1b&sig=lqnJuid27YA20CFSU8h7Ek5hA7c.

———. 2022b. 'Human-Centered AI: Ensuring Human Control While Increasing Automation'. In *Proceedings of the 5th Workshop on Human Factors in Hypertext*, 1–2. Barcelona Spain: ACM. https://doi.org/10.1145/3538882.3542790.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. 'A Study of Translation Edit Rate with Targeted Human Annotation'. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 223–31.

Specia, Lucia, Dhwaj Raj, and Marco Turchi. 2010. 'Machine Translation Evaluation versus Quality Estimation'. *Machine Translation* 24 (1): 39–50. https://doi.org/10.1007/s10590-010-9077-2.

Sperber, Dan, and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. Harvard University Press.

Stanney, Kay, Brent Winslow, Kelly Hale, and Dylan Schmorrow. 2015. 'Augmented Cognition'. In *APA Handbook of Human Systems Integration*, 329–43. APA Handbooks in Psychology®. Washington, DC, US: American Psychological Association. https://doi.org/10.1037/14528-021.

Steigerwald, Emma, Valeria Ramírez-Castañeda, Débora Y C Brandt, András Báldi, Julie Teresa Shapiro, Lynne Bowker, and Rebecca D Tarvin. 2022. 'Overcoming Language Barriers in Academia: Machine Translation Tools and a Vision for a Multilingual Future'. *BioScience* 72 (10): 988–98. https://doi.org/10.1093/biosci/biac062.

Stymne, Sara, Henrik Danielsson, Sofia Bremin, Hongzhan Hu, Johanna Karlsson, Anna Prytz Lillkull, and Martin Wester. 2012. 'Eye Tracking as a Tool for Machine Translation Error

Analysis.' In *LREC*, 1121–26. http://lrec.elra.info/proceedings/lrec2012/pdf/192_Paper.pdf.

Suchman, Lucille Alice. 1987. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge university press. https://books.google.com/books?hl=en&lr=&id=AJ_eBJtHxmsC&oi=fnd&pg=PR7&dq=info:ycy-9HwdLAcJ:scholar.google.com&ots=KtGrjKGJNP&sig=OeFoRGGjo1aT1luTr3e-vTGxmVU.

———. 2007. *Human-Machine Reconfigurations: Plans and Situated Actions*. Cambridge university press. https://books.google.com/books?hl=en&lr=&id=VwKMDV-Gv1MC&oi=fnd&pg=PR7&dq=info:6NuYaNo3jGsJ:scholar.google.com&ots=BQzByKZWiF&sig=Pn7E6-mJcjANh8XUKMqEIYSCJ3U.

Sumita, Eiichiro, and Yutaka Tsutsumi. 1988. 'A Translation Aid System Using Flexible Text Retrieval Based on Syntax-Matching', 14.

Suojanen, Tytti, Kaisa Koskinen, and Tiina Tuominen. 2014. *User-Centered Translation*. Routledge.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. 'Sequence to Sequence Learning with Neural Networks'. *Advances in Neural Information Processing Systems* 27. https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.

Taherdoost, Hamed. 2016. 'Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research'. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.3205040.

Teixeira, Carlos SC, Joss Moorkens, Daniel Turner, Joris Vreeke, and Andy Way. 2019. 'Creating a Multimodal Translation Tool and Testing Machine Translation Integration Using Touch and Voice'. In *Informatics*, 6:13. MDPI. https://www.mdpi.com/2227-9709/6/1/13.

Terribile, Silvia. 2023. 'Is Post-Editing Really Faster than Human Translation?' *Translation Spaces*, December. https://doi.org/10.1075/ts.22044.ter.

Tezcan, Arda, Veronique Hoste, and Lieve Macken. 2017. 'Scate - Taxonomy and Corpus of Machine Translation Errors', 29.

Tiedemann, Jorg. 2009. 'News from OPUS — A Collection of Multilingual Parallel Corpora with Tools and Interfaces'. In *Cilt.309.19tie*. John Benjamins Publishing Company. https://benjamins.com/catalog/cilt.309.19tie.

Toral, Antonio. 2019. 'Post-Editese: An Exacerbated Translationese'. arXiv. https://doi.org/10.48550/arXiv.1907.00900.

