UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Doctoral Thesis

Towards Augmented Translation: Design and Application of a Benchmarking Framework for Post-Editing Environments

Marie Escribe

Supervisor: Prof. Miguel Ángel Candel Mora

Department of Applied Linguistics

Industrial thesis in collaboration with

LanguageWire

Company supervisor: Daphné André

Valencia, July 2025

« Ces transitions d'un idiome à l'autre, du passé à l'avenir,
de la décrépitude à l'aurore, de la mort à la vie, sont laborieuses.
Les traductions y aident, les apprêtent de longue main,
les adoucissent, les facilitent. »

Victor Hugo
Proses Philosophiques – Les Traducteurs (1860-1865)

«Esas transiciones de un idioma al otro, del pasado al porvenir,
de la decrepitud a la aurora, de la muerte a la vida, son laboriosas;
las traducciones las favorecen, las preparan durante un largo proceso,
atenúan sus rigores, las facilitan.»
Traducción de John Jairo Gómez Montoya

"Shifts from one language to another, from the past to the future,
from dusk to dawn, from death to life, are laborious.
Translations help there, they extend a helping hand,
reducing their toughness, making them easy."
Translated by Wikisource Translators

ACKNOWLEDGEMENTS

I am immensely grateful to all the people who have helped and accompanied me throughout the process of writing this thesis.

En primer lugar, quisiera expresar mi sincero agradecimiento a mi supervisor, el profesor Miguel Ángel Candel Mora, por su valiosa orientación a lo largo del proceso de redacción de esta tesis. Durante los tres años que he pasado bajo su supervisión, he tenido el inmenso privilegio de aprender junto a un investigador brillante y apasionado, y de evolucionar a través de fascinantes conversaciones sobre tecnologías de la traducción. Es, sin duda, una de las personas más inspiradoras que he tenido la suerte de conocer, y le estoy infinitamente agradecida por su apoyo y dedicación.

Je suis également particulièrement reconnaissante du précieux soutien que m'a apporté Daphné André tout au long de ces trois années en me supervisant à LanguageWire. Je la remercie d'avoir accepté de se lancer dans cette aventure mêlant recherche académique et application concrète, et de m'avoir aidée à faire le lien entre la théorie et la pratique. Ce fût un grand privilège de pouvoir compter sur l'accompagnement d'une professionnelle aussi passionnée que Daphné.

I would also like to thank my colleagues at LanguageWire for their invaluable support. A special mention goes to the linguists in Daphné André's and Raquel Burgos Pastor's teams for agreeing to take part in the various experiments that were essential to this work. I must also say a million thanks to all the language professionals I had the pleasure of interacting with and those who took the time to fill in the exploratory survey and share their feedback. Thank you for making this research possible. A note of gratitude is also due to my closest colleagues, in particular José Ángel Alegre, Jorge Alcaide Gadea, Mario Puglisi and Agostina Feltrinelli, for their enthusiasm for my research and their encouragement.

My academic journey began before the thesis, and I must also thank the people who inspired me to start a PhD. Among them, I must mention Prof. Ruslan Mitkov and the academic team I had the chance to meet during my time at the University of Wolverhampton, thanks to whom I was able to acquire knowledge that has served as a pillar in my research. Throughout this PhD, I have also taken part in various conferences where I have been able to see these colleagues again, meet new ones, and continue to exchange ideas with them on fascinating topics. Here, a special mention is due to the RANLP, HiT-IT and NeTTT teams.

To my dear friends, who have always shown me their support despite the distance and the time apart: thank you for everything, merci infiniment, muchísimas gracias.

Enfin, je tiens à remercier tout particulièrement mes parents pour leur soutien constant, pour m'avoir toujours encouragée, pour m'avoir réconfortée dans les moments de doute, pour n'avoir jamais cessé de croire en moi. Merci pour tout.

千里之行，始于足下

"A journey of a thousand miles starts beneath one's feet." These words of 老子 (Laozi) are often cited to emphasise the importance of taking the *first step* to begin a long or challenging endeavour. I hope that this thesis is the *first step* towards an equally wonderful adventure; and above all, I sincerely wish that all those who have accompanied me so far will continue to be by my side in the future.

ABSTRACT

The rise of artificial intelligence and the concept of augmenting human capabilities are currently transforming a wide range of fields. The language industry is no exception. In today's globalised world, translation plays a central role in enabling multilingual communication, and technology has become key in meeting the demands for large translation volumes. Recently, the advent of generative artificial intelligence and large language models has marked a major leap forward, and machine translation models have also begun to build on these advances.

However, while machine translation has progressed drastically, computer-assisted translation tools, which are used on a daily basis by many language professionals, have not evolved at the same pace. These tools were indeed conceptualised over 40 years ago and originally revolved around components like translation memories and term bases – which are still relevant but have been supplemented with the integration of machine translation. Today, these tools constitute the primary environment in which machine translation output is post-edited, yet only recent attempts have been made to better adapt them to the post-editing process, which presents significant differences compared to translation from scratch. These recent efforts include the integration of artificial intelligence assistants into translation tools. Such developments represent considerable progress, but they also highlight the continuous reshaping of the interaction between language professionals and translation tools. In this context, further research is essential to understand the real needs of linguists and envision tools that bridge the gap between technological advancements and human needs.

The present work seeks to address this gap by adopting a human-centred approach to establish a framework for evaluating translation technology in the context of post-editing and to formulate recommendations for improvements based on the needs and experiences of post-editors. To that end, a benchmarking framework was designed to assess the post-editing environment based on three core pillars: productivity, user experience and quality. For each pillar, a meticulous assessment model was created based on the results from a series of surveys, experiments and case studies revolving around machine translation quality, post-editing effort and user experience.

The resulting framework was then applied to two real-world scenarios: 1) the assessment of Smart Editor, the proprietary computer-assisted translation tool of LanguageWire (the industrial partner in this research), and 2) the evaluation of prompt-based alternative translations in the Trados Studio translation environment. Beyond testing the framework, these applications provided concrete insights into user needs, which in turn resulted in the formulation of recommendations for developers of translation tools. Such a benchmarking process should be a continuous approach, and the proposed framework should thus serve as a starting point to continue evaluating the tools used by language professionals as they evolve in the future.

RESUMEN

El auge de la inteligencia artificial y la idea de aumentar las capacidades humanas están transformando actualmente una amplia variedad de campos. El sector lingüístico no es una excepción. En el mundo globalizado de hoy en día, la traducción desempeña un papel fundamental a la hora de permitir la comunicación multilingüe, y la tecnología se ha convertido en un elemento clave para satisfacer la demanda de grandes volúmenes de traducción. Recientemente, la llegada de la inteligencia artificial generativa y los grandes modelos de lenguaje han supuesto un gran salto adelante, y los modelos de traducción automática también han empezado a basarse en dichos avances.

Sin embargo, aunque la traducción automática ha experimentado un avance espectacular, las herramientas de traducción asistida por ordenador, utilizadas a diario por muchos profesionales lingüísticos, no han evolucionado al mismo ritmo. De hecho, estas herramientas se concibieron hace más de 40 años y, en un principio, giraban en torno a componentes como las memorias de traducción y las bases de datos terminológicas. Estas siguen siendo relevantes, pero se han complementado con la integración de la traducción automática. Hoy en día, dichas herramientas constituyen el entorno principal en el que se posedita el resultado de la traducción automática, pero solo recientemente se ha intentado adaptarlas más al proceso de posedición, que presenta diferencias significativas con respecto a la traducción desde cero. Entre estos esfuerzos recientes destaca la integración de asistentes de inteligencia artificial en las herramientas de traducción. Dichos desarrollos representan un progreso considerable, pero también ponen de relieve la continua transformación de la interacción entre los profesionales lingüísticos y las herramientas de traducción. En este contexto, es esencial seguir investigando para comprender las necesidades reales de los lingüistas y concebir herramientas que cubran la brecha entre los avances tecnológicos y las necesidades humanas.

El presente trabajo pretende abordar esta brecha adoptando un enfoque centrado en las personas para establecer un marco de evaluación de la tecnología de traducción en el contexto de la posedición y formular recomendaciones de mejora basadas en las necesidades y experiencias de los poseditores. Con este fin, se diseñó un marco de referencia para evaluar el entorno de posedición a partir de tres pilares principales: la productividad, la experiencia del usuario y la calidad. Para cada pilar, se creó un meticuloso modelo de evaluación basado en los resultados de una serie de encuestas, experimentos y casos prácticos que giraban en torno a la calidad de la traducción automática, el esfuerzo de posedición y la experiencia del usuario.

A continuación, el marco resultante se aplicó a dos escenarios reales: 1) la evaluación de Smart Editor, la herramienta de traducción asistida por ordenador propiedad de LanguageWire (socio industrial de esta investigación), y 2) la evaluación de traducciones alternativas basadas en *prompts* en el entorno de traducción de Trados Studio. Además de probar el marco, estas aplicaciones proporcionaron

información concreta sobre las necesidades de los usuarios, lo que a su vez dio lugar a la formulación de recomendaciones para los desarrolladores de herramientas de traducción. Dicho proceso de evaluación comparativa debe ser un enfoque continuo y, por lo tanto, el marco propuesto debe servir como punto de partida para seguir evaluando las herramientas utilizadas por los profesionales lingüísticos a medida que estas evolucionen en el futuro.

RESUM

L'auge de la intel·ligència artificial i la idea d'augmentar les capacitats humanes estan transformant actualment una àmplia varietat de camps. El sector lingüístic no és una excepció. En el món globalitzat de hui dia, la traducció exercix un paper fonamental a l'hora de permetre la comunicació multilingüe, i la tecnologia s'ha convertit en un element clau per a satisfer la demanda de grans volums de traducció. Recentment, l'arribada de la intel·ligència artificial generativa i els grans models de llenguatge han suposat un gran salt avant, i els models de traducció automàtica també han començat a basar-se en estos avanços.

No obstant això, encara que la traducció automàtica ha experimentat un avanç espectacular, les ferramentes de traducció assistida per ordinador, utilitzades diàriament per molts professionals lingüístics, no han evolucionat al mateix ritme. De fet, estes ferramentes es van concebre fa més de 40 anys i, en un principi, giraven entorn de components com les memòries de traducció i les bases de dades terminològiques. Estes continuen sent rellevants, però s'han complementat amb la integració de la traducció automàtica. Hui dia, estes ferramentes constituïxen l'entorn principal en el qual es postedita el resultat de la traducció automàtica, però només recentment s'ha intentat adaptar-les més al procés de postedició, que presenta diferències significatives respecte a la traducció des de zero. Entre estos esforços recents destaca la integració d'assistents d'intel·ligència artificial en les ferramentes de traducció. Estos desenrotllaments representen un progrés considerable, però també posen en relleu la contínua transformació de la interacció entre els professionals lingüístics i les ferramentes de traducció. En este context, és essencial continuar investigant per a comprendre les necessitats reals dels lingüistes i concebre ferramentes que cobrisquen la bretxa entre els avanços tecnològics i les necessitats humanes. El present treball pretén abordar esta bretxa adoptant un enfocament centrat en les persones per a establir un marc d'avaluació de la tecnologia de traducció en el context de la postedició i formular recomanacions de millora basades en les necessitats i experiències dels posteditors. A este efecte, es va dissenyar un marc de referència per a avaluar l'entorn de postedició a partir de tres pilars principals: la productivitat, l'experiència de l'usari i la qualitat. Per a cada pilar, es va crear un meticulós model d'avaluació basat en els resultats d'una sèrie d'enquestes, experiments i casos pràctics que giraven entorn de la qualitat de la traducció automàtica, l'esforç de postedició i l'experiència de l'usuari.

A continuació, el marc resultant es va aplicar a dos escenaris reals: 1) l'avaluació de Smart Editor, la ferramenta de traducció assistida per ordinador propietat de LanguageWire (soci industrial d'esta investigació), i 2) l'avaluació de traduccions alternatives basades en *prompts* a l'entorn de traducció de Trados Studio. A més de provar el marc, estes aplicacions van proporcionar informació concreta sobre les necessitats dels usuaris, la qual cosa va donar lloc a la formulació de recomanacions per als desenrotlladors de ferramentes de traducció. Este procés d'avaluació comparativa ha de ser un

enfocament continu i, per tant, el marc proposat ha de servir com a punt de partida per a continuar avaluant les ferramentes utilitzades pels professionals lingüístics a mesura que estes evolucionen en el futur.

TABLE OF CONTENTS

List of Abbreviations

AI: Artificial Intelligence

ALPAC: Automatic Language Processing Advisory Committee

ALPS: Automated Language Processing Systems

APE: Automatic Post-Editing

ASQ: After-Scenario Questionnaire

ASR: Automatic Speech Recognition

ASTM: American Society for Testing and Materials

BF: Benchmarking Framework

BLEU: BiLingual Evaluation Understudy

CAT: Computer-Assisted Translation

CEN: Centre Européen de Normalisation

CM: Context Match

CSA: Common Sense Advisory

CSAT: Customer SATisfaction

CSUQ: Computer System Usability Questionnaire

DNT: Do Not Translate

DQF: Dynamic Quality Framework

DTS: Descriptive Translation Studies

EAGLES: Expert Advisory Group on Language Engineering Standards

EM: Exact Match

EMT: European Master's in Translation

EUATC: European Union of Associations of Translation Companies

FAHQT: Fully Automatic High-Quality Translation

FEMTI: Framework for the Evaluation of Machine Translation in ISLE

GenAI: Generative Artificial Intelligence

GIS: Geographical Information Systems

GPT: Generative Pre-trained Transformer

HAMT: Human-Aided Machine Translation

HaTS: Happiness Tracking Survey

HCI: Human-Computer Interaction

HTER: Human-targeted Translation Edit Rate

HTML: Hyper Text Markup Language

IA: Intelligence Augmentation

IATE: Interactive Terminology for Europe

ISLE: International Standards for Language Engineering

ISO: International Organisation for Standardisation

KAT: Knowledge-Assisted Translation

LISA: Localisation Industry Standards Association

LQA: Linguistic Quality Assessment

LRE: Linguistic Research and Engineering

LSP: Language Service Provider

MAHT: Machine-Aided Human Translation

MATEO: MAchine Translation Evaluation Online

METEOR: Metric for Evaluation of Translation with Explicit ORdering

MIND: Management of Information through Natural Discourse

MMPE: Multi-Modal Post-Editing

MOOC: Massive Open Online Course

MQM: Multi-dimensional Quality Metrics

MT: Machine Translation

MTLD: Measure of Textual Lexical Diversity

MTPE: Machine Translation Post-Editing

MTUX: Machine Translation User Experience

mWER: multi-reference Word Error Rate

NLP: Natural Language Processing

NMT: Neural Machine Translation

NPS: Net Promoter Score

OCR: Optical Character Recognition

PACTE: Process of Acquisition of Translation Competence and Evaluation

PE: Post-Editing

PEEIs: Post-Editing Effort Estimation Indicators

PET: Post-Editing Tool

POS: Part Of Speech

PSSUQ: Post-Study System Usability Questionnaire

QA: Quality Assurance

QE: Quality Estimation

QUIS: Questionnaire for User Interaction Satisfaction

RBMT: Rule-Based Machine Translation

RQ: Research Question

RUSE: Regressor Using Sentence Embeddings

SAE: Society of Automotive Engineers

SEECAT: Speech & Eye-Tracking Enabled Computer Assisted Translation

SEQ: Single Ease Question

SL: Source Language

ST: Source Text

SMT: Statistical Machine Translation

SUMI: Software Usability Measurement Inventory

SUS: System Usability Scale

TA: Translator's Amanuensis

TAM: Technology Acceptance Model

TAP: Think-Aloud Protocol

TAUS: Translation Automation User Society

TB: Term Base

TEMAA: Testbed Study of Evaluation Methodologies: Authoring Aids

TEnTs: Translation Environment Tools

TEP: Translation, Editing, Proofreading

TER: Translation Edit Rate

TL: Target Language

TM: Translation Memory

TP: Translation Process

TPR: Translation Process Research

TQA: Translation Quality Assessment

TQE: Translation Quality Evaluation

TSS: Translation Support System

TT: Target Text

TTR: Type-Token Ratio

TTS: Text To Speech

TS: Translation Studies

UEQ: User Experience Questionnaire

UGC: User-Generated Content

UI: User Interface

UMUX: Usability Metric for User Experience

UX: User eXperience

WER: Word Error Rate

WPH: Words Per Hour

WYSIWYG: What You See Is What You Get

XLIFF: XML Localisation Interchange File Format

XML: Extensible Markup Language

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

Today, technology and Artificial Intelligence (AI) are becoming integral parts of many humans' lives, thus transforming how individuals interact with tools and perform daily tasks. From voice assistants to text generation, technology continues to evolve, sometimes to the extent of blurring the lines between the real and the virtual, as in the field of extended reality technologies with the shift from *virtual* reality to *augmented* reality, where instead of creating fully simulated digital environments, real environments become supplemented by overlaying digital information onto them.

The idea of *augmenting* human capabilities is also affecting the field of translation, both from the perspective of research and that of industry applications. In the language industry, technology has also become increasingly ubiquitous. Machine Translation (MT), chatbots, and other AI tools are now commonly used and even embedded in various websites and social media platforms. The advent of Generative AI (GenAI), notably with the release of ChatGPT in November 2022, has accelerated this phenomenon and led to widespread research into Large Language Models (LLMs) as well as the integration of AI assistants into diverse tools and workflows. The personal assistants providing instant translations, which were once typical of science fiction, are becoming a reality.

With the advances in content production, the demand for multilingual communication continues to grow, and MT thus plays a crucial role in meeting this need. Industry reports highlight the exponential growth of MT, driven by the increasing volume of content requiring translation. The language industry is now moving towards a Language Operations (LangOps[1]) model, which takes a holistic approach to managing and optimising language services. This model goes beyond traditional localisation methods, emphasising collaboration with AI to streamline workflows and enhance efficiency. Although the technological landscape evolves rapidly, which sometimes challenges traditional processes, the projected growth of the industry is often predicted to remain high. Nimdzi estimated that the language services industry reached USD 67.9 billion in 2023 and predicted a growth of up to USD 95.3 billion by 2028 (Nimdzi, 2024). According to the Slator Language Industry Market Report (Slator, 2024), the MT market alone increased by 31% in 2023 to reach USD 1.55 billion.

However, while technology itself – especially MT – has advanced rapidly, the tools that language professionals are using on a daily basis have not adapted at the same pace. The rapid evolution of translation technology has fundamentally reshaped how translators and post-editors interact with tools. Translation has become a Human-Computer Interaction (HCI) process (O'Brien, 2012), where technology is deeply embedded in every stage of the workflow. Yet, despite these advancements, many

[1] https://langops.org/ (consulted: 30/04/2025)

of the tools currently used by linguists, notably Computer-Assisted Translation (CAT) tools, were conceptualised over 40 years ago, and have not evolved significantly since the integration of MT. As a result, linguists post-edit MT outputs in environments designed for working primarily with Translation Memories (TMs) and Term Bases (TBs), which are less suited to the realities of today's MT systems. This disconnect between the evolution of MT and the tools in which Post-Editing (PE) is performed represents a critical gap in the industry. The MT systems of 2025, powered by deep learning models and LLMs, are indeed vastly different from those of even a decade ago, especially when it comes to output quality. Despite this, the terminology commonly used in the industry – "CAT tools," "MT," and "PE" – has not evolved to reflect these changes. Similarly, translation quality evaluation methods and traditional workflows, such as the Translation-Editing-Proofreading (TEP) process, remain mostly unchanged, though AI integrations are starting to transform the typical translation workflow. This situation suggests a need for novel approaches and updated terminology that better align with the reality of modern translation technology. Recent attempts to integrate AI assistants into translation tools represent a crucial step forward to adapt CAT tools for PE. Nevertheless, further research is needed to understand the real needs of linguists working with today's MT systems and to design tools that bridge the gap between technological advancements and the requirements of post-editors.

Both academic research and translation companies strive to envision innovative solutions to enhance translation processes. However, as technology evolves, linguists must adapt to new toolsets and evolving workflows continuously. In such a dynamic environment, effective communication between key stakeholders is paramount to ensure that technological advances align with the needs of end users. In this case, major contributors are language professionals but also computational linguists, user researchers, product managers and other essential roles in software development (Lagoudaki, 2006; Moorkens, 2020). Therefore, such a collaboration is often challenging, as it stands at the intersection of diverse knowledge fields, with each domain adopting a specific viewpoint. Nevertheless, enhancing usability and ensuring linguistic quality appear as crucial factors to consider when envisioning future technologies.

The present work brings together academia and industry via a collaboration with LanguageWire, a technology-driven Language Service Provider (LSP), which provides multilingual communication services across the globe, and owns a proprietary suite of language products, including Smart Editor, the company's CAT tool. Therefore, by drawing on both academic literature and industry sources – including reports, webpages, and practical insights – this study ensures relevance and currency in a rapidly evolving field. In addition to enriching the theoretical foundation, the use of diverse industry sources also grounds the research in real-world applications, thus providing a more holistic perspective. The primary motivation behind this work stems from the situation of language professionals, who face a constant technological transformation that redefines their work and requires continuous adaptation

(Briva-Iglesias and O'Brien, 2022). Despite these changes, the variety of methods employed to assess translation environments (e.g. Rico, 2001; Zaretskaya, 2016; Krüger, 2016) suggests that no clear consensus has been reached regarding the adoption of a framework to evaluate translation tools from the perspective of their users. This situation greatly complicates any assessment of technology with the goal of augmenting human capabilities, particularly in the context of PE. This research therefore aims at addressing this gap by adopting a human-centred approach. The primary objective is thus to establish a framework for the evaluation of translation technology and formulate recommendations for improvements based on the needs and experiences of post-editors.

To achieve this, this study adopts a benchmarking methodology, chosen for its ability to holistically evaluate complex, interdependent factors in real-world scenarios. A thorough literature review is conducted with a view to identifying the pillars of the framework. It spans across four areas: the historical evolution of translator workstations from early tools (Kay, 1980; Melby, 1982), including the theoretical underpinnings around the interaction between translators and machines (Bar-Hillel, 1960; Licklider, 1960; Hutchins, 1998) to contemporary AI-augmented systems (Alonso and Nunes Vieira, 2017; Herbig et al., 2019; Lee et al., 2021); user experience principles, both in the field of software development (Nielsen, 1994; Sommerville, 2011; Sauro and Lewis, 2016) and in translation technology (Ehrensberger-Dow and O'Brien, 2015; Moorkens and O'Brien, 2017), the intricacies of both the translation process and the PE process (Holmes, 1988; Krings, 2001; do Carmo, 2017; Mossop, 2020) and established methods for assessing translation quality (Görög, 2014; Lommel, Uszkoreit and Burchardt, 2014). Building on this foundation, three main pillars are defined as part of the benchmarking framework: productivity, user experience and translation quality. While the productivity dimension is quantified via post-edited words per hour, the other two pillars required further investigation. To that end, quality was examined through the effort associated with diverse types of MT errors, and a survey was conducted to find out which aspects of the PE experience were the most important to post-editors, which allowed for the creation of a custom questionnaire. The framework was then applied in two empirical studies: a year-long evaluation of Smart Editor's performance for PE, and a focused assessment of how AI-generated alternative translations influence the PE process.

By uniting these objectives and methods, the study advances understanding of PE needs while offering a concrete, industry-relevant framework for evaluation. The dual focus on technological capabilities and human factors responds to calls in Translation Studies for more user-centred approaches to tool assessment, while its application to both established and emerging technologies ensures relevance in a rapidly evolving field. While it is acknowledged that this type of research can quickly become obsolete due to the rapid pace of technological advancement, the development of an evaluation framework remains essential. Such a framework provides a structured approach to assessing current tools. It can also be adapted to take account of future developments, making it useful also in the long term. By focusing on the interplay between usability, linguistic quality, and productivity, this research aims to

contribute to the assessment and design of tools that better support linguists in their work, with a view to enhancing both their efficiency and satisfaction.

This work provides a comprehensive exploration of post-editing environments, beginning with a theoretical foundation and culminating in the design and application of a benchmarking framework. First, a theoretical background is established in chapter II. It explores four key areas that are essential to understanding the PE environment. It begins with an examination of the translator's workstation, tracing the origins and evolution of translation technology from its early developments to the latest advances. This is followed by a discussion of the user experience of translation tools, introducing usability concepts and examining their application to translation technology, especially when PE is involved. The theoretical framework then delves into the PE process, focusing on the shift from traditional translation to PE and studying the resulting requirements in terms of skillsets and guidelines. Finally, it addresses quality evaluation, defining what quality means in the context of translation and reviewing existing methods for assessing translation quality, particularly in relation to MT and post-edited output. Then, the theory is put to practice through a methodology chapter (III) and an analysis and discussion chapter (IV), that form the core of this research. The methodology chapter starts with a detailed exploration of the rationale, research questions, and objectives that situate the work, followed by an introduction to the research context, which presents the industrial collaboration with LanguageWire as a unique opportunity for a practical application. Following this, the methodology is presented as grounded in the principles of benchmarking and tailored to the unique demands of PE environments. This methodology is then executed via a series of experiments. More specifically, the design of the benchmarking framework is presented next, establishing the core pillars of productivity, user experience and quality, and defining how they should be assessed. Finally, the framework is applied to two real-world scenarios: the evaluation of Smart Editor and the assessment of prompt-based alternative translations using AI assistance.

II. THEORETICAL BACKGROUND

This chapter offers an in-depth review of the theoretical foundations underpinning the present work, focusing on four core areas that are critical to understanding the PE environment and its assessment. These areas collectively form the backbone of this research, bridging historical developments, user experience principles, process-oriented insights, and quality considerations. It begins with an exploration of the translator's workstation, tracing the historical evolution of translation technology from its early developments to the state-of-the-art systems in use today. By examining the integration of MT into translation tools as well as the development of other advanced features, this section highlights how technological advancements have transformed the translator's workflow. Next, the focus shifts to the PE user experience, emphasising the critical role of usability in the design and evaluation of translation tools. Drawing on established usability and user experience models, this section identifies the key dimensions that influence post-editors' efficiency, satisfaction, and overall experience. It underscores the value of aligning technological innovations with the needs and expectations of linguists. The third area delves into the transition from translation to PE, exploring the distinctions in processes between translation from scratch and PE. This section examines the implications for post-editor competence, addressing both theoretical and practical challenges inherent in PE. It also considers the evolving role of the post-editor in an increasingly automated landscape, where technological assistance and human expertise are coexisting. Finally, translation quality is presented as a multifaceted concept. The section on quality explores definitions of translation quality, examines how quality is assessed at various stages of the translation process and compares various quality assessment methods. It also considers the impact of MT quality on PE effort and outcomes. By synthesising insights from these four areas, this chapter establishes a robust theoretical foundation for the development of a benchmarking framework centred on productivity, user experience and quality. Beyond providing a contextualisation of the challenges and opportunities in PE environments, it also provides the necessary groundwork for the practical applications explored in subsequent chapters.

1. The Translator's Workstation

"Translation without tools simply does not exist" (Cronin, 2003, p. 24). While this may at first appear as a strong statement, after contemplating the translation practice, it seems to hold true – at least since writing systems have been used in translation. The tools used by translators have evolved significantly over the years, going from mere writing instruments to computers, and finally specific translation tools. Due to major technological advances, translation tools are now revolutionising the translators' workspace. In 2023, the Nimdzi Language Technology Atlas report covered no less than 920 tools,

including Translation Management Systems, Machine Translation, and Quality Management, among other categories (Nimdzi, 2023). This section provides a historical overview of the evolution of tools for translators, from the origins of translation technology to state-of-the-art approaches. In addition to providing practical descriptions, the theoretical implications of the different proposals are also discussed.

1.1. History of Translation Technology

The incorporation of technology in the translation process can be traced back to the early Machine Translation (MT) systems. The evolution of translation technology is one full of twists and turns and has been marked by optimistic phases typically followed by periods of discouragement. The primary objective has regularly oscillated between fully automatic translation and the development of tools to assist translators in their work.

MT is concerned with translating texts automatically, without human intervention. The origins of this field date to the 1940s, thus making this domain one of the oldest applications of computer science (Specia and Wilks, 2016). Although his works were not widely distributed at the time, one of the earliest descriptions reminiscent of MT can be attributed to Petr Petrovich Troyanskii (Hutchins and Lovtskii, 2000). In 1933, this Russian inventor conceptualised a mechanical translation device, in which the central part was a belt allowing for sequentially printing target words together with grammatical information on a tape (ibid.). In 1947, Warren Weaver, a researcher at the Rockefeller Foundation, proposed using computers to perform automatic translation based on information theory and successes in code breaking during the Second World War (Hutchins, 2005). A few years later, in 1954, a public demonstration of the Georgetown-IBM[2] system took place in New York (Hutchins, 2005). This event was described as the "first public demonstration of MT" and although it was a small-scale experiment, it gave a strong impetus to the field. Hutchins (2014) indeed referred to the period between 1954 and 1966 as being the "decade of optimism". Given the historical context, the Soviet Union became particularly interested in automatic translation and embarked upon similar research initiatives, which were summarised in a book by Dmitri Yur'evich Panov in 1955 (Hutchins and Lovtskii, 2000). Five years later, Yehoshua Bar-Hillel (according to Hutchins [2005], the first person to hold an MT researcher position) expressed his concerns as to the reasonableness of fully automatic translation systems, and instead, called for a "machine-post-editor partnership" (Bar-Hillel, 1960, p.97). The same year, Joseph Carl Robnett Licklider (a computer scientist known to be an Internet pioneer), made similar recommendations on a more general level, and advocated for a "symbiotic partnership" between humans and machines (Licklider, 1960, p.4).

[2] IBM: International Business Machines Corporation (a multinational computer and information technology company)

Although early attempts in MT appeared to foreshadow promising research avenues, a significant turning point was reached in 1966, when the Automatic Language Processing Advisory Committee (ALPAC) published a report named "Language and Machines: Computers in Translation and Linguistics" (ALPAC, 1966). In this document, the ALPAC concluded that research was too underwhelming to continue investigating MT systems, as there was, at this time, "no immediate or predictable prospect of useful machine translation" (ibid., p.32). Consequently, instead of focusing on developing fully automatic translation engines, the report encouraged the development of tools to assist the human translation process. Although this change implied a substantial reduction in the financial support MT research had received in the beginning, the report brought a paradigm shift in translation technology research by drawing attention to the development of tools designed to facilitate the work of professional translators. This recommendation was made on the basis of two examples of organisations making use of such technology. The first one was the case of the Federal Armed Forces Translation Agency, which used a system allowing translators to send a query containing the words they would like to search in a glossary. The system performed a morphological reduction on the search words before sending them to the glossary to retrieve corresponding entries. The computer then returned the list of words together with their target equivalents. The use of this database (called a "text-related glossary") proved to bring significant productivity gains, as the translator was provided directly with relevant glossary entries, and thus did not spend additional time searching in conventional glossaries. An experiment conducted at the Federal Armed Forces Translation Agency indeed demonstrated that translators working with traditional glossaries spent more time (on average 66% more) than translators working with text-related glossaries and that such glossaries contributed to reducing translation errors.

The second use case occurred at the Terminological Bureau of the European Coal and Steel Community and stemmed from similar observations regarding the time that translators typically spend elucidating terminological questions. This institution therefore implemented an "automatic dictionary lookup with context included" (ibid., p.27). Similar to the text-related glossary, the lookup functionality received source words as input. The output, however, did not consist of a list of terms; instead, the inputted terms in their context were provided (i.e. occurrences of the search word in previous sentences). Each query was appended to the database, thus making this system increasingly useful each time a new item was added. While this functionality was designed primarily for terminology reasons, the concept of providing search results in context is reminiscent of concordance search, and the accumulation of queries in a database for future consultation is a similar concept to that of Translation Memories (further discussed in 1.2.1 on CAT tools).

Therefore, a clear dichotomy can be observed here. On the one hand, research has initially focused on providing fully automatic translations. On the other hand, real-world applications that were used and recommended by translators were different, with a focus on helping professionals in the rendition of their translations by providing relevant terminological resources. The initial proposal according to which

machines should translate everything (otherwise they would not be useful) was referred to as the "all or nothing syndrome" (Melby, 1982). Translators themselves were in favour of this shift from automatic translation to computer-aided translation. As Hutchins (1998) pointed out, at this time, MT outputs were of inferior quality and produced major syntactic and lexical errors, which entailed heavy editing. The conditions for correcting such texts were tedious, as translators added their corrections by hand on a printout before handing over the annotated versions to a typist. This was a time-consuming and cumbersome process, as human experts needed to adapt their ways of working around the output provided by machines. In contrast, terminological aids, such as the two use cases described above, were well perceived by translators, as they did not disrupt their habitual working process but instead provided valuable support to it. In addition to creating a more pleasant experience, such terminological tools were much needed at the time due to the challenges faced by large organisations to maintain up-to-date terminological resources (ibid.).

These tools, referred to as translators' workstations or workbenches, gained importance after the ALPAC report. Unlike MT systems, which constitute a form of Human-Aided Machine Translation (HAMT), such workstations fall under the category of Machine-Aided Human Translation (MAHT) (Hutchins, 1998). Beyond terminological databases, the translator's workstation gathers in one single place a wide range of functionalities. Hutchins (1998, p.289) indeed provided the following definition:

> *"these workstations make available to the translator at one terminal (whether individual computer or as part of a company network) a range of integrated facilities: multilingual word processing, OCR [Optical Character Recognition] scanning, electronic transmission and receipt of documents, spelling and grammar checkers (and perhaps style checkers or drafting aids), publication software, terminology management, text concordancing software, access to local or remote termbanks (or other resources), TM [Translation Memory] (for access to individual or corporate translations), and access to automatic translation software."*

At first, commercial tools only provided basic features, but over the years, these became more numerous and supported various types of assistance. Although the first commercial workstations supporting a full range of features emerged in the 1990s, they (or their components) were conceptualised in the 1970s and 1980s by different experts in mostly independent proposals (Hutchins, 1998). In 1971, Friedrich Krollmann and Erhard Lippmann presented similar ideas, notably regarding matching entries from a terminological database. Krollmann (1971) described an extension of glossaries into a larger database, which he called a "linguistic data bank". In his proposal, this data bank would centralise several dictionaries and other resources (organised in subbanks) and would thus be useful not only for translators but also for other language specialists (terminologists, lexicographers, language teachers and

documentalists). One of the subbanks would be a "superdictionary" to which users could send their queries, and they would receive results printed in the desired order (e.g. source or alphabetical). Lipmann (1971) presented a proposal for computer-aided translation in which the system would make a certain range of facilities available to translators. He envisioned professionals seated at their terminal who would perform several actions from the same system, including notably editing texts, searching for words in a dictionary, updating files, and accessing remote terminological databases. He described several text processing actions (e.g. search and replace functions) that have nowadays become common in most text processing software. Hutchins (1998) indeed pointed out that these features were still underdeveloped at that time. Apart from having to type certain commands to move the cursor, it was also impossible to consult a dictionary on the same screen in which the text was displayed. Lipmann (1971, p.10) described his suggestion as a way to "enhance and accelerate human translation […] by employing man-machine interaction techniques" and considered this system as an "extension of the capabilities of the user".

A more elaborate proposal on the reuse of translations stored in an archive was presented in the late 1970s. Peter Arthern indeed suggested a "translation by text-retrieval" method (named "TERRIER") at the first Translating and the Computer conference held in 1978 (Hutchins, 1998). Arthern had observed the repetitive nature of certain texts and concluded that the work of translators was not always optimal, as they frequently had to retranslate passages that had already been entirely or partially translated (Arthern, 1978). Depending on the match with previously translated units, the work of translators could sometimes consist in the modification or insertion of names, or the translation of additional information. His proposal is reminiscent of previous suggestions regarding the translator's workstation, in the sense that he also foresaw a centralised system from which translators could perform various actions, such as translating, editing and revising text, and sending queries to a terminological database. His suggestion to store translations in a memory for future reuse was meant to be an integral component of this system. This description remains accurate for TMs as they are implemented today. In addition, Arthern also predicted the integration of MT into the same type of systems, to provide translation suggestions for fragments which were not found in the memory.

During the same decade, Martin Kay had worked on the design of the Management of Information through Natural Discourse (MIND) system (Bisbey and Kay, 1972), a system relying on the translator's input to resolve ambiguities throughout the translation process. In 1980, Kay described what he called the Translator's Amanuensis, a system with an interface displaying the text in two separate windows (the upper part being the source and the one at the bottom being the target), together with various facilities designed to help the translator produce the target text (Kay, 1980). These functionalities included notably dictionary lookup, suggestions retrieved from past translations and a method for marking tentative translations with a view to replacing them at a later stage in the process. In his proposal, the last component would work with brackets that translators could use to denote provisional translations,

which could then be updated at once automatically. Such a suggestion is reminiscent of common functionalities in modern tools (i.e. search and replace, saving as draft and auto-propagation). He also imagined a morphologically intelligent processing technique, which would allow for placing the correct word form when replacing bracketed words (which is similar to morphology-aware term recognition, further detailed in 1.2.1). Beyond the technical applications of this proposal, the Translator's Amanuensis had a strong conceptual reach. Kay strongly advocated for placing the human translator at the centre by developing "cooperative man-machine systems" which would be used under the supervision of the translator, as they would "[be] there to help increase his productivity and not to supplant him" (ibid., p.18). He envisioned a future in which machines would be entrusted with more and more tasks:

> *"I want to advocate a view of the problem in which machines are gradually, almost imperceptibly, allowed to take over certain functions in the overall translation process. First they will take over functions not essentially related to translation. Then, little by little, they will approach translation itself. The keynote will be modesty. At each stage, we will do only what we know we can do reliably. Little steps for little feet!"* (ibid., p.11).

The circulation of such ideas constituted what Hutchins (1998, p.295) considered to be "one of the most decisive moments in the development of the future translator's workstation". At this time, Kay's publication had not been widely disseminated, and another researcher, Alan Melby, presented independent – although highly similar – ideas (Hutchins, 1998). According to do Carmo, (2017), the term "Computer-Assisted Translation" (CAT) was used for the first time in one of Melby's publications. Melby (1982) suggested a workstation with three different levels of assistance between which the translators could switch as they wished. The rationale behind this proposal was that it was not relevant to have all sentences treated equally by the system, as not every sentence would require the same degree of support. The first level supported primarily terminological assistance by providing translators with access to terminology databanks (including a shared term bank) as well as a database of translated texts for research purposes. While the first level did not necessarily require the original text to come in a machine-readable format, this was a prerequisite for the second level. Level two included functionalities based on source text processing. In this level, a "suggestion box" would automatically scan the source text and provide terminology suggestions for each source word. Suggestions could easily be inserted by translators in the desired location with a few keystrokes. Melby also described a "morphological routine" to account for inflected forms based on contextual information found in the source and target sentences (which is similar to morphology-aware term recognition). In addition to these features, level three integrated a "full-blown MT system", which would provide suggestions that translators could easily accept, edit or ignore. Melby suggested an MT system which would come with a "self-evaluation

metric": the system would thus attribute a "quality rating" to each translated sentence (the scale suggested was A-B-C-D, with A being the equivalent of human quality, and D meaning that the output was severely insufficient). The translator would be able to filter the MT suggestions based on this rating, thereby displaying outputs above a certain threshold. This quality rating is reminiscent of Quality Estimation (further detailed in 1.3.2 on Quality Estimation). Melby's proposal highly resembles that of Kay, in that both advocated for empowering translators by letting them be in control of the system they used. However, although the theoretical foundation of their proposals was similar, they suggested different ways of putting it into practice. While in the vision of Kay, machines would be entrusted with gradually complex tasks, Melby imagined distinct levels of assistance integrated into the same system (Hutchins, 1998). In that sense, the approach suggested by Melby is even more human-centric, as the translator would always have the option to select the level of assistance required at any stage of the process. Several years after Melby's proposal, Heyn (1996) also described two methodologies to add MT in the translator's workstation: batch and interactive integration. In the batch scenario, all segments for which no match could be retrieved were passed to an MT engine, whereas in the interactive scenario, translators could send a query to the MT system on an on-demand basis. While the second option was more challenging in terms of implementation, it was more aligned with Melby's vision of letting the translator be in control of the system.

In 1988, Brian Harris provided a more detailed description of the database containing source and translated texts, which he called a "bi-text". He described a workstation in which all translations were stored on the fly in a database, aligned with their source counterparts (Harris, 1987). A "search engine" could browse this bi-text to retrieve occurrences of a specific word in already translated units and display this word in the context of previous translations. More results would be found as the database was fed with translated units, and this bi-text would therefore provide a "memory-perfect exploitation of the translator's own previous experience" (ibid., p.9). Harris also specified that finding occurrences of a non-conventional unit might be difficult: in this case, partially matching units would be provided instead of exact matches. This bi-text database, allowing for retrieval of previous translations with different degrees of similarity with the query, is a correct description of TMs and their fuzzy matches used in today's tools (further detailed in 1.2.1).

Although early proposals of the translator's workstation can be traced back to the 1970s, the hardware available at that time was limited. The first commercial computers indeed appeared in the mid-1980s, and the first implementations of such workstations became available in the 1990s (Bowker and Fisher, 2010), mostly as a response to the increasing demand of corporations and institutions (Garcia, 2014).

According to Hutchins (1998), all components of the translator's workstation (i.e. multilingual word processing, document storage, terminology lookup, access to glossaries and previously translated

segments and MT) had been identified by the end of the 1980s. However, such a system was in practice considered "by translators as something for the future" (ibid., p.301).

The first commercial application of the translator's workstation is known to be the Translation Support System (TSS) launched by Automated Language Processing Systems (ALPS), which was demoed at the 1986 Translating and the Computer conference (Language Monthly, 1986). At this time, it was estimated that this system would "triple the productivity of the individual translator" (ibid., p.10).

The ALPS consisted of three main components. The first one was AutoTerm, a functionality which provided a glossary of terms after scanning the source text. While this is reminiscent of the text-related glossary of the Federal Armed Forces Translation Agency, the resulting terms were not provided in a separate list but displayed in a "reference window" located at the bottom of the screen (Hutchins, 1998). The second component was called "repstraction" (a combination of "repetitions extraction") and was the first implementation of TMs. This functionality relied on a repetitions file, which contained previous translations. The system could then match these translations against the text to generate suggestions. The third component was the TransActive facility, an MT system which would provide suggestions. These could be accepted, rejected or disambiguated interactively to resolve lexical or syntactic ambiguities – which is reminiscent of the MIND system.

Similar commercial tools emerged in the early 1990s. In particular, Hutchins (1998) mentioned four workstations: the TranslationManager/2 (from IBM Corporation), the Transit system (from STAR AG[3]), the Eurolang Optimiser (coming from the GETA[4] and Eurotra MT projects) and the well-known Translator's Workbench (from Trados). To that list, Moorkens (2012) added XL8 and Déjà Vu.

These tools continued to evolve, including more and more advanced features and supporting more file formats. As a matter of fact, Trados launched MultiTerm in 1990, and the Translator's Workbench II (which included a TM with an automatic propagation feature) two years later (Moorkens, 2012). In 1994, a Windows-compatible version of the Workbench was launched, and in 1999 the TagEditor component was added to support the translation of tagged formats, such as Hyper Text Markup Language (HTML; ibid.). In 1993, Déjà Vu stood out as the first tool compatible with Microsoft. In 1996, another version of Déjà Vu with a proprietary interface was launched, which constituted a significant difference compared with other systems working as add-ons (based on macros) in word processing software. This entailed a radically different experience for translators, who, depending on the system being used, would see either the formatted text in a "What You See Is What You Get" (WYSIWYG) format inside a word processing tool, or unformatted text within a specific workstation interface (Moorkens, 2012). This dichotomy persisted, as ten years later, Lagoudaki (2006) provided a classification of CAT tools based on this difference and argued that the choice of one type of tool over the other was a matter of personal

[3] STAR: Software Translation Artwork Recording

[4] GETA: Groupe d'Étude pour la Traduction Automatique (Study Group for Machine Translation)

preference. She identified Wordfast, MultiTrans, Logoport, Metatexis, Trados Studio and Fusion as tools operating inside Microsoft Word, and DéjàVu, Heartsome, memoQ, STAR Transit, SDLX and Across as tools providing their own translation environments (ibid.). The second type of tools gained popularity over the years. In 2013, Trados launched a version which did not support Trados Workbench, which had two editing environments: TagEditor and Microsoft Word (do Carmo, 2017). Trados Workbench was the first tool to integrate a TM and alignment support (via the TAlign component), thus allowing translators to create their own TMs from existing translations. Trados is also known to be the first company to use the term "Translation Memory" (Hutchins, 1998). Following the predictions of Kay and Melby, MT was added to these systems, notably thanks to collaborations with third-party providers, like Systran in the case of Transit (ibid.).

While MT research was considerably slowed down after 1966 to the benefit of translator's workstations, this field of study gradually gained importance over the years. In 1976, a decade after the ALPAC report, the MÉTÉO sublanguage system developed at the University of Montréal provided encouraging results, and the research community regained interest in MT (Specia and Wilks, 2016). The field of MT underwent a series of paradigm shifts over the years. The first engines were Rule-Based (RBMT), thus relying on a set of linguistic rules to transfer a representation of the source text into a target language (ibid.). Improvements in computing capacity in the 1990s allowed for exploiting large corpora, which led to Statistical MT (SMT) systems. These were designed to treat translation as a cryptography problem, generating the target text via a decoding process based on conditional probabilities observed in the data (ibid.). In 1997, several researchers (Ñeco and Forcada, 1997; Castaño and Casacuberta, 1997) formulated an application of neural networks for MT. This theory was inspired by the activation process of biological neurones. Neural MT (NMT) engines transform sentences into vectorial representations, which are decoded into a target language based on probability distributions of vocabulary items derived from the activations encountered while traversing the network (Forcada, 2017). Although NMT could not be implemented at the time due to insufficient technological resources, this technology is commonly integrated into translation workflows today, thus making Post-Editing (PE, i.e. correcting MT outputs) the norm in the translation industry.