———. 2020. 'Reassessing Claims of Human Parity and Super-Human Performance in Machine Translation at WMT 2019'. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 185–94. Lisboa, Portugal: European Association for Machine Translation. https://www.aclweb.org/anthology/2020.eamt-1.20.

Toral, Antonio, and Víctor M. Sánchez-Cartagena. 2017. 'A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions'. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1063–73. Valencia, Spain: Association for Computational Linguistics. https://www.aclweb.org/anthology/E17-1100.

Torregrosa Rivero, Daniel. 2018. 'Black-box interactive translation prediction'. Http://purl.org/dc/dcmitype/Text, Universitat d'Alacant / Universidad de Alicante. https://dialnet.unirioja.es/servlet/tesis?codigo=149939.

Torregrosa-Rivero, Daniel, Mikel L. Forcada, and Juan Antonio Pérez-Ortiz. 2014. 'An Open-Source Web-Based Tool for Resource-Agnostic Interactive Translation Prediction', October. https://doi.org/10.2478/pralin-2014-0015.

Torres-Hostench, Olga, Joss Moorkens, Sharon O'Brien, and Joris Vreeke. 2017. 'Testing Interaction with a Mobile MT Post-Editing App'. *Translation & Interpreting* 9 (2): 138–50. https://doi.org/10.12807/t&i.v9i2.645.

Torres-Hostench, Olga, Ramon Piqué Huerta, Pilar Sánchez-Gijón, Anna Aguilar-Amat, Adrià Martín Mor, Celia Rico Pérez, Amparo Alcina Caudet, and Miguel Ángel Candel Mora. 2016. 'L'ús de Traducció Automàtica i Postedició a Les Empreses de Serveis Lingüístics de l'Estat Espanyol'. https://ddd.uab.cat/record/166753.

Ugas, Mohamed, Meredith Giuliani, and Janet Papadakos. 2024. 'When Is Good, Good Enough? On Considerations of Machine Translation in Patient Education'. *Journal of Cancer Education*, January. https://doi.org/10.1007/s13187-024-02401-4.

Vallor, Shannon. 2024. 'Defining Human-Centered AI'. In , 13–20. https://doi.org/10.1201/9781003320791-3.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. 'Attention Is All You Need'. arXiv. https://doi.org/10.48550/arXiv.1706.03762.

Vermeeren, Arnold P. O. S., Effie Lai-Chong Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila. 2010. 'User Experience Evaluation Methods: Current State and Development Needs'. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, 521–30. NordiCHI '10. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1868914.1868973.

Vieira, Lucas Nunes, Minako O'Hagan, and Carol O'Sullivan. 2021. 'Understanding the Societal Impacts of Machine Translation: A Critical Review of the Literature on Medical and Legal Use Cases'. *Information, Communication & Society* 24 (11): 1515–32. https://doi.org/10.1080/1369118X.2020.1776370.

Vienne, Jean. 1994. 'Toward a Pedagogy of "Translation in Situation"'. *Perspectives* 2 (1): 51–59. https://doi.org/10.1080/0907676X.1994.9961222.

Way, Andy. 2013. 'Emerging Use-Cases for Machine Translation'. In *Proceedings of Translating and the Computer 35*. https://aclanthology.org/2013.tc-1.12.pdf.

———. 2018. 'Quality Expectations of Machine Translation'. In *Translation Quality Assessment*, edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, 1:159–78. Machine Translation: Technologies and Applications. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7_8.

Weinberg, Bruce A. 2004. 'Experience and Technology Adoption'. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.522302.

White, Jules, Sam Hays, Quchen Fu, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. 'ChatGPT Prompt Patterns for Improving Code Quality, Refactoring, Requirements Elicitation, and Software Design'. arXiv. https://doi.org/10.48550/arXiv.2303.07839.

Wilson, Deirdre, and Dan Sperber. 1993. 'Linguistic Form and Relevance'. *Lingua* 90 (1): 1–25. https://doi.org/10.1016/0024-3841(93)90058-5.

Winner, Langdon. 2007. 'Do Artifacts Have Politics?' In *Computer Ethics*. Routledge.