The fact that MT research has grown in the last decades is largely motivated by the need to provide automatic translations. The ALPAC report made it clear that there was "no emergency in the field of translation" (ALPAC, 1966, p.16), since the human workforce could bear translation requests. MT was therefore considered as an experiment – no more than a tool which deserved to be studied "in the name of science". Today, such a perspective is obsolete. With the Internet, globalisation, and language generation technology, the content which is published daily and requires to be translated has reached record amounts, thus creating a pressing need for automatic translations. MT has also become essential

in the work of translators, allowing them to process more text and boost their productivity, which is crucial to manage high volumes and meet tight deadlines.

1.2. The Current Landscape

Today, translation technology is an integral part of the translation workflow. This section describes the current technological landscape by presenting translation tools and their components, and discusses the integration of MT.

1.2.1. CAT Tools

Computer-Assisted Translation – sometimes also called Computer-Aided Translation – (CAT) tools support the translation and revision of content from one language into another. Bowker and Fisher (2010, p.60) defined CAT as "the use of computer software to assist a human translator in the translation process". CAT environments therefore provide assistance throughout the translation process, but translators are responsible for the production of the final text. Several terms have been used in academia and the industry to refer to CAT tools, including notably translator workstations or workbenches, translation support tools, TM tools and translation environment tools (TEnTs; Bowker and Fisher, 2010; Garcia, 2014). "TM tools" in particular, is a widely used name, albeit not entirely accurate. Although the TM is the backbone of CAT tools, calling these TM tools can lead into thinking that only the TM component is being mentioned. Further confusion can occur when TM is interpreted as "Translation Management".

Some CAT tools are integrated as a macro in word processing software (e.g. Microsoft Word), whereas others are standalone applications with their own text processing environment. The former has the advantage of letting users work in a type of software they are likely to be familiar with. According to Garcia (2014), since 2005, this type of CAT tool is not widely commercialised or used anymore (apart from certain exceptions, such as Wordfast Classic). For this reason, the remainder of the chapter focuses on the second type of tool: standalone software. Initially, these tools were desktop-based applications. Since 2006, online platforms have emerged and the industry currently witnesses a strong tendency to adopt online tools, which can be explained mostly by the control and uniformity this modality provides (Garcia, 2014). Working online also offers more collaboration possibilities and allows for instant reuse of confirmed segments by other translators (ibid.). Some tools, like Trados Studio, now even support hybrid modalities, whereby translators can easily shift between the desktop application and the online version of the same tool.

Today, Trados Studio and memoQ are among the most popular tools. While it would be difficult to establish an exhaustive inventory of CAT tools available in the market, it is worth mentioning Wordfast,

Across, Déjà Vu, Star Transit, XTM, Phrase, Wordbee and Smartling. Certain CAT tools are free (e.g. Smartcat and Matecat); some companies, like Wordfast, offer both free and paid licences, and support both desktop and online versions (Wordfast Pro is desktop-based and paid, whereas Wordfast Anywhere is online and free); and some tools are even open source (e.g. OmegaT). Other systems are also designed to allow for translation on the go – in a smartphone, for example, as is the case of Kanjingo (O'Brien, Moorkens and Vreeke, 2014).

Translators rely heavily on several CAT components to ensure they make the best use of linguistic assets and deliver a target text of superior quality. An overview of these components is provided below.

A. Editor

The main window of a CAT tool, in which translators can consult files and edit text, is commonly called the Editor (Garcia, 2014). This workspace integrates other linguistic resources, as further detailed below.

B. Segments

The source text is divided into segments (source segments), to which are associated their target counterparts (target segments). The segmentation process is normally based on sentences and relies on end-of-sentence punctuation markers to delimit segments. However, segments can be composed of any self-sufficient unit, such as titles (Garcia, 2014). The interface component in which the segments are visualised is the central part of the Editor. In the most common layout, source and target texts occupy two main spaces on the screen, divided vertically: source segments are on the left, and their target equivalents are aligned on the right. Nonetheless, several tools allow customisation of this setup to visualise notably target segments below the source.

C. Translation Memory & Concordance

One of the core components of CAT tools is the Translation Memory (TM), a database in which previously translated segments are stored (Bowker and Fisher, 2010). When a similar segment is detected in the current text to be translated, exact or partial ("fuzzy") matches from the TM can be retrieved, which helps translators to maintain consistency. Each TM entry (sometimes called translation unit) often comes with metadata, including the author, time, date, and other relevant information (Garcia, 2014) which can then be provided during a translation project and may help the translator make decisions (e.g. selecting the most recent wording).

When a match is found in the memory, it is used as a pre-translation to populate the target segment. This pre-translation comes with a matching score, which indicates the degree of similarity between the original segment and the one found in the TM. This score provides valuable information to the translator, who can refer to it as an estimation of the editing effort involved to transform the pre-translation into a definitive version. In addition to the similarity with previously translated segments, this score also

accounts for contextual information, as the order in which segments are stored is also considered. Therefore, while identical segments have a 100% score (Exact Matches [EM]), this score might be higher if the segments stored before and after in the memory are the same as in the current text to be translated (Context Matches [CM]). A dedicated window can appear to display the retrieved TM match, highlighting differences with the current source segment. This helps the translator identify which parts of the pre-translation should be edited. Most CAT tools come with a configuration setting allowing either the project managers or the translators themselves to specify a fuzzy match threshold. TM matches below this threshold do not appear as pre-translations, which avoids generating noise (i.e. partial matches which would require a high editing effort). Nevertheless, since the matching mechanism is based on the segment level, certain expressions are not retrieved when the overall segment score is lower than the threshold, which results in failing to account for conventional phraseology (Garcia, 2014). To address this issue, some CAT tools support subsegment matching (e.g. fragment matches in Trados Studio[5]).

In addition, a concordance search functionality allows translators to browse the TM. Translators can search for source or target words, and the concordance feature returns a list of previously translated segments containing the search entry. This list provides contextual information (i.e. context in which the search word appeared and its corresponding source or translated equivalent), which is valuable for gaining a deeper understanding of previous lexical choices and thus producing consistent translations.

D. Term Base

The Term Base (TB) is a glossary of source terms aligned with their corresponding translated terms. In a CAT tool, terms are generally highlighted at the segment level, which allows translators to easily visualise if terms are missing in a target segment. If terms are found in the current segment, a dedicated terminology window also provides additional information from the TB, such as a definition, part of speech, usage status, and sometimes images. Depending on the tool, a TB lookup feature may be available, allowing translators to consult the TB as they complete a project. In case a recurrent term is not already in the TB, or if translators notice a deprecated term, they may modify the TB entries. In certain cases, this change triggers a validation workflow, during which the translator's suggestions must be approved by a terminologist.

E. Quality Assurance

Quality Assurance (QA) checks perform automatic verifications on the target text. These verifications encompass only a limited amount of quality issues for which detection can be automated (e.g. spotting

[5] https://docs.rws.com/fr-FR/sdl-trados-studio-783545/file-options-editor-fragment-matches-422233 (consulted: 30/04/2025)

mismatched numbers, incorrect spacing, identical source segments translated differently, inconsistent use of terminology, etc.). Therefore, while certain checks can cater to language correctness (spelling and grammar), other linguistic factors (fluency and accuracy) cannot be considered. For this reason, these quality checks can be referred to as "formatting" QA checks (Makoushina, 2007) or "formal quality" checks (Segarra Beneyto and Huertas Abril, 2017). When a QA check is triggered, an icon appears next to the affected segment. The translator can also consult a summary of the different warnings detected and jump to the corresponding segments by clicking on the errors in the list.

Due to their automated nature, these checks are often very generic and may not be suitable for all language pairs or types of projects. As a matter of fact, end punctuation differs across languages: [;] is a semicolon for most languages but a question mark in Greek, and certain languages, like Thai, do not use any punctuation. These discrepancies inevitably create issues when trying to apply a generic logic. Certain tools allow for the adaptation of QA checks (e.g. considering non-breaking spaces before certain punctuation marks in French). Without customisation, QA checks favour recall over precision (do Carmo, 2017). In other words, they detect as many errors as possible, regardless of the accuracy of detected errors, which often results in a high proportion of false positives. Consequently, users often complain about the limited capabilities of QA components and the noise they generate when the number of false positives is high (Makoushina, 2007), which calls for further research on automatic verifications.

The first QA tools were standalone software (Garcia, 2014) which could take a file in the XML [Extensible Markup Language] Localisation Interchange File Format (XLIFF) as an input and perform a quality analysis. Some of these tools are known to provide more extensive functionalities and be more customisable than QA checks in a CAT tool. They are commonly used by translators and revisers to double-check the quality of a translation before delivery. Examples of such applications include ApSIC XBench, Yamagata QA Distiller, ErrorSpy, Okapi CheckMate, Acrocheck and Verifika. Although some of these tools are still used separately today, QA is now integrated into CAT tools and can thus be used within the same software.

F. Preview

Depending on the CAT tool and the type of source file, a dedicated preview panel can show the original file, which is useful for determining the placement of a specific segment in the resulting document (for instance, differentiating paragraphs, headers, footnotes, etc.). Certain tools even support dynamic navigation between the preview and the segments (i.e. clicking on a section in the preview takes the translator to the corresponding segment, and vice versa), and a target preview (in which the source text is automatically replaced with target content on the fly).

G. Text Processing Functionalities

CAT tools support common text processing functionalities, which are typically found in Microsoft Word and other text editing software. Spellchecking, search and replace, autocomplete, formatting, symbols insertion and keyboard shortcuts are relevant examples of such functionalities. Further text processing features are specific to CAT tools. These include, in particular, copying source content to the target side (which can be useful when the source contains only proper names, for example) and merging and splitting segments (which can be required to avoid discrepancies in the TM, when a sentence in the target corresponds to two source segments, for instance). Moreover, apart from traditional formatting (i.e. applying bold, italics, etc.), CAT tools also come with tag handling features, which are essential when the source file processing generates tags (i.e. special symbols which contain formatting information or hyperlinks).

H. Revision and Quality Assessment

Certain functionalities are available only for specific jobs outside translation itself. During revision (sometimes also called proofreading, i.e. when a second person checks the translation), Linguistic Quality Assessment (LQA) features might become available. LQA consists of annotating errors found in the translation based on a predefined error typology. Each annotation is assigned a severity level, which corresponds to a certain number of penalty points. Upon completion of the annotation process, a quality score can be calculated and compared against a threshold to determine if the translation meets quality requirements. As for QA checks, although this feature is natively supported in some CAT tools, standalone applications specialise in LQA. Relevant examples include ContentQuo and TQAuditor.

I. User Interface

Apart from CAT tool features themselves, the way functionalities and linguistic resources are presented at the interface level varies from one tool to another. In particular, the information visible on the screen as well as the functionalities available can differ considerably, and users tend to have better experiences with leaner interfaces (Kappus and Ehrensberger-Dow, 2020). This suggests that finding a balance between empowering the users by making advanced functionalities available to them and keeping the interface attractive and intuitive is a challenging task, which would require further investigation. Moreover, most CAT tools also integrate project management features which are used by both translators and project managers to handle translation requests. Apart from the Editor view, CAT tools come with a project window which lists all current projects and allows for navigating between one assignment and the other (Garcia, 2014). Certain tools also include additional functionalities related, in particular, to term extraction and text alignment.

1.2.2. MT: the Neural Paradigm and Large Language Models

The recent advances in deep learning led to the creation of state-of-the-art models, and today NMT is considered to reach high performance levels, especially for resource-rich language pairs and domains (Chu and Wang, 2018). Nevertheless, despite the productivity gains attributed to the use of NMT, PE is not necessarily an easier task, compared with translation from scratch. Indeed, it can be a demanding exercise, as it requires a different skillset and, particularly, a strong ability to pay attention to detail. As a matter of fact, MT is prone to certain errors (Castilho et al., 2017) and PE might involve a greater cognitive effort (Xiao and Muñoz Martín, 2020).

It should also be mentioned here that a shift from NMT to LLM-based MT seems possible today. In terms of technology, while both NMT and LLM-based MT rely on deep learning principles, the latter incorporates a broader range of training data, allowing the models to capture more nuances. In particular, LLMs typically display a better handling of long-range dependencies and context, which can lead to the production of more accurate translations compared to traditional NMT. When comparing the quality of translations produced by these systems, LLM-based MT proved to achieve superior performance in specific aspects, notably those related to context handling. Notably, Wang et al. (2023) showed that LLMs achieved better results for document-level translation; Raunak et al. (2023) found that LLMs tended to produce less literal translations; and Briva-Iglesias, Cavalheiro Camargo, and Dogru (2024) found that LLMs were rated by human evaluators as producing contextually adequate and fluent translations, sometimes outperforming traditional NMT systems in the legal domain. However, while a possible shift to LLM-based MT could be considered by MT providers, the main challenges lie in scalability and latency (Zeng et al., 2024). These observations could lead to hypothesising that LLM-based MT involves a reduction in cognitive effort for post-editors; however, since this technology emerged recently, further research is needed to confirm this claim.

1.2.3. MT + TM integration

As Melby (1982) foresaw when describing the third level of the translator's workstation, MT is now commonly integrated into CAT tools. The different CAT tools available in the market offer either their own automatic translations or integrations with third-party MT providers. Several options exist to implement MT in CAT. Zaretskaya, Corpas Pastor and Seghiri (2015) suggested classifying the different integration mechanisms by distinguishing internal and external integration. Internal combining processes include, among other methods, segment assembly (i.e. using retrieved TM fragments to substitute corresponding source fragments) and using MT to repair fuzzy matches (i.e. using retrieved TM fragments and applying MT to the rest of the segment). In contrast, in an external integrated environment, one or more suggestions are provided in addition to TM matches and can be generated

online or offline. Kanavos and Kartsaklis (2010) observed increased productivity when MT was integrated in addition to TMs into a CAT environment, especially when the MT suggestions were produced in real time. More recently, Caro Quintana (2021) proposed studying the benefits of external integration by comparing different scenarios (namely, MT alone, TM alone and the external integration of MT and TM). It should be pointed out here that the translator often does not have much control over the incorporation of MT suggestions in CAT tools. It is indeed common for translators to be provided with a pre-translation composed of TM matches and MT suggestions. This clashes with the ideas of Melby (1982) and the concept of interactive MT integration presented by Heyn (1996), which both described systems in which translators would be in control and could decide on the level of assistance needed at all times. This limitation commonly imposed in the modern translation workflow thus appears as a deprivation of translators' freedom.

1.3. Latest Advances

Due to continuous technological advances, CAT tools are constantly evolving. At the conceptual level, the idea of the "Translator's Amanuensis 2020" (TA2020) was introduced by Alonso and Nunes Vieira (2017) as an adaptation of Kay's Translator's Amanuensis to the task of PE. This PE environment would incorporate a set of modern functionalities, going from a knowledge feature (i.e. identifying keywords and providing definitions), an effort prediction function (i.e. estimating the remaining time to completion for the current task) to a 3D visualisation (i.e. displaying the source and target content in two layers). This environment could also identify when the post-editor encounters difficulties (using gaze data, for example) and provide tailored assistance. This section provides an overview of various innovative ideas that have emerged in recent years to optimise translation environments.

1.3.1. TMs & TBs

With TM matching based on entire segments, subsegment matching has been introduced to capture fragment matches regardless of the overall matching score of the segment (Kappus and Ehrensberger-Dow, 2020), helping translators produce consistent content. Indeed, numerous studies have questioned the matching mechanisms behind TMs and argued that more elaborated retrieval mechanisms, such as linguistic transformation (Djabri and Caro Quintana, 2021) or semantic textual similarity (Ranasinghe, Orăsan and Mitkov, 2020) would be beneficial for translators.

A common issue when using both TMs and MT is the discrepancy of the lexicon used depending on the pre-translation origin (Heyn, 1996). This phenomenon can be solved by forcing an MT system to use TB entries. For example, Dinu et al. (2019) proposed augmenting the input text with inline terminology annotations which specify the target term to be used. Apart from ensuring TB entries are being used,

this also ensures that terminological consistency is maintained in the translated text. In addition to ensuring a standard use of terms, recent developments have also focused on improving the terminology matching mechanism. Although Kay and Melby already advocated for morphology-aware processing of terms, the term recognition algorithms have remained rather rudimentary until recently. Terms were therefore recognised only if they matched exactly TB entries (or a part of TB entries if sub-term matching was allowed). This logic inevitably leads to issues when an inflected form is used in the target and not recognised, triggering more QA warnings than necessary (false positives). More importantly, when an inflected term is used in the source and is not detected as a term, there will be no warning to notify the translator that a TB entry must be used in the target (false negative). Some CAT tools support a certain degree of flexibility via fuzzy matching of terms (Garcia, 2014), which partially addresses the limitations of term recognition. However, setting a low matching threshold may result in a high number of false positives, and conversely setting a high threshold may entail leaving certain terms undetected if an inflected form significantly different from the TB entry is used. Hence, morphology intelligence appears as an optimal solution. Morphology-aware term recognition has recently been implemented in CAT tools, such as Phrase[6] and Smart Editor[7].

1.3.2. Quality Estimation

The score of TM fuzzy matches provides a good indication of the amount of editing necessary for a given segment. Similar information can be shown in CAT tools for segments in which MT is applied, using Quality Estimation (QE). Indeed, QE models aim to predict the quality of an MT output without any access to a reference. They can thus be used to generate MT confidence scores, which linguists could rely on to estimate the editing effort and decide if retranslating from scratch would be a better option. The use of confidence scores has been suggested early on, notably by Melby (1982), and has been advocated in several publications. Notably, Moorkens and O'Brien (2013) found that post-editors would like to be presented with confidence scores. According to Vardaro, Schaeffer and Hansen-Schirra (2019, p.4), "quality estimation models can be regarded as a parallel to the human process of error recognition". It is thus understandable how such models can assist translators, notably by allowing them to identify errors more rapidly. Béchara et al. (2021) discovered that displaying confidence scores improved efficiency and reduced cognitive load. Nevertheless, the performance of this method is highly dependent on the capacity of QE models to provide accurate results, and it can also be argued that such

[6] Phrase comes with a "Rich Morphology" option for terms: https://support.phrase.com/hc/en-us/articles/5709733389084-Term-Morphology-TMS (consulted: 30/04/2025)

[7] In Smart Editor, terminology matching is based on lemmatisation: https://languagewire.zendesk.com/hc/en-gb/articles/16996035474717-Terms (consulted: 30/04/2025)

scores might introduce unnecessary bias. Confidence scores have also proved to be successfully implemented to filter out MT results before sending them to linguists for PE, notably at Unbabel (Moorkens, 2020).

As part of the APE-QUEST project, Alva-Manchego et al. (2021) investigated the trade-off between speed, cost and quality when applying a quality gate in an MT workflow to filter out segments which have a high QE score. This study also explored the effects of different QE thresholds and compared them to a traditional setting (PE of the entire MT output), and to an MT-only setting (no PE). The results showed that the quality gate allowed for reducing both time and cost, while preserving quality levels similar to the traditional setting. Such results seem highly encouraging for the implementation of QE in industrial scenarios. However, one crucial point should be pointed out regarding the quality of the resulting translations. The primary objective of having a quality gate lies in bypassing the human PE step, and instead, sending for PE only the sentences which receive a QE score below a predefined threshold. This process leads to PE of isolated sentences, which is likely to be detrimental for certain types of texts, especially when contextual information is lost as a result of this filtering. Alva-Manchego et al. (2021) measured quality relying on acceptability ratings for individual sentences. Arguably, examining the impact of the quality gate on a document level and paying special attention to phenomena occurring beyond the sentence level would be a relevant follow-up to this study. Wei, Utsuro and Nagata (2022) argued that traditional word-level QE, which consists in predicting whether MT words are "OK" or "BAD", provides limited assistance due to the insufficient information in the "BAD" category. To remedy this situation, these authors proposed a refined QE model based on extended word alignments, which can classify MT words into more informative tags (namely, "REP" for replacement, "INS" for insertion and "DEL" for deletion).

An early example of QE integration in CAT tools is IntelliCAT (Lee et al., 2021), which supports sentence-level and word-level QE. Sentence-level QE provides a quality score in a percentage format for each segment, while word-level QE allows for highlighting potential issues at the word level (namely, incorrect words and locations of missing words). More recently, during 2023, several integrations of QE technology in CAT tools were announced by well-known players in the language industry. Relevant

examples include Phrase[8], STAR Transit NXT (Hamm and Klein, 2023) and the TAUS[9] DeMT™ Estimate API[10] which can be connected to memoQ[11].

1.3.3. Automatic Post-Editing

Apart from MT, deep learning has also been applied to automate the PE process itself. Automatic Post-Editing (APE) is indeed concerned with the automatic correction of an MT output (do Carmo et al., 2021). Although the margin for improvement is rather narrow given the quality of NMT, neural networks have proven successful in this task as well. APE outputs can thus be integrated into CAT tools as extra translation suggestions (e.g. Pal et al., 2016). Incremental adaptation based on PE input can even be envisaged to further tailor the APE prediction to the ongoing PE task, as suggested by Landwehr, Steinmann and Mascarell (2023) (see more detailed examples of APE in 3.4.1 on Technological Assistance for Post-Editing). APE can also be combined with NMT and QE to enhance translation workflows. RWS[12] Evolve[13] is a recent application of such a technology, as it relies on a QE model to filter segments sent to APE, and also includes an auto-adaptation mechanism, whereby the APE output is fed back to the NMT and QE models for self-improvement, as illustrated in Figure 1.

[8] https://phrase.com/blog/posts/phrase-quality-performance-score/ (consulted: 30/04/2025)

[9] TAUS (Translation Automation User Society) is a think tank gathering users and providers of translation technologies and services

[10] An API (Application Programming Interface) is a set of rules and protocols that allows different software applications to communicate, exchange data, and perform functions with each other in a standardised and controlled way.

[11] https://www.taus.net/resources/blog/taus-memoq-partnership-enables-machine-translation-quality-estimation-in-memoq (consulted: 30/04/2025)

[12] RWS: Randall Woolcott Services (the company now owning Trados Studio).