Wright, Peter, John McCarthy, and Lisa Meekison. 2004. 'Making Sense of Experience'. In *Funology: From Usability to Enjoyment*, edited by Mark A. Blythe, Kees Overbeeke, Andrew F. Monk, and Peter C. Wright, 43–53. Human-Computer Interaction Series. Dordrecht: Springer Netherlands. https://doi.org/10.1007/1-4020-2967-5_5.

Wu, Yunhan, Daniel Rough, Anna Bleakley, Justin Edwards, Orla Cooney, Philip R. Doyle, Leigh Clark, and Benjamin R. Cowan. 2020. 'See What I'm Saying? Comparing Intelligent Personal Assistant Use for Native and Non-Native Language Speakers'. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, 1–9. Oldenburg Germany: ACM. https://doi.org/10.1145/3379503.3403563.

Yue, Thomas, David Au, Chi Chung Au, and Kwan Yuen Iu. 2023. 'Democratizing Financial Knowledge with ChatGPT by OpenAI: Unleashing the Power of Technology'. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4346152.

Zens, Richard, Franz Josef Och, and Hermann Ney. 2002. 'Phrase-Based Statistical Machine Translation'. In *KI 2002: Advances in Artificial Intelligence*, edited by Matthias Jarke, Gerhard Lakemeyer, and Jana Koehler, 18–32. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-45751-8_2.

Zhang, Hong, and Olga Torres-Hostench. 2022. 'Training in Machine Translation Post-Editing for Foreign Language Students'. https://scholarspace.manoa.hawaii.edu/handle/10125/73466.

Zhang, Mike, and Antonio Toral. 2019. 'The Effect of Translationese in Machine Translation Test Sets'. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, edited by Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, et al., 73–81. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5208.

Zhong, Junhao, Yilin Zhong, Minghui Han, Tianjian Yang, and Qinghua Zhang. 2023. 'The Impact of AI on Carbon Emissions: Evidence from 66 Countries'. *Applied Economics* 0 (0): 1–15. https://doi.org/10.1080/00036846.2023.2203461.

Zhuo, Terry Yue, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. 'Red Teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity'. arXiv. https://doi.org/10.48550/arXiv.2301.12867.

APPENDIX A. DCU ETHICS APPROVAL

Faculty of Humanities & Social Sciences
DUBLIN CITY UNIVERSITY
03 May 2023
CONFIRMATION OF RESEARCH ETHICS APPROVAL FOR A PROJECT
Application Reference:
Project Title:
Project contact(s):
DCU-FHSS-2022-020
Translator User Experience: The Forgotten Element in Machine Translation
vicent.brivaiglesias2@mail.dcu.ie

This project was originally approved on 10 March 2022, and was re-approved on 03 May 2023, on notification of amendments to the proposed research.
Let this letter certify that the proposed project identified above has been reviewed by the *Humanities & Social Sciences Faculty Research Ethics Committee* (F-REC) and has been approved as a low-risk project. The application was found to that comply with university requirements and best practices for research ethics, and with GDPR guidelines and requirements where personal data is processed in the project.

A copy of the application, including appended documents related to participant consent, is archived under the reference above. Queries about this project's approval may be directed to the F-REC Chair.

Sincerely,
Dr Dónal Mulligan
donal.mulligan@dcu.ie
Chair, Faculty Research Ethics Committee
Faculty of Humanities & Social Sciences
Dublin City University

APPENDIX B. RECRUITMENT JOB AD FOR THE MAIN STUDY – TRANSLATORS AND REVIEWERS

For translators

To Whom It May Concern,

We are looking for 11 junior English-Spanish translators specialized in the legal domain to participate in a paid research study called "Machine Translation User Experience: The Forgotten Element in Machine Translation". This study compares the user experiences of translators interacting with two different Machine Translation Post-Editing (MTPE) modalities, namely traditional post-editing and interactive post-editing. The texts to be translated are from the legal domain in the English-Spanish language combination, and translators will be paid at an hourly rate of €20. Your entire participation will take no longer than 11.5 hours of your time.

We are interested in the experience of the translator when using different MTPE modalities. This study is part of a research project funded by the SFI CRT of Digitally-Enhanced Reality (d-real). It is being carried out by Vicent Briva-Iglesias under the supervision of Dr Sharon O'Brien and Dr Benjamin R. Cowan.

If you accept to participate, you will be asked to complete a pre-task and a post-task questionnaire, and to post-edit different English legal texts using two different MTPE modalities. We will make sure to keep the collected data anonymous and confidential at all times. These are the requirements for being eligible to participate:

- You are a junior translator, that is, you have less than 5 years' full-time experience in the language services industry.
- You are a native Spanish speaker, and your main language combination is English-Spanish.
- You have translation experience in the legal domain.

ONLY IF YOU MEET THE REQUIREMENTS and are interested in participating, please send us an email at the address below expressing your interest and attaching your CV with the following SUBJECT: "[MTUX Study Translator]". Please only apply to participate if you are willing to

participate in the entire study. Please also note that only those who complete the entire set of tasks will be paid. Non-completion will result in no payment.

Thank you in advance for your cooperation,

Vicent Briva-Iglesias (vicent.brivaiglesias2@mail.dcu.ie)

For reviewers

To Whom It May Concern,

We are looking for three senior English-Spanish reviewers specialized in the legal domain to participate in a paid research study called "Machine Translation User Experience: The Forgotten Element in Machine Translation". This study compares the user experiences of translators interacting with two different Machine Translation Post-Editing (MTPE) modalities, namely traditional post-editing and interactive post-editing. Three reviewers will assess the translation quality of texts translated by professional translators over some iteration rounds. They will be paid at an hourly rate of €20. Two reviewers are expected to work around 3 hours. The third reviewer is expected to review a higher volume of work, and their entire participation will take around 55 hours of their time.

We are interested in the experience of the translator when using different MTPE modalities. This study is part of a research project funded by the SFI CRT of Digitally-Enhanced Reality (d-real). It is being carried out by Vicent Briva-Iglesias under the supervision of Dr Sharon O'Brien and Dr Benjamin R. Cowan.

If you accept to participate, you will be asked to assess and score different translations made by professional translators by following strict Translation Quality Assessment (TQA) guidelines. We will make sure to keep the collected data anonymous and confidential at all times. These are the requirements for being eligible to participate:

- You are a senior reviewer, that is, you have more than 5 years' full-time experience in the language services industry.
- You are a native Spanish speaker and your main language combination is English-Spanish.

- You have translation experience in the legal domain.

ONLY IF YOU MEET THE REQUIREMENTS and are interested in participating, please send us an email at the address below expressing your interest and attaching your CV with the following SUBJECT: “[MTUX Study Reviewer]”. Please only apply to participate if you are willing to participate in the entire study. Please also note that only those who complete the entire set of tasks will be paid. Non-completion will result in no payment.

Thank you in advance for your cooperation,

Vicent Briva-Iglesias (vicent.brivaiglesias2@mail.dcu.ie)

APPENDIX C. PRE-TASK QUESTIONNAIRE

Questions regarding machine translation post-editing (MTPE)

We understand MTPE as the process where a translator corrects raw machine translated output according to specific guidelines and quality criteria.

Q1. Do you have experience with MTPE tasks?

Yes / No

Q1.1. How long have you engaged with MTPE tasks? Give an approximate time of use with months or years and months (e.g., 1 year and 6 months).

Q2. On a scale of 1-7, where 1 is “Strongly Dislike” and 7 is “Strongly Like”, please rate your perception of doing MTPE tasks in professional translation projects.

Q3. On a scale of 1-7, where 1 is “Not trustworthy at all” and 7 is “Very trustworthy”, please rate if you can trust MTPE to help you successfully delivery a professional translation project.

Q4. Please rate how much you agree or disagree with this statement: “Machine Translation is a threat to the sustainability of the translation profession. [1 is “Completely disagree” and 7 is “Completely agree”]

Q5. Please rate the following statement: “When I am doing MTPE tasks, I find them [SCORE]”.

[1 is “Boring” and 7 is “Engaging”]

APPENDIX D. MTUX QUESTIONNAIRE

This research aims to study Machine Translation User Experience (MTUX), a new concept in Translation Studies, which we consider to be the user experience of translators interacting with machine translation (MT).

The initial hypothesis of this work is that a positive MTUX may influence the translation process, resulting in increased productivity in the long term and a more enjoyable task for MT users in the short term, as well as in a final product (translation) of better quality.