[13] https://www.rws.com/artificial-intelligence/evolve/ (consulted: 30/04/2025)

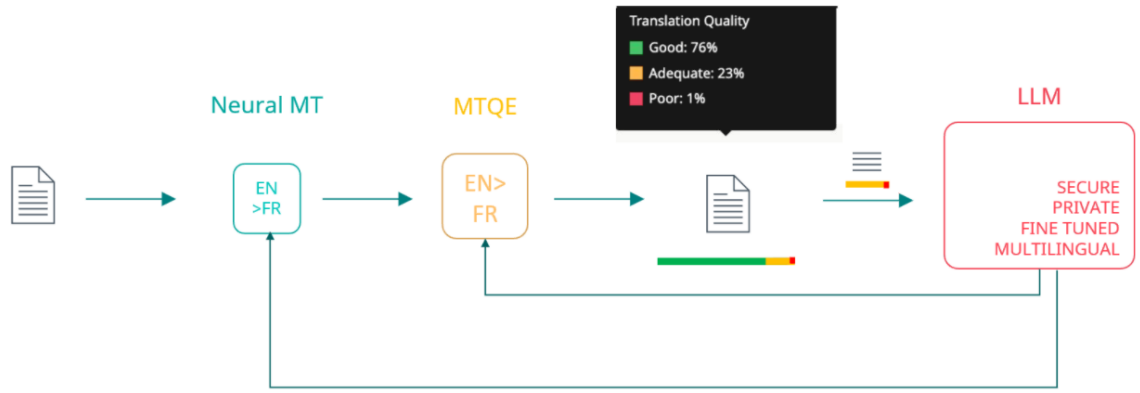


Figure 1: The RWS Evolve workflow.[14]

1.3.4. Interactive Technology

A common issue related to traditional MT systems is that post-editors often have to correct the same errors repeatedly, which results in cognitive friction (see a more extended definition in 3.3.4 A Cognitive Perspective), as pointed out by Ehrensberger-Dow and O'Brien (2015). In PE settings, human-computer interaction (HCI, further discussed in 3.3.2 on Augmented Translation) is limited, because linguists fix outputs without feeding their corrections back to any database – as opposed to TMs, which are populated on the fly with confirmed segments. Green, Heer and Manning (2015) argued that while HCI was the main paradigm in the early days of translation technology research, this perspective was progressively discarded. Although Bar-Hillel (1960) and Licklider (1960) strongly advocated for a partnership between humans and machines, research has tended to focus on providing fully automatic translations. However, adaptive systems can learn from human input and provide informed suggestions in real time. This is precisely how Interactive MT (IMT) models work: they produce a prefix-constrained output based on the input entered by the translator, who can then accept or modify the suggested prefixes. IMT models regained popularity after the introduction of the TransType system (Foster, Langlais and Lapalme, 2002), and several recent projects have proposed methodologies for applying IMT to neural models (e.g. Santy et al., 2019; Peris and Casacuberta, 2019). This technology has also been integrated into CAT tools, such as LILT (Daems and Macken, 2019) and UDAAN (Maheshwari et al., 2023).

Apart from IMT models, interactive PE systems could also be envisaged, especially to address the repetitive editing actions. For example, Simard and Foster (2013) took inspiration from how TMs recycle previous translations and introduced Post-Edit Propagation (PEPr), a model which uses online

[14] Source: RWS webinar "The Evolution of Translation with Linguistic AI", held on 5 December 2023: https://www.rws.com/language-weaver/resources/the-evolution-of-translation-with-linguistic-ai/ (consulted: 30/04/2025)

methods to learn from human corrections on the fly. Similarly, Similarly, Escribe and Mitkov (2023) proposed an incremental approach allowing for a PE model to learn dynamically from user interaction and Terribile (2024a) presented a rule-based algorithm designed to automate repetitive PE actions. Several studies have also found that interactive scenarios significantly reduced the PE effort (Alves et al., 2016; Karimova, Simianer and Riezler, 2018; Domingo et al., 2020), which suggests that integrating interactive technologies could contribute to enhancing the experience of translators.

1.3.5. AI Assistance

Interactive generation of translation suggestions based on the translator's input is also reminiscent of certain disambiguation functionalities in early CAT tools (such as the TransActive facility in TSS or the MIND system). While interactive disambiguation seems not to receive as much interest as IMT, there are some exceptions. For instance, Fairslator[15] (Měchura, 2022) asks for the user input to clarify gender ambiguities. Apart from actively asking for user input, alternative translations can also be retrieved from the MT decoder, as the output is the best ranked hypothesis of an n-best list of candidates. These alternatives can thus be provided upon request from the translator. This is the case in IntelliCAT (Lee et al., 2021), which provides alternatives conditioned on both left and right contexts. This functionality can also be available in standalone MT systems, as is the case in DeepL, which provides alternative word translations. It is also possible to impose certain conditions to an MT system. In the same way that the output can be domain-adapted, the level of formality can be adjusted. In DeepL[16], for example, a functionality allows users to select if they would like the output to be of a formal or informal register. Natural Language Processing (NLP) methods can also be helpful for re-writing translations, and inserting figures of speech for example, which can contribute to yielding similar effects in the target language when the source has a complex style (Toral, 2023). Such textual analyses can also be integrated into QA checks. For instance, XTM announced the release of AI-powered QA checks, designed identify inappropriate language[17]. Following OpenAI's launch of ChatGPT in November 2022, a number of commercial tools rapidly began to implement integrations with Generative Pre-trained Transformer (GPT) technology, and CAT tools were no exception. In May 2023, Matecat introduced an integration allowing for in-context research[18], a feature which provides explanations of fragments of source text based on the context of the source document, directly in the target language. In August 2023, Smartcat

[15] https://www.fairslator.com/ (consulted: 30/04/2025)

[16] https://deepl.com/ (consulted: 30/04/2025)

[17] https://slator.com/xtm-sets-a-new-standard-for-inclusive-language-withai-powered-quality-checks/ (consulted: 30/04/2025)

[18] https://translated.com/matecat-gpt-4 (consulted: 30/04/2025)

announced that AI-based actions (namely, rephrasing, shortening, translating with GPT-3.5, and fixing punctuation and grammar) were made available for users[19]. Similarly, in Trados Studio, the AI Professional plug-in[20] can provide alternative translations based on custom-prompts, thanks to a connection to both OpenAI and Azure's language models.

1.3.6. Multi-Modality

Further interaction patterns between users and tools can also be subject to transformations. Switching attention between the keyboard and the screen can make translation or PE tedious for certain individuals (García-Martínez et al., 2014). Furthermore, with TMs and MT, translators mostly edit pre-translations, which involves extensive usage of certain keyboard commands, thus pausing physical problems and potentially leading to repetitive strain injury (Ehrensberger-Dow and O'Brien, 2015). Such issues could be solved by introducing dictation functionalities (ibid.). The Speech & Eye-Tracking Enabled Computer Assisted Translation (SEECAT) project (García-Martínez et al., 2014) examined the integration of Automatic Speech Recognition (ASR) for PE into the CasMaCat (Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation) tool (Alabau et al., 2013). The results of this study showed that a combination of typing and ASR increased PE productivity. Commercial applications of ASR in current CAT tools currently include the speech recognition feature in Wordfast (Web Speech in Wordfast Anywhere[21]) and the Hey memoQ[22] dictation add-on for memoQ.

As far as research applications are concerned, a few initiatives are also worth mentioning. In order to extend the interaction modalities beyond mouse and keyboard, Teixeira et al. (2019) introduced a tool supporting touch and speech input. During testing, participants expressed positive opinions about the speech recognition feature, however the touchscreen, which was deemed relevant for dragging words (i.e. accounting for the "reordering" PE action, further explained in 3.2.2 on the Phases of the PE Process) received more mixed comments. The Multi-Modal Post-Editing (MMPE) interface presented by Herbig et al. (2020) constitutes another relevant example, as it supports voice commands as well as pen and touch interactions. Overall, participants in this study shared positive feedback on the interface. It was also found that the usefulness of certain types of interactions seemed to correlate with the type of edit to be performed. For instance, pen and touch interactions were efficient for deleting and reordering, but not so much for inserting, especially long fragments of text. This diversification has proven to be particularly efficient for certain types of edits. A possible explanation is that eliminating discomfort

[19] https://www.smartcat.com/release-notes/ (consulted: 30/04/2025)

[20] https://appstore.rws.com/Plugin/200 (consulted: 30/04/2025)

[21] https://www.wordfast.net/wiki/Wordfast_Anywhere_6_Manual (consulted: 30/04/2025)

[22] https://www.memoq.com/product/hey-memoq/ (consulted: 30/04/2025)

contributes to reducing cognitive friction, thus easing cognitive processing (Ehrensberger-Dow and O'Brien, 2015) and improving the overall experience of linguists.

The consequences of these different input modalities should however be highlighted. The use of ASR has proved to increase productivity (which is understandable as the average production rate is higher when speaking compared with typing) and to yield more natural-sounding texts (Chereji et al., 2022). Dictation can also be an attractive solution for disabled users who find it difficult to type on the keyboard for prolonged periods of time. Despite these benefits, ASR also comes with certain drawbacks, such as a higher risk of introducing errors and making more informal translation choices (which is less frequent when writing) and might even increase the cognitive load (ibid.). Other aspects of PE can also be challenging when using ASR. In particular, punctuation, capitalisation and tag placement are more tedious to handle by dictation. Modifying a sentence after having dictated most of it can also entail a higher effort compared to editing it with the keyboard. This observation also holds true for the PE task in general: ASR might be less relevant when having to edit a translation suggestion compared to translating from scratch. Dictation might also entail introducing errors which are different from translation errors, such as words incorrectly transcribed by the ASR system. Such errors will exist in a dictionary (as opposed to misspellings when typing a translation, for example), hence translators need to adapt their methodology in the revision phase to ensure such errors are addressed. Apart from inputting text, multi-modal interactions have also been explored for revising, especially by listening to the target text instead of reading it, which can increase focus and allow to detect certain errors. The Text To Speech (TTS) plug-in for Trados Studio[23], for example, allows for listening to the audio of both source and target segments while working in the tool.

1.3.7. Modularity

In addition to native functionalities, a range of add-ons can now be integrated into both desktop and online tools. Hey memoQ is an example of a such an add-on for ASR. While it was developed specifically for memoQ, other non-proprietary plug-ins can also be integrated into CAT tools, particularly if they are online. In the case of ASR for instance, the InVoice plug-in[24] can be integrated into web-based CAT tools like Matecat or Smartcat. The same goes for certain QA checks, especially

---

[23] https://appstore.rws.com/Plugin/93?tab=documentation#:~:text=Introduction-,Trados%20TTS%20(Text%20to%20Speech)%20is%20a%20plugin%20which%20allows,to%20generate%20the%20audio%20file. (consulted: 30/04/2025)

[24] https://chromewebstore.google.com/detail/voice-in-speech-to-text-d/pjnefijmagpdjfhhkpljicbbpicelgko (consulted: 30/04/2025)

for those related to spelling, grammar and style. LanguageTool[25] is a popular option. It supports a wide range of languages and can spot errors which are left unidentified by QA checkers in CAT tools and standalone QA software, as these typically do not support part of speech tagging, and therefore cannot identify grammatical issues (Miłkowski, 2012). This plug-in is made available directly by certain CAT tools (e.g. Trados Studio) or can be integrated to a browser and used in online CAT tools (e.g. Matecat). External terminology can also be added via plug-ins. This is the case of the Interactive Terminology for Europe (IATE) database[26], which can be incorporated into CAT tools (as done in Wordfast Anywhere or via the IATE Real-Time Terminology plug-in for Trados Studio[27]). Similarly, certain CAT tools allow for configurating functionalities to run external searches in the web, which can be useful to rapidly check a given term or expression in a dictionary or concordance. Relevant examples include Web Search for memoQ[28] and Web Lookup for Trados Studio[29].

[25] https://languagetool.org/chrome (consulted: 30/04/2025)

[26] https://iate.europa.eu/home (consulted: 30/04/2025)

[27] https://appstore.rws.com/plugin/30 (consulted: 30/04/2025)

[28] https://docs.memoq.com/current/en/Workspace/memoq-web-search.html (consulted: 30/04/2025)

[29] https://appstore.rws.com/plugin/4 (consulted: 30/04/2025)

1.4. Summary & Outcomes of this Section

Although translation technology finds its roots in MT, the development of workstations has soon appeared as a valuable research direction to optimise translators' work. CAT tools have greatly evolved since they were introduced on the market, going from basic word processing and terminology support to the integration of TMs and, more recently, MT.

The evolution of translation technology suggests a general pattern whereby the belief in full translation automation fluctuates with technical advances. During periods of optimism, full translation automation seems an achievable goal. Such an ambitious objective has resulted in disappointment, which in turn has led to research focusing on the development of tools to assist translators instead of automating translation. Taking inspiration from Hutchins (2014), Candel-Mora (2021) proposed to divide the evolution of technology for translators into four main periods from the point of view of translators: the Origins (1947-1990), Development (1990-2007), the Google Era (2007-2016) and the Big Data Era (starting in 2016). Therefore, the current focus seems to be on techniques which rely on data, namely machine learning approaches. The vast amount of publications on neural networks for MT and APE indeed indicate a renewed optimism regarding the full automation of translation (Candel-Mora, 2021). However, this wave of optimism seems to be more cautious this time, as the limitations of APE are regularly put forward (do Carmo et al., 2021), which suggests that although a major part of the research is centred on machine learning, this does not reach an extent to which the "all or nothing syndrome" (Melby, 1982) is the only approach.

While science fiction and the media regularly predict the total replacement of translators by MT engines (Candel-Mora, 2021), foreseeing the future of translation technology remains a difficult – if not impossible – task. Pigott (1986) indeed predicted the integration of dictation tools before the 2000s; however, this technology seems to be still under development today. Similarly, although the last level of Melby's Translator Workstation (MT integration) seems to have become standard today, some features of the first level, in particular morphology-aware search and replace (Melby, 1992), are still not available – or not fully developed – in most CAT tools. In the face of such uncertainty, it is worth remembering Melby's observations (Melby, 1992). He recalled that the success of MT since the 1950s has largely depended on the type of source text and the intended use of the translation. Given that there is a predominant need for translation of publishable quality for special-language texts, Melby considered that the Translator's Workstation appears well suited to this situation. Special-language texts are characterised by a certain lexical flexibility (i.e. feature of general language) associated with a restricted terminology (i.e. feature of a sublanguage). With such a combination, the Translator's Workstation appears as an ideal solution, as MT is typically beneficial when working with restricted language, and the remaining problematic parts can be conveniently amended by the translator in an integrated environment. Assuming that correction by a human translator will remain necessary when the aim is to

distribute the translated text, Melby considers workstations as a "long-term solution to international communication problems" (p.164).

With the advances in machine learning, NMT technology has reached unprecedented quality levels and is now commonly integrated into CAT tools. While this integration was predicted several decades ago by Melby (1982), this concretisation comes with several implications. While early proposals, such as those by Licklider, Kay and Melby, aimed at using technology to empower translators, it appears that the use of the same technology is mostly imposed on them today, as CAT tools and NMT are standard practises in the industry, and not adapting to those might have detrimental consequences for a freelancer. Mossop (2006) even argued that the adoption of technology in translation served business purposes, notably the need to produce translations at an increasingly fast pace.

Consequently, a concerning phenomenon can be noted, as despite the integration of MT and the resulting standardisation of PE, very few PE-specific features have been included in CAT tools. Among these, confidence scores (e.g. Lee et al., 2021) and alternative translations (e.g. with the AI Professional plug-in for Trados Studio), including APE suggestions (e.g. Herbig et al., 2019; Landwehr, Steinmann and Mascarell, 2023) seem to be the most common additions for the PE environment, but rather than adaptations to the needs of post-editors, they emerged mostly as consequences of the advent of machine learning techniques behind these methods. This underscores an urgent need to systematically assess the experience of post-editors in translation tools.

2. Assessing the Post-Editing User Experience in Translation Tools

Technological advances have revolutionised the translation environment, and continue to do so today. In this context, assessing the user satisfaction with translation tools is paramount. This section explores how the user's perception is typically assessed in translation technology research and compares typical methodologies employed with the frameworks developed in user experience research.

2.1 Practises for Assessing Translation Technology

Following the rapid technological developments, a plethora of user-centred studies has been conducted to investigate how translators interact with tools and to identify user needs. These studies typically employ a wide range of methods, depending on the objective and resources available. Lagoudaki (2008) listed seven major elicitation methods: observation, think-aloud protocols, scenarios, prototyping, focus group discussions, interviews and surveys. In the field of software development, the process of understanding needs and defining objectives is known as requirements engineering (Sommerville, 2011). While a part of the requirements is purely technical (system requirements), another aspect is more practical and user-oriented (user requirements). Requirements elicitation (also called requirements discovery) plays a significant role in requirements engineering. Sommerville (2011) distinguished four main methods for requirements elicitation: interviews, scenarios, use cases and ethnography. Based on these classifications and considering various user studies, user-centred methods for evaluating translation tools can be classified into five main categories, as described below.

2.1.1 Comparative Analyses

Acquiring a solid knowledge of different translation tools and studying their differences (e.g. which features they offer and how effective they are, as done for QA checking in Makoushina [2007]) appears as a relevant first step. Conducting comparative analyses allows for gaining further insight into the spectrum of technological solutions available to translators. In addition, another outcome of comparative studies is the identification of features present in most tools and other functionalities which are less common and could thus require further investigation. For instance, Nunes Vieira and Specia (2011) compared translation tools from the perspective of PE based on a set of criteria comprising:

- interface intuitiveness
- existence of a spell/grammar/style checker
- use of MT outputs from multiple systems
- integration between TM and MT
- existence of an "auto-complete" function

- preservation of source text formatting
- existence of a QA function
- possible log from user's feedback
- data confidentiality
- existence of scores for TM fuzzy matches

While all these categories seem relevant, this study did not provide a thorough justification for the criteria definition. Finding which criteria are the most important to users is not a trivial task. Although certain aspects may seem evident (e.g. "interface intuitiveness"), others may be more difficult to define in an optimal way (e.g. in the case of "existence of a QA function", the QA feature may exist but may not satisfy certain user requirements). In addition, not including users directly at the criteria definition stage implies the risk of failing to include some criteria which are considered crucial for users.

2.1.2 User Feedback Collection

One of the most straightforward user-centred methods consists of directly asking translators for their feedback. This can be achieved via the dissemination of surveys, or by conducting interviews. The answers obtained in such contexts are relevant for understanding how translators interact with the tools they use (e.g. which features are essential to them) and for identifying their perceived difficulties and pain points (e.g. which features could be improved). This information is important for identifying gaps between the available technology and real user needs (Candel-Mora, 2017). They can also be used to collect opinions on possible innovations. Similarly, it is common for companies to hold focus group discussions (also known as brainstorming sessions) to collect opinions and exchange ideas between relevant stakeholders (Lagoudaki, 2008). Some examples of user feedback collection in the literature include a user survey about TMs by Lagoudaki (2006), a survey about CAT tools by Zaretskaya, Corpas Pastor and Seghiri (2017), and a survey centred on translation-oriented terminology management by Candel-Mora (2017). Similarly, the user survey in Candel-Mora (2016) provided insights into various components of CAT tools via a feature scoring.

2.1.3 Scenarios

Scenarios constitute another way of identifying requirements based on simulations. They typically involve outlining a series of the most common actions in a specific context, which is a valuable method to discuss ideas based on real-life examples. User stories are often used in the software development process to describe the typical behaviour of users interacting with a product via modelling the different actions taken, and therefore, to identify their needs (Lagoudaki, 2008). Similarly, use cases are employed to define possible situations in which users interact with a system (Sommerville, 2011). Prototyping also

belongs to the scenario-based requirement discovery process, as this technique entails designing a prototype of the desired system and presenting it to potential users to collect their opinions (Lagoudaki, 2008).

2.1.4 Observation

In certain cases, it might be advisable to conduct empirical research based on observation (or "ethnography", in the words of Sommerville [2011]). Such experiments typically explore the differences between performing a translation job with distinct types of technological assistance. For instance, this research may involve comparing PE with and without confidence scores, internal and external integration of MT and TMs, or even different CAT tools (e.g. Ehrensberger-Dow and O'Brien, 2015). A wide range of methodologies could be applied to conduct such comparisons. Among various indicators, the PE effort (which can be temporal, cognitive or technical) is a relevant one (Krings, 2001; for a more detailed definition, see section 3.2.3 Studying the Post-Editing Process).

Other types of data can also be collected. For instance, Ehrensberger-Dow and O'Brien (2015) relied on screen recordings, which could be relevant for mapping and understanding a specific series of actions. Another option consists in evaluating the quality of translations produced in different scenarios. The primary aim of the proposed research was to formulate recommendations to enhance translation tools rather than evaluating the resulting translations. Although quality itself can be an elusive concept, evaluating translations obtained in different settings could allow for determining whether translators are prone to making certain types of errors in a specific environment. Such evaluations are generally performed based on commonly used models (further detailed in 4.2 on Approaches to Translation Quality Assessment), such as the Dynamic Quality Framework (Görög, 2014) or Multidimensional Quality Metrics (Lommel, Uszkoreit and Burchardt, 2014).

2.1.5 Standardised Evaluation Models

It appears that, although numerous studies have endeavoured to evaluate translation tools, they have adopted a variety of methods rather than followed a standard procedure. However, some attempts to propose standardised evaluation processes have been introduced. The Expert Advisory Group on Language Engineering Standards (EAGLES), a group associated to the Linguistic Research and Engineering (LRE) programme of the European Commission, designed one of the foundational models for language technology evaluation. The objective of EAGLES was to establish a common evaluation framework for NLP systems, which resulted in the publication of an extensive Evaluation of Natural Language Processing Systems Final Report (EAGLES Evaluation Working Group, 1996). The EAGLES model finds its basis in the ISO 9126 standard "Software engineering — Product quality (Part

1: Quality model)" (ISO/IEC, 1991)[30], which was further adapted for the evaluation of language technology.

In ISO 9126, the concept of quality was divided into six characteristics (functionality, reliability, usability, efficiency, maintainability and portability), to which EAGLES added a seventh dimension (customisability). The EAGLES model revolved around a sound identification and description of the elements to be studied, which included the users (embedded in a context of use), the system under evaluation, and the measures and methods to be employed. A special focus was placed on the user dimension, since the description of the user profile was expected to result in the attribution of weights for each system feature (King and Maeggard, 1998). Among the evaluation procedures described in this report, one is *feature inspection*, which can be completed based on *feature checklists* and is relevant to compare the capabilities of different systems.

Furthermore, the initial work of EAGLES was extended in EAGLES-II (EAGLES Evaluation Working Group, 1999), which also elaborated a quick recipe comprising seven steps to follow for the evaluation of language technology systems. The recommended steps were as follows:

1. Why is the evaluation being done? (determine the purpose of the evaluation and the specify system to be evaluated)
2. Elaborate a task model (identify the users of the system and the tasks they perform in the system)
3. Define top level quality characteristics (identify the features of the system to be evaluated and specify their importance)
4. Produce detailed requirements for the system under evaluation, on the basis of 2 and 3 (find a way to measure the performance of features considered important)
5. Devise the metrics to be applied to the system for the requirements produced under 4 (identify the measure and method to be employed in order to obtain this measure for each item and sub-item; define expectations in terms of satisfactory scores)
6. Design the execution of the evaluation (find the test materials and define the test circumstances)

[30] While this version of the standard is first discussed as it was the latest one when EAGLES was developed, it should be mentioned that this standard was subsequently withdrawn and replaced by ISO 25010:2011 on "Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models" (ISO/IEC, 2011), which as in turn replaced by three different standards, namely:

- ISO/IEC 25002:2024 on "Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model overview and usage" (ISO/IEC, 2024)
- ISO/IEC 25010:2023 on "Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Product quality model" (ISO/IEC, 2023)
- ISO/IEC 25019:2023 on "Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality-in-use model" (ISO/IEC, 2023)

7. Execute the evaluation (carry out the measurements, compare them with expectations, summarise the outcomes)

The EAGLES initiative led to the creation of a flexible evaluation framework, which could be applied to various types of NLP systems. For example, the Testbed Study of Evaluation Methodologies: Authoring Aids (TEMAA) tested this framework for spellcheckers (Paggio and Underwoord, 1998). Similarly, a subgroup of EAGLES, centred on translation tools, conducted a study of translation aids. Following the EAGLES framework, this study first described user profiles, provided an exhaustive description of the functionalities found in translation tools, and identified a feature checklist. The evaluation system proposed was designed to account for diverse situations thanks to the concept of parametrisable testbeds, according to which user profiles could be used to determine weights.

The work of EAGLES was also extended by another related European programme, the International Standards for Language Engineering (ISLE) project, which, among other contributions, resulted in the creation of the Framework for the Evaluation of Machine Translation in ISLE (FEMTI). Following the EAGLES framework, FEMTI was also an application of the ISO guidelines for software evaluation, specifically designed for MT evaluation. Consequently, FEMTI followed the same principles and proposed a classification of the elements defining the context of use, a classification of the quality characteristics of the MT system, and a mapping between these two classifications allowing to determine the importance of each quality characteristic based on the context of use (Hovy, King and Popescu-Belis, 2002).

Although it was initiated over two decades ago, the EAGLES initiative remains unique in that it brought together various institutions on a multinational level in a shared effort to standardise the evaluation of language technologies. A project of this scale has not been renewed since then, which is regrettable in view of the rapid technological developments. However, as the concept of standardisation remains at the heart of EAGLES' principles, one can hypothesise that many aspects of this framework remain relevant and useful today. In fact, it has served as a foundation for various subsequent studies. Notably, Rico (2001) also stressed the need to establish a user-centred evaluation model for translation tools. According to this author, flexibility is a key element for such a model, as "there is no such thing as a best system, but a best system for a particular situation" (ibid., p.2). She also discussed the difficulties typically faced when handling subjective evaluation statements, as these may result in less credible conclusions. Consequently, she proposed a CAT tool evaluation model inspired by EAGLES. To account for the various aspects influencing the overall assessment, four main categories were considered (client, product, process and system), and further divided into more specific items, as shown in Table 1.

Stakeholder	Feature	Weight	Score (100)
Client	• Compliance with corporate conventions and style • Compliance with corporate terminology • Communication means • Deadlines • Average volume • Required quality • Exchange of materials (glossaries, translation databases) • Pricing model		
Product	• Format specifications • Content specifications • Publishing needs • Reusability • Quality • Portability • Tailoring content • Linguistic consistency • Technical consistency • Front-end • Back-end • Languages involved (does the system support them?)		

Process	• Project size
	• Operational environment
	• Timeframe
	• Budget
	• Future expansion
	• Quality assurance requirements
	• Translation team
	  ○ Teletranslation
	  ○ Training required
	  ○ Material sharing
	• Reviewing needs
	• Quality control
	• Terms and conditions
	• Outsourcing needs
	• Running projects in parallel
	• Quotation
	• Management of updates
	• Testing results
	• Scalable work process		

System	• Accuracy (precision and recall) • Interoperability • Compliance with standards • Security • Portability • Usability • Efficiency • Maintainability • Backup and service • Pricing policy • Investment • Customisation • Update policy		

Table 1: Evaluation checklist in Rico (2001, p.7).

The items under “Client” comprised several client details, as well as compliance with specifications. The “Product” dimension referred to the translated texts, and thus included quality measures, whereas the “Process” dimension was more centred on project management. The “System” category had to do with software qualities and was based on principles described in the EAGLES report. Each value could be assigned a different weight depending on its relevance in a specific scenario, and a performance rating depending on the satisfaction with the performance (which could be optimum, partially satisfactory, or poor). Once this process was completed, a final evaluation score could be calculated (out of a maximum of 100 points). It should be noted that although this model is based on numerical values, it does not entirely eliminate subjectivity, which is most likely reflected in the assignment of weights and performance values. However, this is not necessarily a negative phenomenon, given the importance of user perception.

Zaretskaya (2016) also proposed a method for evaluating CAT tools based on the work of EAGLES and the ISO 9126 standard for software quality. This evaluation method revolved around three pillars derived from these previous works: functionality (features designed to meet user needs), interoperability (ability to operate in multiple systems) and adaptability (capacity to adjust to a determined situation). These pillars were directly related to CAT tools, by linking functionality to translation and word processing features (e.g. TM matching) and by understanding interoperability as the ability to work with various file formats, and adaptability as the adjustments of certain settings. The resulting attributes are exemplified in Table 2.

This method therefore relied on a feature checklist, associated with a weighting system. Unlike Rico’s model, in this case the weights were pre-determined, based on the popularity they had in a user survey. A second metric then specified the presence or absence of each item (0: absent, 1: partial integration, 2: present). This metric was then multiplied by the weight. The numbers thus obtained for all items were added, which gave a total score. While the evaluation model allowed for flexibility by adjusting the weights, relying on a feature checklist comes with certain consequences. The most salient one is that tools can evolve rapidly, and as new features are launched, other ones can become deprecated, which threatens the replicability of any feature-based evaluation method. In that regard, the EAGLES report specified that a feature checklist had to be “both standardised in the sense that it should not be constrained by particular situational variables, and open, in the sense that it can cover different approaches to a problem without being prescriptive in nature” (EAGLES Evaluation Working Group, 1996, p.35). Arguably, providing a precise description of the functionalities considered while still allowing for flexibility can be a challenging task. In that regard, focusing on task objectives could be a relevant solution. In addition, as acknowledged by Zaretskaya (2016), although this method is strongly user-centred (given that the weights are derived from the importance of each item based on the results of a user survey), it fails to account for any perception of usability. In other words, an important feature

(i.e. maximum weight) can be present as defined in the specifications (maximum presence metric value), but its utilisation might be cumbersome or not intuitive enough.

		Metric	Weight	[Tool 1]	[Tool X]
Functionality	Concordance	0-no, 2-yes	3		
	Auto propagation	0-no, 2-yes	2		
	Aligner	0-no, 2-yes	1		
	Storing TM in the cloud	0-no, 2-yes	1		
	Real-time QA	0-no, 1-not real-time, 2-yes	2		
	Access to online TM	0-no, 1-with plugin, 2-yes	1		
	Access to online term. res.	0-no, 1-with plugin, 2-yes	2		
	Sub-segment suggestions	0-no, 1-with plugin, 2-yes	1		
	Real-time target preview	0-no, 1-generate preview, 2-real-time	2		
	Good grammar checker	0-no, 2-yes	1		
	Merge TMs	0-no, 2-yes	1		
	Easily add terms	0-no, 1-not easy, 2-yes	1		
	Segment assembly	0-no, 2-yes	1		
	Dictation	0-no, 1-works w/ Domain Name System, 2-integration	1		
	Simple handling of tags	0-no, 1-some tags, 2-all tags	1		
	Terminology management	0-no, 2-yes	3		
	Machine translation	0-no, 1-with plugin, 2-yes	1		
	Work with > 1 TM	0-no, 2-yes	1		
Total Functionality					
Adaptability	Different Operating System	0-no, 2-yes	1		
	Adjustable keyboard shortcuts	0-no, 2-yes	2		
	Adjustable segmentation	0-no, 1-merge/split segments, 2-yes	2		
	Adaptable/modular interface	0-no, 1-some modules, 2-yes	1		
	Web-based version	0-only desktop, 1-only web, 2-both	1		
Total adaptability					
Interoperability	Share TM	0-no, 2-yes	1		
	Number of TM formats	0-0, 1-1 to 2, 2->3	3		
	Number of document formats	0-<40, 1-41 to 50, 2->50	3		
Total interoperability					
Total score					

Table 2: Evaluation table in Zaretskaya (2016, p.157).

Unlike other studies, Candel-Mora and Ramírez Polo (2015) placed their focus one step ahead in the evaluation process. They emphasised the challenges around the implementation of translation technology in institutional settings and the lack of a standard process. Consequently, they introduced a

decision-making framework to guide the development of translation tools. This work identified key elements to be considered when making development choices and provided a framework based on these factors. It resulted in the creation of a questionnaire, which led in turn to the design of a decision-making matrix. This matrix was composed of five main criteria (namely, users, texts, equipment, resources and budget), for which weights could be adjusted depending on the use case. The proposed framework would be highly relevant to test in real-life scenarios in order to examine its impact on decisions and their effects on translation tools. It would also appear beneficial to extend this matrix to other fields besides institutional settings.

2.2 Usability and UX

It can be observed from the previous section that user perception and overall experience with the tools have played a significant role in translation technology research. Several methods for user studies have been employed, ranging from comparative analyses, user feedback collection, observation and scenario-based studies to more standardised evaluation models. However, observing how a user interacts with a product is the central focus of an entire field of study – User Experience (UX) research. This section therefore presents the concepts of usability and UX, together with common UX evaluation frameworks, before discussing how these have been applied to assess translation technology.

2.2.1 Definitions

Usability is concerned with the degree of ease with which products allow their users to achieve their objectives. UX adopts a more comprehensive approach, considering the overall experience of a user, including a more emotional dimension, thus considering the enjoyment related to the use of a system (Rusu et al., 2015; see also the definition of the ISO 9241 standard below). The distinction between these two concepts is sometimes subject to debate. In that regard, Bevan (2009) posited that they can be distinguished by their objective: while the aim of usability methods is to improve human performance, that of UX methods lies in improving user satisfaction.

Related notions, such as usefulness and utility, should also be clarified. According to Nielsen (1994, p.24), usefulness is the "issue of whether the system can be used to achieve some desired goal". This concept implies two dimensions, namely utility (i.e. extent to which the system solves user needs), and usability, which can be further divided into several factors, as mapped in the Figure 2 on system acceptability.

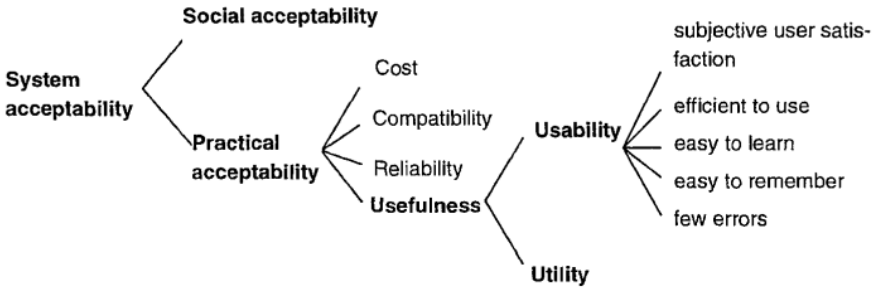


Figure 2: System acceptability dimensions. Source: Boisen, Bygholm and Hejlesen (2002, p.828, adapted from Nielsen [1994]).

Boisen, Bygholm and Hejlesen (2002) built on Nielsen's model and applied Activity Theory to propose a representation of the relations between the subject, the case and the system. Activity Theory is a conceptual framework for understanding human behaviour and consciousness beyond the perspective of individual actors (Vygotsky, 1978). It describes human endeavours as grounded in purposeful, socially situated activities. According to this framework, human endeavour ("activity") takes the form of interaction between humans ("subjects") and the objectives they would like to accomplish ("case") via the use of artifacts ("systems"). Based on this description, the authors derived three types of relations (i.e. usability, utility and copability), as represented in Figure 3.

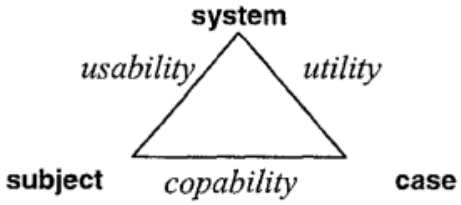


Figure 3: The three types of relations in the Activity Theory model. Source: Boisen, Bygholm and Hejlesen (2002, p.829).

In this model, usability refers to the ability of a subject to operate a system. Utility represents the appropriateness of the system to address a case. A third relation, copability, describes the capacity of a subject to cope with a case. This last relation is paramount, as high usability and utility are not relevant in case the subject is unable to handle the challenge posed by the case.

According to the ISO 9241 standard on Ergonomics of Human-System Interaction (ISO, 2018; more specifically Part 11 on "Usability: Definitions and concepts"), usability can be defined as 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". The three pillars of usability identified in this definition are further defined as follows.

- Effectiveness: "accuracy and completeness with which users achieve specified goals"

- Efficiency: “resources used in relation to the results achieved (typical resources include time, human effort, costs and materials)”
- Satisfaction: “extent to which the user’s physical, cognitive and emotional responses that result from the use of a system, product or service meet the user’s needs and expectations”

In addition, the context of use refers to the “combination of users, goals and tasks, resources, and environment”. In the same standard, UX is defined as the “user’s perceptions and responses that result from the use and/or anticipated use of a system, product or service”. It is also specified that the user’s perceptions and responses include “the users’ emotions, beliefs, preferences, perceptions, comfort, behaviours, and accomplishments that occur before, during and after use”. Therefore, UX is a broader concept, and usability can be viewed as a dimension belonging to it. This is reflected in Guo (2012)’s proposal to divide UX into four main pillars, namely:

- Value: is it useful?
- Usability: is it easy to use?
- Adoptability: is it easy to start using?
- Desirability: is it fun and engaging?

Other frameworks have been presented in the literature, and although significant overlaps can be found in the different definitions, certain differences are relevant to mention. Nielsen (1994) identified five usability attributes. Two of them can also be found in the ISO 9241 standard (ISO, 2018):

- Efficiency, which Nielsen related to productivity
- Satisfaction, which Nielsen described as a subjective appreciation

The three other dimensions comprise:

- Learnability: “the system should be easy to learn so that the user can rapidly start getting some work done with the system”
- Memorability: “the system should be easy to remember, so that the casual user is able to return to the system after some period of not having used it, without having to learn everything all over again”
- Errors: “the system should have a low error rate, so that users make few errors during the use of the system, and so that if they do make errors they can easily recover from them”

Quesenbery (2003) also presented five dimensions of usability, gathered in a model called the “5 Es”, according to which a system with a high usability should be

- Effective: “completeness and accuracy with which users achieve their goals”
- Efficient: “speed (with accuracy) with which users can complete their tasks”
- Engaging: “degree to which the tone and style of the interface makes the product pleasant or satisfying to use”

- Error tolerant: “how well the design prevents errors, or helps with recovery from those that do occur”
- Easy to learn: “how well the product supports both initial orientation and deepening understanding of its capabilities”

2.2.2 Assessment Methodologies

While the aforementioned frameworks contributed greatly to defining usability and UX, they do not prescribe any specific evaluation method for real-life scenarios. Thence, it becomes relevant to explore how usability and UX are investigated in practice. Several studies have defined *ad hoc* frameworks adapted to their needs. On certain occasions, those frameworks can be generalised and adapted to different use cases. For instance, Wachowicz et al. (2008) designed a dynamic usability framework tailored to spatial domain applications, particularly web mapping. This framework is based on a set of abstraction levels, namely:

- Usability hypothesis: specific proposition to be tested
- Usability typology: this dimension is defined by four components (the content, interaction models, interaction styles, and interaction objects)
- Usability variables: values which can be assigned during testing, decomposed in four components (users, purpose, tasks and skills)
- Usability elements: measurable attributes (e.g. clarity)
- Usability measures: quantifiable parameters (e.g. time to learn)

For illustration purposes, Table 3 provides further details of how this framework was applied by Wachowicz et al. (2008) in the case of web mapping applications. Apart from providing guidance on which attributes to select, this work also clarified the overall structure of a usability framework, and more specifically, how the theoretical objectives (the initial usability hypothesis) are articulated and derived in more and more practical dimensions (variables and elements) until reaching concrete values (measures). Arguably, such a framework could be beneficial to several fields outside the spatial domain, including translation technology.

USABILITY FRAMEWORK
USABILITY HYPOTHESIS Web mapping applications exhibiting a higher degree of usability will be associated with greater perceived user satisfaction.
USABILITY TYPOLOGY 1. Content: route display (static, interactive, or animated), route description, route information services with maps. 2. Interaction modes: visual, auditive and tactile. 3. Interaction styles: direct manipulations (zooming, panning, scrolling, resizing, pointing), querying, printing, command language (or command entry), form fill -in, menus (pull down, pop-up, hierarchical). 4. Interaction objects: buttons, sliders, menus, progress bars, alerts, wizards, undo, form or field validation, cancel, hyperlinks
USABILITY VARIABLES 1. Users: Who are the users? 2. Purpose: What are their goals? 3. Tasks: What kind of tasks the users need to perform in order to achieve their goals? 4. Skills: What kind of skills do the users have?
USABILITY ELEMENTS 1. Readability – All essential components of the route, especially the roads should be visible and easily identified. 2. Clarity – Routes must be clearly marked and readily distinguishable at a quick glance. The map should only contain the amount of information necessary to highlight the chosen route. 3. Completeness – All information necessary for navigation must be provided. 4. Convenience – For route maps used while travelling, they should be easy to carry and to manipulate; also, for route maps in the browser window, maps should not require scrolling or resizing
USABILITY MEASURES 1. Time to learn: How long does it take for users to learn how to use the commands that are relevant to a set of tasks? 2. Speed of performance: How long does it take to carry out a task or a set of tasks? 3. Rate of errors: How well do users maintain their knowledge after set period of time? (perhaps related to frequency of use)

Table 3: Abstract levels and their attributes in the usability framework proposed and applied in Wachowicz et al. (2008, p.402).

While this framework provided valuable guidance in terms of structure, the usability elements to be included may differ depending on the case at hand and may be more or less difficult to define depending on the system to be evaluated. Furthermore, as discussed above, several usability models have been introduced, and although various dimensions can be found in multiple models, some discrepancies remain. More recently, in a usability study of Geographical Information Systems (GIS), Rahman et al. (2022) performed an exhaustive literature review and extracted 68 usability elements, grouped by three variables and 11 dimensions, as summarised in Table 4.

Variables	Dimensions
Quality	Functional suitability
	Performance efficiency
	Compatibility
	Usability
	Reliability
	Security
	Maintainability
	Portability
User satisfaction	/
Individual work performance	Task performance
	Adaptive performance

Table 4: Usability dimensions in Rahman et al. (2022).

It should be noted that each dimension contains between 4 and 9 items. For example, "Functional suitability" is further divided into six items (Rahman et al., 2022, p.70):

- "The functions provided in GIS are suitable for me.
- The functions provided in GIS meet with my work requirements.
- The information provided in GIS is accurate.
- The results of the analysis provided from GIS are precise.
- Functions in GIS facilitate me in the business process.
- Functions provided in GIS assisting me in meeting organisational goals."

These models thus demonstrates that usability is a complicated concept encompassing a seemingly endless list of attributes, which entails difficulties when it comes to assessing usability in real-world applications. However, various methodologies have been developed to address this issue. Several metrics exist to quantify usability or UX based on answers to questionnaires. Sauro and Lewis (2016) distinguished two main types: post-study questionnaires, which are used at the end of a study to evaluate the overall usability of a product, and post-task questionnaires, which focus on a more local level by assessing the perceived usability of software to accomplish a specific task and can therefore provide more detailed information. Several examples of commonly employed questionnaires are detailed below based on Sauro and Lewis (2016). An early example of such a questionnaire is the Technology Acceptance Model (TAM; Davis, 1987), which revolves around the notions of perceived usefulness and perceived ease of use. The items of this questionnaire are evenly distributed across these two dimensions, as detailed in Table 5.

Usefulness	1	Using [this product] in my job would enable me to accomplish tasks more quickly.
	2	Using [this product] would improve my job performance.
	3	Using [this product] in my job would increase my productivity.
	4	Using [this product] would enhance my effectiveness on the job.
	5	Using [this product] would make it easier to do my job.
	6	I would find [this product] useful in my job.
Ease of use	7	Learning to operate [this product] would be easy for me.
	8	I would find it easy to get [this product] to do what I want it to do.
	9	My interaction with [this product] would be clear and understandable.
	10	I would find [this product] to be flexible to interact with.
	11	It would be easy for me to become skillful at using [this product].
	12	I would find [this product] easy to use.

Table 5: List of 12 items of the TAM[31].

One of the most widely used post-study questionnaires is the System Usability Scale (SUS) introduced by Brooke (1996), a method that also relies on the usability definition of the ISO 9241 standard. Brooke argued that low-cost assessments of usability were necessary in the industrial context at the time, and elaborated SUS as a "quick and dirty" method – as per the wording used by Brooke himself – for assessing usability. SUS consists of a questionnaire of 10 items (see Table 6 below), in which positive and negative affirmations come one after the other. Respondents select an answer for each item in a Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree).

1	I think that I would like to use this system frequently.
2	I found the system unnecessarily complex.
3	I thought the system was easy to use.
4	I think that I would need the support of a technical person to be able to use this system.
5	I found the various functions in this system were well integrated.
6	I thought there was too much inconsistency in this system
7	I would imagine that most people would learn to use this system very quickly.
8	I found the system very cumbersome to use.
9	I felt very confident using the system
10	I needed to learn a lot of things before I could get going with this system.

Table 6: List of the 10 items of SUS, adapted from Brooke (1996).

[31] Adapted from https://measuringu.com/tam/ (consulted: 30/04/2025)

The final score is a number out of 100. To calculate it, 1 is subtracted from the score (i.e. answer) of odd-numbered items, 5 is subtracted from the score of even-numbered items, and the resulting values are added up and multiplied by 2.5. It is generally recognised that the benchmark is 68, and that values above 80 indicate high usability, whereas values below 50 are a sign that usability should be drastically improved (Sauro and Lewis, 2016).

Although SUS is among the most popular metrics, other usability questionnaires have been proposed. Sauro and Lewis (2016) listed commonly used standardised usability questionnaires. For example, the Questionnaire for User Interaction Satisfaction (QUIS) comes in two lengths (in version 7, short: 41 items, long: 122 items) and distinguishes nine interface factors, including the following elements:

- screen factors,
- terminology and system feedback,
- learning factors,
- system capabilities,
- technical manuals,
- online tutorials,
- multimedia,
- teleconferencing,
- and software installation

The Software Usability Measurement Inventory (SUMI) is a lengthier questionnaire consisting of 50 items, divided into five aspects (comprising 10 questions each):

- "Efficiency: The degree to which the software helps users complete their work
- Affect: The general emotional reaction to the software
- Helpfulness: The degree to which the software is self-explanatory, plus the adequacy of help facilities and documentation
- Control: The extent to which the user feels in control of the software
- Learnability: The speed and facility with which users feel they mastered the system or learned to use new features"

The Post-Study System Usability Questionnaire (PSSUQ[32], Table 7), originally developed as an internal metric at IBM, comprises 16 items (version 3, answers on a 7-point scale) which are distributed across four categories (one general and three subscales):

- Overall
- System usefulness
- Information quality

[32] https://uiuxtrend.com/pssuq-post-study-system-usability-questionnaire/ (consulted: 30/04/2025)

- Interface quality

The Computer System Usability Questionnaire (CSUQ) is a variation of the PSSUQ, which contains the same questions, but with slightly different wording.