You will now complete one MTUX questionnaire (User Experience Questionnare), so we can analyse your UX when interacting with MT. In the questionnaire, you will find word pairs that are intended to aid you in assessing the product that you have just become acquainted with. The word pairs represent extreme opposites, with seven graduations possible between them. An example:

Please mark the box to acknowledge that you have read the instructions of the questionnaire and that you commit to complete it completely.

Adjective pairs:

Annoying-Enjoyable

Not understandable-Understandable

Creative-Dull

Easy to learn-Difficult to learn

Valuable-Inferior

Boring-Exciting

Not interesting-Interesting

Unpredictable-Predictable

Fast-Slow

Inventive-Conventional

Obstructive-Supportive

Good-Bad

Complicated-Easy

Unlikable-Pleasing

Usual-Leading edge

Unpleasant-Pleasant

Secure-Not secure

Motivating-Demotivating

Meets expectations-Doest not meet expectations

Inefficient-Efficient

Clear-Confusing

Impractical-Practical

Organized-Cluttered

Attractive-Unattractive

Friendly-Unfriendly

Conservative-Innovative

APPENDIX E. FLUENCY DIFFERENCE BETWEEN TPE AND IPE

This appendix contains a step-by-step explanation of how the TPE and IPE modalities impact translation fluency. The IPE process has been reproduced in a table after thorough examination of the screen recordings of the main longitudinal study. The fragment selected is extracted from the segment 11 of Text 6. The TPE was produced by T0010, while the IPE was produced by T0008.

Source	WHEREAS, it is reasonable, prudent and necessary for the Company contractually to obligate itself to indemnify, and to advance expenses on behalf of, such persons to the fullest extent permitted by applicable law so that they will serve or continue to serve the Company free from undue concern that they will not be so indemnified;
Raw MT	CONSIDERANDO que es razonable, prudente y necesario que la Compañía se obligue contractualmente a indemnizar y anticipar los gastos en nombre de dichas personas en la medida máxima permitida por la ley aplicable para que sirvan o continúen sirviendo a la Compañía sin preocupación indebida de que no serán indemnizados;
Target TPE (T0010)	CONSIDERANDO QUE, es razonable, prudente y necesario que la Compañía esté obligada contractualmente a indemnizar y adelantar los gastos en nombre de dichas personas en la medida máxima permitida por la ley aplicable para que sirvan o continúen sirviendo a la Compañía sin preocupación de no ser indemnizados;
Target IPE (T0008)	Es razonable, prudente y necesario que la Empresa se obligue por contrato a indemnizar y a adelantar los costes en nombre de estas personas dentro de los límites legales para que trabajen o sigan trabajando para la Empresa con la garantía de que serán indemnizadas;

Table E.1.

Table E.1 contains the source language text, the unedited MT system proposal, and the final translation proposals for TPE (T0010) and IPE (T0008). The colour coding of Table E.1 is as follows:

- Fragments highlighted in yellow are preferential changes that have been made to the source text. That is, changes that are neither correct nor incorrect, but which the translator has deemed necessary.
- Fragments highlighted in red are errors. In this case, they only appear in the segment post-edited by TPE and are errors in the MT that the translator did not change.
- Fragments highlighted in green are appropriate and correct translation solutions. These fragments can be caused either by a change of the translator or by an update of the interactive system's MT proposal.

Thus, we can see that, in the TPE example, T0010 makes a number of preferential changes and leaves much of the TPE proposal identical to the TA proposal. The final evaluation of this translation is Adequacy 3 and Fluency 3, as the fragments highlighted in red detract from the quality of the translation.

In contrast, the IPE proposal has a rating of 4 both in Adequacy and Fluency. The different iterations carried out after reviewing the recordings are detailed and explained below. The colour coding of the tables below is:

- In green, you can see the words that the translator has validated in the IPE modality.
- In red, you can observe the deletions that the translator has made in a segment.
- In square brackets, the additions that the translator has made to a segment can be seen.
- Next to the bracketed word, and underlined, the word selected in purple by Lilt is observed (in other words, the next word that could be automatically added with the hotkey) (see Figure 6.2).
- In bold, you can observe the words that are updated in the MT proposal when the translator makes a change.