1	Overall, I am satisfied with how easy it is to use this system.
2	It was simple to use this system.
3	I was able to complete the tasks and scenarios quickly using this system.
4	I felt comfortable using this system.
5	It was easy to learn to use this system.
6	I believe I could become productive quickly using this system.
7	The system gave error messages that clearly told me how to fix problems.
8	Whenever I made a mistake using the system, I could recover easily and quickly.
9	The information (such as online help, on-screen messages, and other documentation) provided with this system was clear.
10	It was easy to find the information I needed.
11	The information was effective in helping me complete the tasks and scenarios.
12	The organization of information on the system screens was clear.
13	The interface of this system was pleasant.
14	I liked using the interface of this system.
15	This system has all the functions and capabilities I expect it to have.
16	Overall, I am satisfied with this system.

Table 7: List of the 16 items of PSSUQ, adapted from Sauro and Lewis (2016).

A more recent questionnaire, the Usability Metric for User Experience (UMUX), was designed to be consistent with SUS, but contains only four items (Table 8). An even shorter version, UMUX-LITE comprises only two items (Table 9). It should be pointed out that these two components correspond to the two dimensions proposed in the TAM (usefulness and ease of use).

1	This system's capabilities meet my requirements.
2	Using this system is a frustrating experience.
3	This system is easy to use.
4	I have to spend too much time correcting things with this system.

Table 8: List of the four items of UMUX, adapted from Sauro and Lewis (2016).

1	This system's capabilities meet my requirements.
2	This system is easy to use.

Table 9: List of the two items of UMUX-LITE, adapted from Sauro and Lewis (2016).

As far as post-task questionnaires are concerned, relevant examples include the After-Scenario Questionnaire (ASQ) which includes three items (Table 10) and the Single Ease Question (SEQ) comprising only one item (Table 11).

1	Overall, I am satisfied with the ease of completing the task in this scenario.
2	Overall, I am satisfied with the amount of time it took to complete the task in this scenario.
3	Overall, I am satisfied with the support information (on-line help, messages, documentation) when completing the task.

Table 10: List of the three items of ASQ, adapted from Sauro and Lewis (2016).

1	Overall, this task was [very difficult] […] [very easy].

Table 11: Item of SEQ, adapted from Sauro and Lewis (2016).

Apart from post-study and post-task questionnaires, other metrics exist. For instance, the Happiness Tracking Survey (HaTS; Müller and Sedley, 2014) was created at Google following best practises in the field in order to track user attitudes and collect open-ended feedback at a large scale. The survey items have been carefully designed to be as representative of the reality as possible, thus avoiding common biases (e.g. satisficing behaviour). The order of the questions follows a funnel approach, going from overall satisfaction and likelihood to recommend, to perceived frustrations and satisfaction with specific product attributes and specific tasks performed in the product. A more detailed overview of this survey is provided in Table 12.

Survey item	Answer format
1. Overall, how satisfied or dissatisfied are you with [product]?	7-point scale
2. How likely are you to recommend [product] to a friend or colleague?	10-point scale
3. (Optional) What, if anything, do you find frustrating or unappealing about [product]? What *new* capabilities would you like to see for [product]?	Open-ended
4. (Optional) What do you like best about [product]?	Open-ended
5. How satisfied or dissatisfied are you with [product] in the following areas?	7-point scale for each area defined
6. In the last month, which of the following tasks have you *tried* to accomplish with [product]?	List of tasks (randomised order)
7. How satisfied or dissatisfied are you with doing the following tasks in [product]?	7-point scale for list of task (randomised order)
8. How many [weeks/months] ago did you start using [product]?	Enter a number
9. In the last [weeks/months], on about how many days have you used [product]?	Enter a number

Table 12: List of items of HaTS, adapted from Müller and Sedley (2014), italics in original.

These questionnaires are relevant to measure the subjective satisfaction of users and are commonly used by UX researchers in scenario-based or prototype testing settings. These methods are easily replicable and allow for establishing comparison (between tools, users, or against a benchmark) and for reporting quantified results (Sauro and Lewis, 2016).

Among other relevant examples, the Net Promoter Score (NPS) aims at assessing customer loyalty, the Customer Effort Score measures the ease with which customers can interact with a product, and the Emotional Metrics Outcomes evaluate the emotional response of an interaction. The Customer Satisfaction (CSAT) score is another widely used metric in business to measure customer satisfaction with a product, service, or interaction. It originated as a survey methodology to gauge customer sentiment and has become a key performance indicator for customer service and product quality across various industries. CSAT involves a question, such as "How satisfied were you with [experience/ product/service] today?". Responses are typically ratings going from "Very satisfied" to "Very Dissatisfied". The resulting CSAT score is then calculated as the number of positive responses out of the total number of responses, multiplied by 100. Consequently, this score serves as a good metric for determining the number of satisfied customers. These formats are shorter and commonly used for regular satisfaction monitoring.

In addition, other questionnaires do not measure the agreement with predefined statements but rather rely on semantic differential scales to evaluate specific UX criteria. According to Hassenzahl, Burmester, and Koller (2003), common factors include

- Attractiveness: overall experience

- Pragmatic quality: ease of use of the tool
- Hedonic quality: extent to which the experience with the tool is enjoyable

The same authors argued that existing UX evaluation methods were too focused on pragmatic quality and tended to neglect hedonic quality, which is a crucial aspect of UX as it encompasses relevant needs. Therefore, they suggested dividing the hedonic quality aspect into two factors:

- Stimulation: extent to which the features are engaging, motivating and novel
- Identity/Identification: indicates the extent to which the product enables the user to identify with it

Based on this hypothesis, Hassenzahl, Burmester, and Koller (2003) introduced AttrakDiff 2, a UX questionnaire encompassing several hedonic aspects. This questionnaire is composed of 28 bipolar items (Table 13) that should be assessed based on a 7-point Likert scale after interacting with the system being evaluated.

It is essential to point out here that the qualities described here are independent of each other: a tool can be highly pragmatic but poorly hedonic, and vice versa. However, pragmatic quality tends to correlate with other factors (external to the questionnaire), such as effort measures (ibid.).

Factor	Item
Attractiveness	Unpleasant - Pleasant
	Ugly - Attractive
	Disagreeable - Likeable
	Rejecting - Inviting
	Bad - Good
	Repelling - Appealing
	Discouraging – Motivating
Pragmatic quality	Technical - Human
	Complicated - Simple
	Impractical - Practical
	Cumbersome - Straightforward
	Unpredictable - Predictable
	Confusing - Clearly structured
	Unruly - Manageable
Hedonic quality – identification	Isolating - Connective
	Unprofessional - Professional
	Tacky - Stylish
	Cheap - Premium
	Alienating - Integrating
	Separates me - Brings me closer
	Unpresentable - Presentable
Hedonic quality – stimulation	Conventional - Inventive
	Unimaginative - Creative
	Cautious - Bold
	Conservative - Innovative
	Dull - Captivating
	Undemanding - Challenging
	Ordinary - Novel

Table 13: AttrakDiff structure and items[33].

[33] Source: https://www.researchgate.net/publication/328856866_Understanding_user_representations_a_new_development_path_for_supporting_Smart_City_policy_Evaluation_of_the_electric_car_use_in_Lorraine_Region (consulted: 30/04/2025)

Laugwitz, Held and Schrepp (2008) presented the User Experience Questionnaire (UEQ), which consists of a series of 26 items to be rated in a bipolar scale based on a 7-point Likert scale (Figure 4) following an interaction with the tool being evaluated. This questionnaire is based on similar factors as AttrakDiff but aims to focus more on pragmatic aspects. In this questionnaire, the factors of pragmatic quality and hedonic quality are further divided into more specific subfactors:

- Pragmatic quality
    - Perspicuity: ease of use and learnability
    - Efficiency: speed and practicality
    - Dependability: degree of control perceived by the user
- Hedonic quality
    - Stimulation: extent to which the tool is motivating to engage with
    - Novelty: degree of innovation

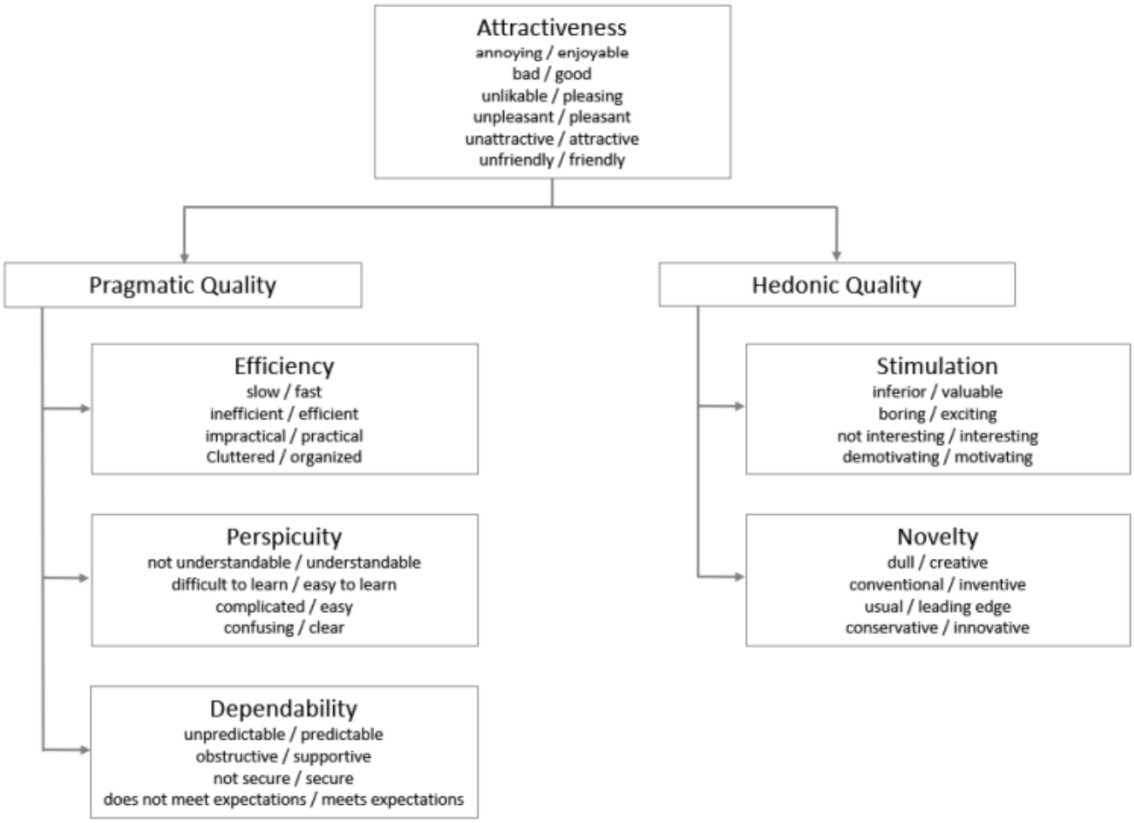


Figure 4: Structure and items of UEQ. Source: User Experience Questionnaire Handbook (Schrepp, 2023, p.3).

The position of the positive and negative terms is randomised in order to avoid having all positive or negative terms on the same side. A condensed version of the UEQ, called UEQ-S (Table 14), consisting of eight items (four belonging to pragmatic quality and four to hedonic quality) was introduced later and

proved to be useful for predicting the results of the extended version (Schrepp, Hinderks and Thomaschewski, 2017).

1	Obstructive – Supportive
2	Complicated – Easy
3	Inefficient – Efficient
4	Clear – Confusing
5	Boring – Exciting
6	Not interesting – Interesting
7	Conventional – Inventive
8	Usual – Leading edge

Table 14: List of items of UEQ-S. Source: Schrepp, Hinderks and Thomaschewski (2017, p. 105).

2.2.3 Applying UX Principles for Translation Technology Assessment

In light of the UX principles and methodologies described above, it appears pertinent to find out to which extent these are applied to translation technology assessment. It should be noted here that the UX perspective was adopted here over usability because UX is a more comprehensive concept. However, usability frameworks are also referred to when deemed relevant.

The model of Activity Theory can be used to describe PE in a CAT tool. More specifically, the system is the CAT tool, the subject is the linguist, and the case is the translation – or more specifically here, the PE project (Figure 5). In this schema, usability refers to the ease with which the linguist can navigate the CAT tool, utility has to do with the suitability of the CAT tool to complete the PE project, and copability is the ability of the linguist to carry out the PE project. While the latter has to do with competence and is independent of the system, it appears relevant to assess the utility and usability dimensions.

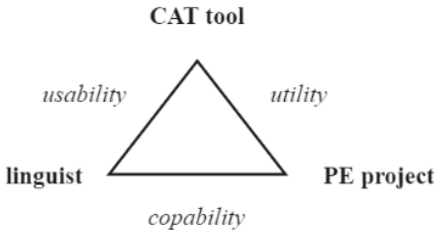


Figure 5: Application of the Activity Theory model to CAT tools for PE.

This representation emphasises that usability is embedded in a broader vision, where utility and copability also play important roles. Along the same lines, it appears relevant to mention here that

translation and PE are not to be considered in isolation, but rather as situated activities. Situatedness in translation, as conceptualised by Risku (2002), describes translation as a context-bound activity shaped by the interplay of social, cultural, and material environments. This perspective diverges from traditional cognitive models that treat translation as an isolated mental process, focusing on how the translator's brain processes linguistic information, such as decoding the source text and encoding it into the target language. These models often emphasise internal mechanisms like memory, problem solving, and linguistic analysis, largely detached from external influences. For example, Göpferich and Jääskeläinen (2009) explored translation competence through cognitive models, emphasising mental processes like problem solving, decision making, and memory retrieval, thus treating translation as an individual cognitive activity, largely independent of external factors. In contrast, framing translation as a situated activity (Risku, 2002) involves recognising that translation does not occur in a vacuum but is deeply embedded in real-world contexts. Translators interact with their physical environment (e.g., CAT tools), social dynamics (e.g., clients, colleagues), and cultural norms (e.g., expectations of the target readership). This approach considers translation as a socially distributed activity, shaped by the translator's engagement with both human actors (like collaborators) and non-human factors (like technology). The present work acknowledges both dimensions but places the focus on the interaction between post-editors and the environment in which they work.

Several studies have explored the usability or UX of MT output alone. For example, Castilho et al. (2014) investigated whether post-edited translations were more usable than raw MT; Doherty and O'Brien (2014) assessed the usability of raw MT; and both Koponen et al. (2020) and Karakanta et al. (2022) evaluated the usability of MT for subtitling from the perspective of post-editors using the UEQ. While these studies offer valuable insights into the usability of MT for post-editors, they do not account for the environment in which PE takes place. Therefore, it appears essential to further analyse the interaction between post-editors and the PE environment they are working in.

In order to assess this interaction, several evaluation methods have been proposed and were presented in 2. on Assessing the Post-Editing User Experience in Translation Tools (namely, comparative analyses, user feedback collection, scenarios, observations and standardised models). These may encompass some UX dimensions, depending on how they are designed. For example, observation-based studies are likely to bring valuable insights regarding the intuitiveness (e.g. learnability and memorability) of a tool. Other methodologies, like comparative analyses and user surveys, are more flexible and may encompass different usability characteristics depending on the use case at hand.

Moreover, in the field of translation technology, some frameworks have also been introduced for evaluating the usability of translation tools. Krüger (2016) built on the definition of usability provided in the ISO 9241 standard (ISO, 2018) to propose a usability model specifically designed for CAT tools (CAT Tool Usability Model [CATTUM]), taking a user-oriented perspective (see Figure 6). This

usability model is embedded into a context of use and comprises four dimensions, three of which are identical to those presented in the standard, namely:

- Effectiveness: "how well a given user can achieve specified goals or perform specified tasks using the software", described as a "qualitative dimension"
- Efficiency: "effort required to achieve these goals or to perform [certain] tasks", described as a "quantitative dimension"
- Satisfaction: "user's satisfaction with the functions and characteristics offered by the CAT tool", described as a "subjective dimension"

In addition to the three usability pillars mentioned above, the author added

- Learnability, which reflects "how easily new users can familiarise themselves with a given software system and which resources they can draw on in this process" (ibid., p.132).

While the four dimensions of usability in Krüger (2016) are clearly defined, the author acknowledged that a certain degree of interdependency exists between them. For example, one could easily imagine that a CAT tool feature which improves efficiency would in turn have a positive impact on satisfaction. This model is meant to be used as a reference when it comes to testing the usability of CAT tools. Such testing is crucial, as usability determines the impact of translation tools on the translator's cognitive performance.

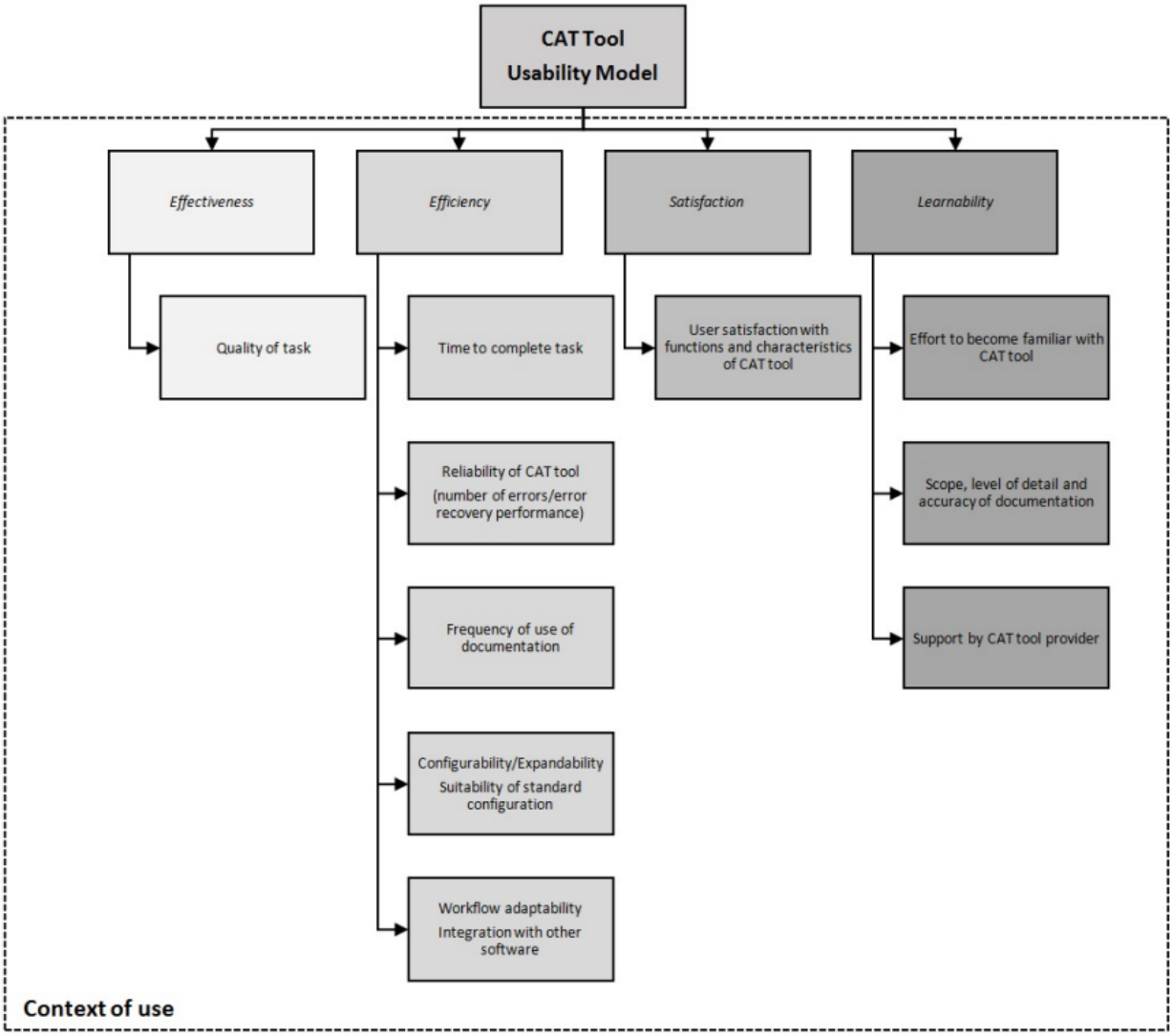


Figure 6: CAT tool usability model (Krüger, 2016, p.131).

It appears important to clarify the differences between the dimensions of effectiveness and efficiency. On the one hand, the dimension of effectiveness can be understood as the way in which the tool provides the necessary elements for the translator to produce a work of high quality. In addition to the CAT tool usability model, Krüger (2016) also introduced a TM system usability model (Figure 7), which extends the CAT tool usability model, particularly for the effectiveness dimension. More precisely, it is divided into three aspects: the representation of translation units (including previous and following units for contextual information), completeness and precision of the elements retrieved from the TM (including concordance search results) and the TB, and the performance of integrated QA checks (and the compatibility with external QA tools). These components are all likely to have an impact on the resulting translation. For example, the lack of contextual information in the visualisation of translation units can cause cohesion issues affecting the quality of the translation. The same goes for the performance of TB and TM retrievals, and QA checks. As a matter of fact, Krüger (2016) suggested that effectiveness is

inextricably connected to translation quality and recommended the use of common frameworks (such as SAE J2450, which is further detailed in 4.2 Approaches to Translation Quality Assessment) to evaluate it. On the other hand, efficiency appears to be more related to productivity, which is primarily measured as the time to complete a translation project. However, Krüger (2016) highlighted that time should not be considered in isolation, as one needs a certain amount of time in order to render a translation of high quality. Among the factors that can affect time and therefore efficiency, Krüger (2016) identified the following: reliability of the tool (errors and time to recover), frequency of use of documentation, configurability (e.g. adapting the layout) and expandability (e.g. adding new functionalities via plug-ins), workflow adaptability (e.g. predefined workflows), speed and ease of use of linguistic assets (e.g. concordance search) and treatment of placeable and localisable elements (e.g. tags).

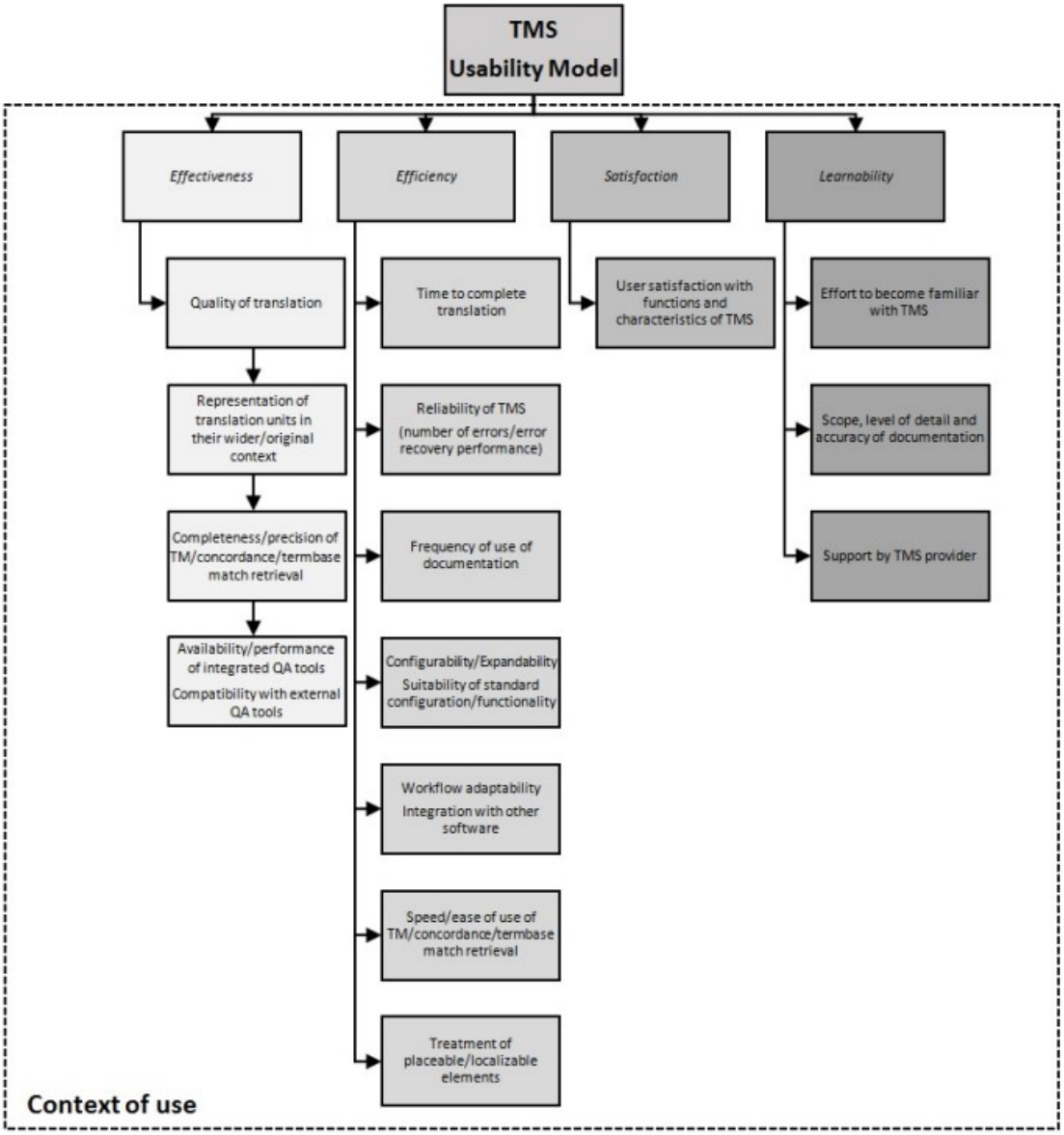


Figure 7: TMS usability model (Krüger, 2016, p.135).

As far as the dimension of satisfaction is concerned, Krüger (2016) suggested using SUS as a template and adapting the items according to the system to be evaluated and the relevant context of use. Finally, learnability has to do with the difficulty or ease with which the user becomes familiar with the tool (which depends on system-internal aspects, such as intuitiveness of the different functionalities), the quality of the documentation provided (scope, level and accuracy) and the support provided by the company which is developing the tool. As Krüger (2016) remarked, there is a natural connection between "learnability" and the "frequency of use of documentation" aspect described in the dimension of "efficiency". A more recent version of the CATTUM was introduced by Krüger (2019). This updated version reflects advancements in usability theory and addresses emerging technological and regulatory demands. The main pillars (effectiveness, efficiency, satisfaction and learnability) remained, but with some adjustments in their definitions, and a fifth element (data security) was introduced (Figure 8).

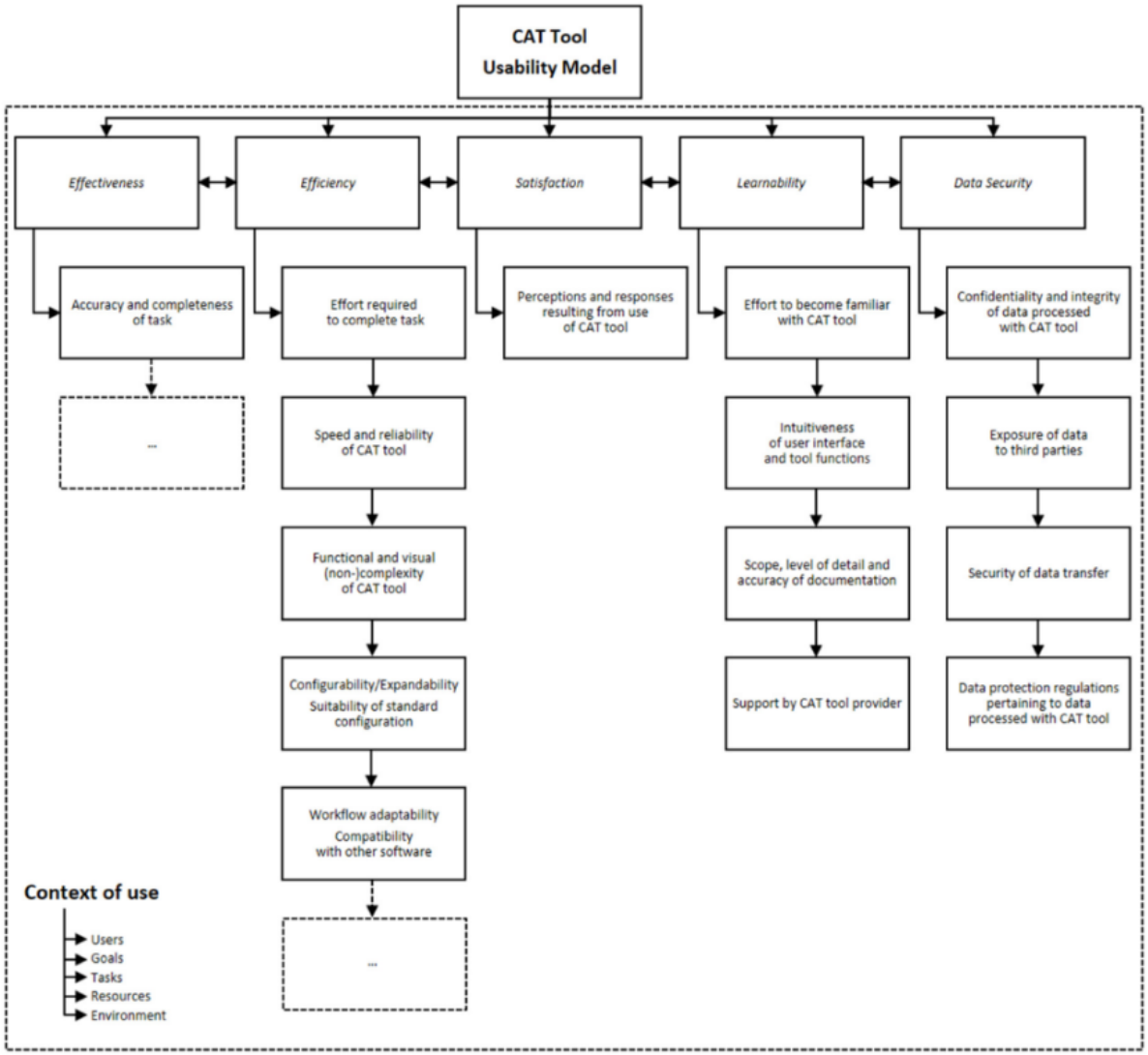


Figure 8: Updated CAT tool usability model (Krüger, 2019).

This expansion addresses contemporary concerns about confidentiality (e.g., exposure of data to third parties), compliance (e.g., adherence to regulations), and technical safeguards (e.g., secure data transfer). Such updates respond to the industry's shift towards cloud-based tools and AI-driven tools, where data privacy and integrity have become critical customer requirements. The revised model also refines its approach to task evaluation. While the first referenced "quality of task" as a sub-element falling under the "effectiveness" dimension, the updated version specifies "accuracy and completeness of task". Similarly, the metric under "satisfaction" was further specified and went to from "user satisfaction with functions and characteristics" to "perceptions and responses resulting from use of CAT tool". Under "efficiency", the metric "time to complete task" shifted to "effort required to complete task", suggesting that task completion time alone is not sufficient to measure usability, and the sub-element "intuitiveness of the user interface" was included as a determining factor of the "learnability" dimension. The updated version also expanded the "context of use" considerations, enumerating factors such as "users, goals, tasks, resources, and environment". The interdependency between all five dimensions of this model is also stressed in the updated version via the use of bidirectional arrows. Furthermore, the dashed-line boxes indicate that the "effectiveness" and "efficiency" dimensions have to be further specified based on the tool being studied.

More recently, Briva-Iglesias and O'Brien (2023) argued that MT was evaluated in terms of productivity and quality, without much focus on the overall interaction experience of MT users. Based on this observation, they suggested adopting a UX-centred approach to evaluate users' perceptions of MT before, during, and after the interaction with an MT system. To address this gap, they introduced the concept of Machine Translation User Experience (MTUX), which is defined as "[a] person's perceptions and responses resulting from the use and/or anticipated use of MT" (ibid., p.338). They tested two questionnaires (AttrakDiff and UEQ) as possible metrics to evaluate MTUX and found that AttrakDiff was centred more on the emotional response and pleasure of users (hedonic quality), whereas UEQ focused more on efficiency (pragmatic quality). Given the importance of speed in the industry, the authors recommended using UEQ based on its pragmatic dimension. Although the justification is relevant, such guidance is somewhat counterintuitive, as the authors emphasised the need for evaluating user interactions with MT beyond productivity but still recommended a metric with a strong focus on efficiency. Arguably, a broader UX perspective encompassing not only the MT output itself but also the environment in which the interaction with this output occurs would provide a more holistic view.

Significant similarities can thus be identified across the usability models presented by the ISO 9241 standard, Nielsen, Quesenbery and the CAT tool-oriented model introduced by Krüger. An overview of the usability dimensions considered in each model appears in Table 15.

Model	Effectiveness	Efficiency	Satisfaction	Learnability	Memorability	Errors
ISO 9241	✓	✓	✓	×	×	×
Nielsen	×	✓	✓	✓	✓	✓
Quesenbery	✓	✓	✓ (“engaging”)	✓	×	✓
Krüger	✓	✓	✓	✓	×	×

Table 15: Comparison of dimensions encompassed in different usability models.

It can be seen from Table 15 that these models all have two dimensions in common: efficiency and satisfaction. Effectiveness and learnability are present in all except one, whereas memorability and errors are included in two and one model, respectively. Furthermore, the ISO 9241 standard and Nielsen’s and Krüger’s models all emphasised the importance of the context of use.

As far as questionnaires are concerned, many options are available, and selecting one is not always a straightforward choice. Moreover, one could question the correlation between the scores obtained from the different questionnaires. To address this concern, Schrepp, Kollmorgen and Thomaschewski (2023) conducted a comparative analysis of SUS, UMUX-LITE and UEQ-S. They found that the three methods provided comparable results when it came to ranking various products. Nevertheless, if the aim is to study the strengths and weaknesses of a product, each questionnaire provides different information. The central focus of both SUS and UMUX-LITE is pragmatic quality, whereas UEQ-S attributes equal importance to pragmatic and hedonic aspects. However, the elements related to pragmatic quality were found to correlate across the three questionnaires. Therefore, the selection of a specific questionnaire in this case should be guided by the importance of hedonic quality attributed in the evaluation.

It also appears pertinent to point out certain discrepancies between usability models and UX questionnaires. The dimensions of efficiency and learnability are found directly in several questionnaires (e.g. they constitute items of the UEQ). “Memorability” is strongly related to “learnability”; thus one could hypothesise that this aspect is indirectly encompassed in the “perspicuity” category of UEQ or in the “Cumbersome – Straightforward” item of Attrakdiff. Similarly, the errors dimension could be linked to “Unruly – Manageable” (Attrakdiff) and “Unpredictable – Predictable” (Attrakdiff and UEQ). Effectiveness is less evident but could be related to the degree to which the tool meets the users’ expectations to help them achieve their goals (UEQ). These five attributes (effectiveness, efficiency, learnability, memorability and errors) are therefore closely related to pragmatic quality. The remaining dimension of usability models, satisfaction, is, however, closer to the hedonic quality components, as it has to do with factors revolving around fulfilment and enjoyment. Therefore, usability models are centred around pragmatic quality, while the UX questionnaires cover a broader perspective, including several aspects of satisfaction (hedonic quality). This observation is not surprising, given that UX is a broader field compared to usability.

With several usability models available in the literature, and one specifically targeted at CAT tools, conducting usability assessments of translation tools seems a particularly promising research avenue. Since the purpose of CAT tools should be to meet the needs of their users, conducting usability studies is paramount. However, in 2008, Lagoudaki pointed out that such research has been scarce. Similarly, Rabadán (2008) adopted the perspective of Descriptive Translation Studies and discussed "applied extensions". In her view, the academia regards anything which is "applied" as more related to users, and fails to place enough emphasis on this aspect, which eventually affects users. She argued that several problems encountered by users have to do with cognetics (i.e. the engineering of tools that accommodate the human thought process). Based on this observation, two pillars are recommended: usefulness (i.e. whether tools can solve practical problems) and usability (i.e. whether tools allow users to optimise their work). Arguably, the situation has evolved since then, as some new developments, like IMT, have fostered user-centred studies (e.g. Briva-Iglesias, 2024). Despite this, Zaretskaya (2017, p.65) suggested that more interest is typically attached to the feature coverage than to usability (she clearly stated that "software developers often pay attention to functionality at the cost of usability"). She explained this phenomenon by the fact that "usability is a more abstract concept, and its evaluation is not that straightforward" (ibid.). One of her conclusions is that CAT tools should be evaluated based on time and effort, since they should help translators by reducing the effort while increasing speed and productivity. Although this suggestion appears to be in line with the CAT tool usability model dimension of efficiency, Krüger's model is more complex and encompasses several other dimensions that seem relevant to evaluate. This reinforces the need for usability studies of translation tools. It can be seen from the literature that some CAT tool evaluation studies began to rely on usability and UX metrics. For example, IntelliCAT (Lee et al., 2021), Intellingo (Coppers et al., 2018) and Escriba (Porto Veloso, 2013) were all evaluated based on SUS. A usability study of web-based CAT tools for blind translators (Rodríguez Vázquez, Fitzpatrick and O'Brien, 2018) employed the CSUQ. Other CAT tool usability studies, notably one focusing on Arab translators (Alotaibi, 2020) and another on students (Vargas-Sierra, 2019) relied on SUMI. Similarly, Briva-Iglesias, O'Brien and Cowan (2023) investigated the UX of post-editors, comparing interactive and traditional PE, also using the UEQ.

2.3 Summary and Outcomes of this Section

This section revolved around principles and methodologies to assess translation tools from a user perspective. In translation technology research, a wide range of methods have been employed (e.g. comparative analyses, user feedback collection, scenarios and observation-based studies), and despite some attempts to develop standardised models (e.g. EAGLES), the field mostly continued to utilise different frameworks. While each study has specific needs and requires evaluation methods to be adjusted accordingly, the absence of a standard practice entails evident replicability and comparability problems. On the other hand, in the field of UX, a plethora of models and assessment methodologies (particularly questionnaires) has been introduced. Although these frameworks differ, several primary usability principles are common to most of them (e.g. efficiency, satisfaction, learnability). This implies better compatibility between methods and suggests that several UX pillars are necessary regardless of the evaluation's focus.

Several UX methodologies have been applied to assess translation technology; however, it appears that there is no consensus on which UX framework to apply when working on translation tools. In addition, since they are meant to be applicable to a wide range of products, these UX frameworks are rather generic. Knowing that several studies have revealed that translation or PE environments do not always meet the requirements of linguists, it would appear relevant to define a UX model specifically adapted to assess translation tools. To that end, gaining first an in-depth understanding of the PE process and its intricacies is paramount.

3. From Translation to Post-Editing: Understanding the Post-Editing Process

While usability and UX assessment is essential, having a thorough understanding of the process being assessed is at least equally important. This section thus revolves around the PE process. It aims at defining it and identifying the differences between translation and PE via an analysis of the phases of the PE process, the PE actions, PE guidelines and PE competence models. The typical environments in which PE tasks take place (i.e. CAT tools) are then discussed with a view to discover whether they are well suited to the PE activity.

3.1 The Translation Process

As a starting point, this section endeavours to define the Translation Process, by contextualising it withing Translation Studies, examining its phases, and how it is affected by the use of technology. This section adopts the perspective of human translation. It provides a summary of a vast body of research, which aims to serve as a foundational reference rather than an in-depth exploration, offering a clear contrast to the technology-driven discussions that follow in subsequent sections.

3.1.1 Origins and Definitions

Translation Studies (TS) constitute a broad discipline, since translation can be researched from a wide range of perspectives. According to Holmes (1988), TS consist of two primary branches: Pure (i.e. pure research) and Applied TS (i.e. practical vision of the translation process, typically concerned with translator training and translation tools). Pure TS can in turn be further divided into Theoretical Translation Studies, which are concerned with understanding and anticipating translational phenomena, and Descriptive Translation Studies (DTS), which aim at describing the phenomena of translation and translating based on empirical, interdisciplinary and target-oriented methods (Holmes, 1988; Toury, 2012; Bassnett and Lefevere, 1995). The definitions provided by Holmes (1988) served to give shape to the DTS ramification by subdividing it into three branches:

- Product-oriented DTS, which provide descriptions of existing translations (such descriptions can be centred around individual translations or can adopt a comparative approach)
- Function-oriented DTS, which adopt a more sociological point of view and produce descriptions of the function of translations in their target context

- Process-oriented DTS, which rely on a psychological or cognitive perspective and aim to describe the act of translation itself (in other words, what occurs in the minds of translators while they are translating).

The Translation Process (TP) studied in process-oriented DTS is admittedly a complex object of study, as it revolves around the "cognitive activity of producing a target text in one language, based upon a source text in another language" (Englund Dimitrova, 2010, p.406). Such cognitive activity is often referred to as the "translators' black box" (Holmes, 1988; Englund Dimitrova, 2010), due to the complexity of observing phenomena taking place in the mind of translators. Holmes (1988, p.73) emphasised this aspect by quoting Ivor Armstrong Richards, who wrote the following about translation in an anthology on Chinese philosophical concepts: "We have here indeed what may very probably be the most complex type of event yet produced in the evolution of the cosmos".

Although the TP is not directly observable, several research methods have been adopted in the field of Translation Process Research (TPR) to gather data and therefore describe and analyse it. Such methods may rely on introspective analysis or on data collection. Introspective analysis can take the form of Think-Aloud Protocols (TAPs), during which a subject is asked to describe their thought process while accomplishing a task, or verbal reporting, which consists of a post-task reflection (Englund Dimitrova, 2010). Data collection generally relies on computer programmes to log various kinds of information. Typical methods comprise keystroke logging (i.e. storing typing and deleting actions, pauses, etc.), eye tracking (recording eye movements on a screen, such as fixations and saccades) and video recordings of task performance (for example, recording the screen with a view to analyse the actions performed by a translator) [ibid.]. Less frequently, other types of physiological data can also be collected, using, for instance, electroencephalograms and functional magnetic resonance imaging. One of the main challenges in TPR lies in running ecologically valid experiments, as it is often difficult to reproduce real situations (for instance, TAPs might affect the TP itself, and eye tracking might be invasive depending on the equipment used). For this reason, it is often recommended to triangulate data (i.e. to combine various methods and draw conclusions based on more than one type of data; Alves [2003]).

3.1.2 Phases of the Translation Process

There is a general consensus in TPR that the TP consists of three main phases, although the nomenclature and the boundaries between certain tasks may differ from one model to another. The first TP models were established around the assumption that translation consisted of converting strings of words individually into their target counterparts, before the sentence became the unit for analysis (Holmes, 1988). One of these early representations is Nida's model of the TP, which was split into:

- Analysis
- Transfer

- Restructuring

Sager (1994) proposed to divide the TP into

- Reading comprehension phase (centred on the Source Language [SL])
- Translation phase (bi-directional)
- Revision phase (centred on the Target Language [TL]).

Austermühl (2001) built on the TP phases introduced by Holmes (1988) to provide an updated model, divided into the following phases:

- Reception phase (retrieval of knowledge required to perform the translation and Source Text [ST] analysis)
- Transfer phase (conversion of the ST map into target context)
- Formulation phase (production of the Target Text [TT])

In addition to this classification, Austermühl (2001) emphasised the crucial role of electronic resources at certain stages (for instance, terminological databases can be helpful during reception to gain a thorough understanding of the ST) and mentioned a post-translation phase during which translated texts are stored in electronic archives. Pym (2011) proposed the following parts of the TP:

- Problem recognition
- Alternative solutions generation
- Selection of one solution

Englund Dimitrova (2010) identified three major phases which are generally studied in TPR:

- Initial phase consisting of planning, orientation and reading
- Drafting / Generation phase, during which the TT is produced
- Revision phase in which the TT is checked

These are reminiscent of the phases found in Carl, Dragsted and Jakobsen (2011):

- Initial orientation
- Translation drafting
- Final revision

do Carmo (2017) distinguished similar phases, and argued that the TP consists of the following:

- Orientation, which includes planning and reading
- Drafting, which consists of generating the TT
- Revision, in which the TT is checked (also called self-revision, further detailed in 3.2.1)

While some notable discrepancies can be pointed out across these models (e.g. Nida's, Austermühl's and Pym's models did not include a revision phase), a clear pattern emerges from the comparison of these representations. In all models, the TP appears to include an analysis of the ST, a transfer stage (which can be separated or not from the generation stage during which the TT is produced), and optionally a revision phase. For clarity, and because this author has extensively studied the topic of the

PE process (defined in 3.2), the TP is considered based on phases defined by do Carmo (2017) in the remainder of this work. It should be noted here that translation is far from being a linear activity, as it is rather an iterative process. During the drafting phase in particular, translators are likely to edit their renditions repeatedly until they are satisfied with the result. do Carmo (2017) argued that most of the decisions are intermediary steps which are often not reflected in the final text. It is indeed common to go back to an already translated sentence to amend it based on knowledge gained at a later point in the TP (Sager, 1994). Some parts of the process are also not directly observable, as post-editors also spend time considering whether to implement changes (as shown by Koehn [2009, p.258]: "most of the time is spent on contemplating changes, but very little on executing them"). These considerations make it impossible to study the TP based only on the final translated text (do Carmo, 2017). Furthermore, it appears important to mention that there is no single way of performing the phases of the TP. In practice, differences are observed based on individual behaviour. Carl, Dragsted and Jakobsen (2011) identified various behavioural preferences for each of the three phases. For example, translators might read the entire ST carefully or just skim through it during the orientation phase; they might focus on different spans of context during drafting; and they might revise constantly or perform an end revision.