Iteration 0	MT	CONSIDERANDO que es razonable, prudente y necesario que la Compañía se obligue contractualmente a indemnizar y anticipar los

		gastos en nombre de dichas personas en la medida máxima permitida por la ley aplicable para que sirvan o continúen sirviendo a la Compañía sin preocupación indebida de que no serán indemnizados;
Iteration 1	User	~~CONSIDERANDO que~~ Es razonable, prudente y necesario que la **[Empresa]**Compañía se obligue contractualmente a indemnizar y anticipar los gastos en nombre de dichas personas en la medida máxima permitida por la ley aplicable para que sirvan o continúen sirviendo a la Compañía sin preocupación indebida de que no serán indemnizados
	MT	Es razonable, prudente y necesario que la **Empresa** se obligue contractualmente a indemnizar y anticipar los gastos en nombre de dichas personas en la medida máxima permitida por la ley aplicable para que sirvan o continúen sirviendo a la **Empresa** sin preocupación indebida de que no serán indemnizados

Table E.2. IPE iteration 1

In Iteration 1, T0008 deletes “CONSIDERANDO que”, validates a series of words, and makes a preferential change of “Empresa” instead of “Compañía”. When this happens, the IPE system re-runs itself and reproduces the change of “Empresa” later in the MT proposal. Should this change have happened in the TPE workflow, T0008 would have needed to change “Compañía” twice.

Iteration 2	User	Es razonable, prudente y necesario que la Empresa se obligue **[por]**contractualmente a indemnizar y anticipar los gastos en nombre de dichas personas en la medida máxima permitida por la ley aplicable para que sirvan o continúen sirviendo a la Empresa sin preocupación indebida de que no serán indemnizados
	MT	Es razonable, prudente y necesario que la Empresa se obligue **por contrato** a indemnizar **y a adelantar** los **costes** en nombre de **estas** personas en la medida máxima permitida por la ley aplicable para que sirvan o continúen sirviendo a la Empresa sin preocupación indebida de que no serán indemnizados

Table E.3. IPE iteration 2

In Iteration 2, T0008 introduces “por” and then the MT proposal makes a series of changes highlighted in bold. In this case, the changes done by the MT system are only preferential and do not add anything positive or negative to the MT proposal.

Iteration 3	User	Es razonable, prudente y necesario que la Empresa se obligue por contrato a indemnizar y a adelantar los costes en nombre de estas personas **[dentro de]**en la medida máxima permitida por la ley aplicable para que sirvan o continúen sirviendo a la Empresa sin preocupación indebida de que no serán indemnizados
	MT	Es razonable, prudente y necesario que la Empresa se obligue por contrato a indemnizar y a adelantar los costes en nombre de estas personas **dentro de los límites legales** para que sirvan o continúen sirviendo a la Empresa sin preocupación indebida de que no serán indemnizados

Table E.4. IPE iteration 3

In Iteration 3, T0008 introduces “dentro de” and guides the IPE system to change the main structure of the MT output, which replaces “en la medida máxima permitida por la ley applicable” (a very literal translation proposal of the source text in English) for “dentro de los límites legales” (a more fluent and natural collocation in Spanish).

Iteration 4	User	Es razonable, prudente y necesario que la Empresa se obligue por contrato a indemnizar y a adelantar los costes en nombre de estas personas dentro de los límites legales para que **[trabajen]**sirvan o continúen sirviendo a la Empresa sin preocupación indebida de que no serán indemnizados
	MT	Es razonable, prudente y necesario que la Empresa se obligue por contrato a indemnizar y a adelantar los costes en nombre de estas personas dentro de los límites legales para que **trabajen o continúen trabajando para** la Empresa **con la garantía** de que serán **indemnizadas**

Table E.5. IPE iteration 4

Finally, in Iteration 4, T0008 changes “sirvan” (a non-adequate verb in Spanish, which is a literal translation from the English source text) into “trabajen”, and the IPE system re-runs

itself once again to implement different changes in the MT proposal. This time, the system learns from all the previous validated words and changes “con la garantía” (a more appropriate translation solution than previously suggested) and “indemnizadas” (correcting a mistake, as the previous MT proposal was in masculine, while it should be in feminine).

As a consequence, we can see that in the IPE workflow, the system adapts to the proposals of the translator. This does not only allow for better translation solutions (and higher adequacy) in this case, but also to a higher fluency because the final translation is less similar in terms of syntactic structure than the raw MT output edited in the TPE proposal, and reads more natural in Spanish.