3.1.3 The Translation Process in CAT Tools

Apart from studying the TP itself, TPR also encompasses the use of tools to support the translation activity. Since CAT tools are frequently used by translators, it appears relevant to examine how technology affects the TP. It can be argued that the use of CAT tools brought three primary transformations to the translation flow:

- it encourages translators to process the text on a segment-by-segment basis (due to the text segmentation),
- it allows for reusing previously approved translations from the TMs, which are automatically used to pre-fill the target side in the case of high matches,
- and it centralises the information (translators can perform a TM lookup via the concordance search, consult the TB and check spelling in the same interface).

Translators are thus faced with several types of segments:

- Pre-translated segments:
    - CMs or EMs (≥ 100%)
    - High fuzzy matches (75-99%)
- No matches (empty target and usually TM matches <75%)

Four editing actions can be implemented when it comes to processing fuzzy matches: deleting, inserting, moving, and replacing (do Carmo, 2017). Since the fuzzy match score indicates the overlap between a given segment and previous translations, it can be assumed that the TM editing effort increases when

this score decreases. The usefulness of TM matches is usually questionable past a certain threshold, and when too many changes are required, it can be preferable to ignore the TM suggestion and retranslate from scratch. To avoid populating the target side with such matches, several CAT tools allow to set a minimum match value. Therefore, translating in a CAT environment constitutes a combination of translation from scratch and editing of TM matches. do Carmo (2017) contended that in such settings, three diverse ways of writing are mixed: "typing TL sentences from scratch, overwriting SL sentences and/or editing TL sentences". Arguably, constantly switching between these modalities can contribute to increasing the cognitive load. According to Mossop (2020), the work of translators who use TMs is closer to that of a reviser/editor compared to that of a composer. Working with fuzzy matches indeed affects the drafting phase, as the production stage is bypassed when translators are already presented a TM match. In this case, they concentrate their efforts on editing this suggestion rather than producing a translation. In that regard, Pym (2011, p.1) posited that the production phase is the most impacted by the use of TMs, as the "translating activity is enhanced in its generative moment, yet potentially retarded in the moment of selection, where the values of intuition and text flow become difficult to recuperate". Given that the TMs are the cornerstone of CAT tools (they are even sometimes even called "TM systems"), it is not surprising that the main view of the interface in which translators are likely to spend most of their time working on translations is, in fact, often called the Translation *Editor*.

3.2 The Post-Editing Process

In addition to CAT tools, MT is also frequently used in translation workflows today, and PE has now become a standard job type in the industry. This transformation profoundly affects the nature of translators' profession, which seems to be shifting towards PE – a change which was already observed in Pym [2013]). To understand the PE process and how it differs from translation, it is paramount to clarify certain terms. In particular, revision and editing entail ensuring that a text is of a satisfactory linguistic quality. Although these tasks share similarities with PE (mostly because the main focus in on reading texts and spotting errors), they must be clearly defined, and they should not be confused with PE.

3.2.1 Definitions

Since in PE, target segments are populated with a pre-translation (MT output), it could be tempting to conceptualise PE as a form of revision. However, these are two distinct tasks, and before delving into the PE process itself, it is essential to clarify several terms.

A. Revision (self, other)

In particular, before focusing on the difference between PE and revision, it should be noted that there are two forms of revision: self-revision and other-revision (Mossop, 2020). The former is an integral part of the TP, whereas the latter refers to revisions performed outside of it, generally as an additional quality inspection step in the workflow. Consequently, another crucial distinction is that other-revision is not performed by the person who delivered the first translation, but by another expert (ISO, 2015). In both cases, since the translation has already been produced, revision consists mainly of reading the target text (Mossop, 2020). Nevertheless, contrary to self-revision, other-revision is a complete process, and it can be divided into phases. do Carmo (2017) argued that a certain degree of orientation can be found in self-revision, albeit to a lesser extent compared to the orientation in other-revision. According to the same author, after this orientation step, other-revision entails two main tasks: revising (exhaustive verification of all segments) and checking (performing remaining corrections). A certain degree of overlap occurs here, as actions performed in other-revision are supposed to have been performed in the self-revision phase already (e.g. use of QA checks).

In the industry and in the literature, other-revision is sometimes named "review" or "proofreading". For clarification purposes, below are the definitions provided by the ISO17100:2015 standard (ISO, 2015):

> *"Revision: bilingual examination of target language content against source language content for its suitability for the agreed purpose*
>
> *Review: monolingual examination of target language content for its suitability for the agreed purpose*
>
> *Proofread: examination of the revised target language content and applying corrections before printing"*

While revision and proofreading both rely on an SL and a TL (as opposed to review, which is monolingual), revision entails an extensive verification of the translation, and proofreading takes the form of a final check to ensure that required corrections were implemented before delivery. As do Carmo (2017) also pointed out, translation contains a revision phase, but revision does not contain a translation phase. If translators find themselves recreating translations during revision, they are not translating but retranslating.

B. Editing

The main difference between revision (self/other) and editing is that revision entails checking a translation, whereas editing is applied to texts which are not translations. In the latter, the concept of an SL is therefore irrelevant. In practice, however, similarities can be found in the work of editors and revisers. As a matter of fact, Mossop (2020) listed the editing functions of word processors which are

relevant for both tasks. These include spelling, grammar and style checkers, find and replace, displaying changes and inserting comments, among others. Nevertheless, apart from checking linguistic correctness in the TL, revisers have the additional responsibility of ensuring that the ST was accurately translated.

C. PE

This definition of the editing task already throws into question the use of the word "editing" in PE, given that editing alone appears to be a monolingual task, which cannot be said of PE. The term "post-editor" appeared after the introduction of the first MT systems, as it was coined in 1950 by Erwin Reifler (Krings, 2001). Reifler provided the following definition of PE: "to select the correct translation from the possibilities found by the computer dictionary and rearrange the word order to suit the target language" (Krings, 2001, p.44). More recently, the ISO 18587:2017 standard (ISO, 2017) defined PE as follows:

> *"Post-edit: and correct machine translation output".*

In a section dedicated to the PE process, the same standard provides a more detailed definition:

> *"Post-editing is performed on MT output for the purpose of checking its accuracy and comprehensibility, improving the text, making the text more readable, and correcting errors"* (ISO, 2017).

The present work adopts a more extensive definition of PE, based on that provided by do Carmo (2017, p.173):

> *"Post-editing is a generic name that describes a set of tasks by which a translator modifies language content that has previously been converted from a Source Language into a Target Language by a Machine Translation system, in order to make it conform to the objectives defined for the Target Text. [The set of tasks required for the modification of the machine-translated content may include translating, editing and revising.]*
>
> *Post-editing may be identified as being composed only of editing, but this is only possible if the purpose of the Target Text can be achieved by performing only the four technical actions (deleting, inserting, replacing and moving) over the machine-translated content, within a defined editing effort threshold."*

D. Translation vs PE

Generally, post-editors do not need to produce translations, as suggestions are already available to them. In that sense, their work appears close to that of revisers (as opposed to that of composers, as highlighted

by Mossop [2020]) since they focus on finding the right solutions rather than generating new ones – the same happens when editing TM matches (Pym, 2011). In other words, post-editors spend more time reading than writing, the drafting phase is therefore the main difference between translation from scratch and PE. When an MT output is available, it may take the function of a "first draft" (do Carmo, 2017). This phenomenon directly impacts the translators' thought process. When no MT output (or TM match) is presented, translators generate mental representations of potential suggestions to construct translations. A translation suggestion already available on the target side thus interferes with this process, and the generation stage is bypassed (do Carmo, 2017). Nonetheless, this observation may not hold true for all segments in a PE job. If post-editors deem that a certain segment would require applying too many changes, they are likely to reject the MT suggestion and retranslate the segment from scratch.

3.2.2 Phases of the PE Process

The PE process is also composed of various phases, which can be established on the basis of the ones defined for the TP model. Therefore, PE also includes the following phases (based on do Carmo, 2017):

- Orientation, during which the post-editor consults the ST and TT
- Drafting, during which the MT output is checked and amended if necessary (or rejected and retranslated from scratch)
- Self-revision, during which the TT is checked

In this model, the phases of the PE process even bear the same names as the ones of the TP. While this establishes a common reference, the use of "drafting" for the PE process might lead to confusion, as drafting is typically understood as writing a first version of a text. This does not appear appropriate here, since the first version is already provided in the case of PE. Instead, this phase could be called, "adjustment". "Correction" also appears as a good description of this phase; however the task of correcting would imply that the MT output systematically contains errors, which is not necessarily the case, especially with NMT. Certain sentences can sometimes be accepted directly, without performing any type of "correction". The term "validation" would then seem more accurate; however it comes with the opposite bias since validating implies that the output can always be trusted, which is not true. Consequently, "adjustment" is a more balanced and neutral option, as this term describes the task of applying changes in order to obtain a result which fits the requirements, which is a more accurate definition of this central phase in PE.

The challenges faced in finding a clear definition of PE and establishing boundaries with other related tasks, such as translation and revision, demonstrate the complexity of this process. This phenomenon is also reflected in the tools in which these tasks take place. According to do Carmo (2017, pp.66-67), "revisers work with virtually the same interfaces as translators, sometimes the only difference being the features they use more often". In Trados Studio, for example, a bilingual file can be opened "for

translation" or "for review" (which is normally used for revision assignments). Two major observations can be made from here. First, the translation view typically corresponds to translation or PE assignments, which entails that there is no differentiation between these types of tasks at the interface level. Second, there are only minor differences between the translation and the revision views. More specifically, when opening a document for revision, the track changes mode is automatically enabled, and the segments are given a different status after being confirmed. This low contrast in the interface between the different tasks emphasises their proximity and calls for further investigation as to whether more differentiated environments would be beneficial. do Carmo (2017, p.208) concluded that "PE should be studied and approached in practice as a form of translation, rather than a form of revision".

The ISO 18587:2017 standard on PE (ISO, 2017), also identified three main tasks that should be performed by post-editors:

- *"reading the MT output and evaluating whether a reformulation of the target language content is necessary*
- *using the source language content as reference in order to understand and, if necessary, correct the target language content*
- *producing target language content either from existing elements in the MT output or providing a new translation".*

It is commonly recognised that there are four main editing actions: deleting, inserting, moving and replacing (do Carmo, 2017). These operations are triggered by errors in the MT output (for example, a word might be moved if its initial position in the output is incorrect). Blain et al. (2011) argued that the four editing actions (which they called "PE actions") do not constitute an exact description of the post-editor's work because of their mechanical nature. Instead, they proposed "logical edits", which are motivated by a sounder linguistic reasoning and provide a more accurate representation of the editing decisions and how they are related to other mechanical edits. Indeed, substituting a noun for another may entail a series of edits if the new noun has a different gender, as it will be necessary to adjust determiners and adjectives accordingly, for instance. In this case, several mechanical edits would be necessary as a propagation of the "main" edit (substitution of a noun), however these do not reflect the intent of the post-editor. Consequently, they proposed a typology of logical PE actions based on linguistic units (for example, noun phrase for lexical changes, verbal phrase for grammatical changes, preposition changes, etc.). This typology was then used as a model of rules in order to automate the analysis of PE actions.

According to do Carmo (2017), the different processes and the phases they include determine the resulting product. Hence, if self-revision is omitted in a PE process, the result is a post-edited text, but if it is included, the result is a translated text. Consequently, in the second case, the product is the same

as it was produced during a TP. Following this logic, a revised text can be defined as a text which was checked by a second translator. It should also be noted here that the input for an other-revision is a translated text, which can be the result of a TP or a complete PE process. do Carmo (2017) also noted that the result of editing actions are not editions, since after editing a translation, the resulting product is still a translation. In addition, the fact that the result of a PE process is a translation and not a post-edited text "not only implies that, during PE, the translator performed a TP, but it also defeats the idea that the translation had already been completed by the MT system" (ibid., p.179).

3.2.3 Studying the Post-Editing Process

It is widely recognised that PE boosts productivity (compared to translation from scratch). The results reported by Guerberof Arenas (2008) even showed that translators are more productive on PE tasks than when editing TM matches. The experience of translators using MT and the quality of the resulting translations are popular research topics, which have attracted the attention of experts over the last few years. One of the factors impacting the experience is the PE effort, as this measure allows to gain insights into the PE process. It can be defined in several ways, as this effort can indeed be temporal, technical or cognitive (Krings, 2001), and different metrics are typically used to measure each aspect. More specifically, the temporal effort corresponds to the PE time; the technical effort comprises operations performed as part of PE (i.e. word insertion, substitution or deletion) and can be measured by logging keystrokes; and the cognitive effort refers to cognitive processing involved in the activity and can be estimated by think-aloud protocols or eye-tracking data (i.e. gaze time, fixation count and duration). Cognitive effort is therefore the only dimension that is not directly observable. It is also worth noting that the three dimensions of the PE effort do not necessarily correlate with each other. A change might require a high technical effort (e.g. changing an entire sequence of words) but a low cognitive effort (especially in case a translation error was evident, and the correction was easy to find and implement).
With the increasing interest in measuring the PE effort, dedicated tools have emerged. The Post-Editing Tool (Aziz, Castilho and Specia, 2012) allows for collecting PE effort metrics (time, timestamped edits, keystrokes statistics and edit distance) and for assessing the MT output quality (via a score entered by the post-editor). Similarly, Postediting Calculeffort (Candel-Mora and Borja-Tormo, 2017) was designed based on recurrent categories in MT error typologies and allows for obtaining estimations of time and effort for a PE task based on an annotation of MT errors. Matecat, an online CAT tool, also provides an editing log which stores and analyses editing actions and provides, notably, the PE time and technical effort for each segment. Such tools allow for investigating, among other phenomena, whether specific MT errors or language directions influence the PE effort. In such an experiment, Zaretskaya et al. (2016a) found that MT errors resulted in different effort intensity based on the language pair.

Krings (2001) also suggested a distinction between the absolute PE effort (i.e. derivation from a "zero" value, in which no editing would be required) and the relative PE effort (i.e. relative effort compared to the translation effort; in practice, it takes the form of a ratio between the absolute PE effort and the translation effort). Therefore, a relative effort of zero means that the MT output did not require editing (this value cannot be obtained in practice, since a minimum effort is always required for reading). A relative effort value of one implies that PE and translation required equivalent effort. A value below one indicates that PE required less effort than translation from scratch (i.e. PE entails a productivity increase), whereas a value above one indicates that translation from scratch is more efficient. It should be mentioned that several factors might affect the effort. One of these is the exposure to MT output and the habit of working with a specific MT engine. Arnold et al. (1994, p.33) pointed out that "familiarity with the pattern of errors produced by a particular MT system is […] an important factor in reducing PE time".

The types of PE actions also have an impact on the PE effort. Popović et al. (2014) investigated the relation between PE operations and effort. In this study, the cognitive effort was directly related to the quality of an SMT output, as human annotators had to assign a quality level to each sentence, and each quality level was mapped to an effort measurement (for example, when the output was almost acceptable, it was deemed easy to post-edit). This study identified the most demanding edit operations: reordering and lexical edits were both the most cognitively demanding, and lexical edits also entailed the higher PE time. Although the measurements of cognitive effort relied exclusively on the annotators' perception (which might be biased compared to other methods like eye-tracking, for example), the results provided patterns which served for establishing correlations between editing types and effort.

Comparing PE and translation from scratch is an angle frequently adopted in the literature. As a matter of fact, Jia, Carl and Wang (2019) compared PE of NMT outputs and translation from scratch in translation students. PE was found to entail a lower cognitive load (reflected by shorter and fewer pauses) independently of the text type and to be significantly faster in the case of specialised texts. The authors relied on the relevance theory to explain the lower cognitive effort in PE. According to Alves and Gonçalve (2013, p.111), "the principle of relevance establishes an optimal balance between the estimated effort to be spent and the effects to be obtained, always looking for a positive relation". When the MT output quality is satisfactory, drafting is eliminated, and translators are less likely to contemplate alternative translations. The authors concluded that PE is less demanding from a cognitive perspective compared to translation from scratch.

Other experts have expressed different views. This is notably the case of do Carmo (2017), who argued that MT outputs act as another type of reference and contribute to increasing the number of elements that should be checked by translators, which results in investing more time and effort in research. This

author however indicated that a compensation phenomenon occurred, as the "increased cognitive reading load has to be balanced with the decreased technical load" (ibid., p.182).
Apart from approaching the PE experience from measured values, the perception of post-editors should not be neglected. In that regard, Moorkens et al. (2015) investigated the relation between perceived and actual PE effort via a comparison of predicted effort (indicated by raters) and actual effort (cognitive, temporal and technical). The findings revealed only a moderate correlation. As far as perception is concerned, it is also worth mentioning that the type of errors found in MT outputs would rarely correspond to errors typically made by translators, and correcting such errors repeatedly can be a tedious task. In that sense, PE can be perceived as less pleasant compared to translation from scratch, as it involves continuously fixing "unintelligent errors" (O'Brien, 2020). As far as quality is concerned, Jia, Carl and Wang (2019) found that the products of translation from scratch and PE had equivalent quality levels, considering both fluency and accuracy.

It also appears relevant to mention here that when the effort exceeds a certain limit, the PE task becomes, in fact, more of a translation task. The score of TM fuzzy matches can be directly connected to the editing rate, as there exists a direct inverse relationship between these values. For instance, it can be assumed that a fuzzy match with a score of 80% will entail an editing rate of 20% (i.e. editing the part of the segment which does not correspond to the TM entry). do Carmo (2017) summarised this phenomenon across the different bands of TM scores in Table 16 below.

	Fuzzy score	Editing rate	Resource	Task
Full match	100%	0%	TM	Checking
Fuzzy match	99-95%	1-5%	TM	Editing
Fuzzy match	94-85%	6-15%	TM	Editing
Fuzzy match	84-75%	16-25%	TM	Editing
No match	74-0%	26-100%	MT	Editing

Table 16: Fuzzy match bands and editing rates for TMs and MT (do Carmo, 2017, p.211).

While the fuzzy match score allows for establishing the corresponding editing rate for TM matches, estimating this value in the case of PE is not straightforward. The TM threshold for a match to be considered usable is commonly set to 75% in the industry. In that sense, it is implied that an editing rate above 26% requires translation instead of editing. Consequently, below this 75% value, lower-scoring matches are typically left out and MT is applied instead. This scenario is problematic because an editing rate of 100% would entail that the MT output was not useful at all, and that the post-editor had to retranslate from scratch. Following this rationale (i.e. 0-25% = editing; 26-100% = translation), do Carmo (2017, p.210) introduced the MT editing threshold, which defines the line "above which the

complexity of editing involved in a PE project exceeds the level that is considered reasonable for an editing task", based on the values presented in Table 17 below.

	Fuzzy score	Editing rate	Resource	Task
MT hypotheses		0-25%	MT	Editing
MT hypotheses		26-100%	MT	Translating

Table 17: The MT editing threshold (do Carmo, 2017, p.211).

Acknowledging this concept would imply dramatic changes in the translation industry. This threshold can indeed be applied both to the segment level and to the document level. Therefore, projects in which over 25% of the content had to be edited should be considered translation jobs, rather than PE jobs – which would have important economic repercussions (ibid.). Assigning a value to this threshold remains, however, an arduous task, as the effort is sometimes not accurately reflected by the editing rate alone. A seemingly low edit rate may indeed have entailed a high research effort (e.g. checking terminology or looking for a definition), and a seemingly high edit rate may have been fast to implement. According to do Carmo (2017), the highest editing effort is commonly found in the intermediate values of the editing threshold (and not in the upper and lower bounds).

3.2.4 Post-Editing Guidelines

The study of PE guidelines also sheds some light on the tasks which are expected to be performed in a PE assignment. Guidelines are indeed essential to define the objectives and to provide a set of best practises to post-editors. However, drafting PE guidelines is an arduous task, and there is no international standard or recommendation regarding the content that such guidelines should cover (Gene and Guerrero, 2021). Moreover, requirements may vary significantly from one project to another, based on factors such as quality requirements, targeted readership or the duration of use of the final text (Nitzke and Hansen-Schirra, 2021). For this reason, although generic guidelines have been introduced in the industry, it should be borne in mind that they are always subject to adaptations to fit a particular PE scenario (Massardo et al., 2016).

Generally, the industry makes a distinction between two types of PE based on the expected quality: light and full PE. Despite some criticism, this differentiation is well established (Nitzke and Hansen-Schirra, 2021). The difference between these two types of PE lies in the degree of quality expected in the final text. While light PE aims at ensuring adequacy and correcting errors which would otherwise affect comprehension ("good enough" quality), the objective of full PE is to produce a translation of human quality, or which can be deemed equivalent to human quality (sometimes referred to as "publishable" quality). The distinction is therefore not based on the PE effort, but on the final quality (Gene and

Guerrero, 2021). More specifically, Massardo et al. (2016), provided the TAUS guidelines below to perform PE depending on the quality requirements:

"Guidelines for achieving "good enough" quality:

- *Aim for semantically correct translation.*
- *Ensure that no information has been accidentally added or omitted.*
- *Edit any offensive, inappropriate or culturally unacceptable content.*
- *Use as much of the raw MT output as possible.*
- *Basic rules regarding spelling apply.*
- *No need to implement corrections that are of a stylistic nature only.*
- *No need to restructure sentences solely to improve the natural flow of the text.*

Guidelines for achieving quality similar or equal to human translation:

- *Aim for grammatically, syntactically and semantically correct translation.*
- *Ensure that key terminology is correctly translated and that untranslated terms belong to the client's list of "Do Not Translate" terms.*
- *Ensure that no information has been accidentally added or omitted.*
- *Edit any offensive, inappropriate or culturally unacceptable content.*
- *Use as much of the raw MT output as possible.*
- *Basic rules regarding spelling, punctuation and hyphenation apply.*
- *Ensure that formatting is correct."*

While these guidelines have a reference status (which can be attributed primarily to the excellent reputation of TAUS in the industry), other guidelines have been proposed, and although they share some similarities with TAUS, certain instructions can seem contradictory. Hu and Cadwell (2016) indeed compared five sets of PE guidelines and identified overlaps and discrepancies. Namely, in the case of light PE, content correctness plays a significant role, whereas grammar, syntax and style are not deemed essential. However, only certain guidelines specify that terminology should be consistent. Apart from being difficult to follow, in particular because of the frustration caused by leaving certain errors unedited (Hu and Cadwell, 2016), the usefulness of light PE is sometimes questioned. As far as full PE is concerned, most guidelines share the same vision about accuracy, terminology and grammar, but express different views on style.

In addition, one instruction seems to be shared among most guidelines: the recommendation to use as much of the MT output as possible (Hu and Cadwell, 2016; Massardo et al., 2016; Nitzke and Hansen-Schirra, 2021), in other words, to avoid implementing preferential changes.

In the industry, the PE standard is ISO 18587:2017 (ISO, 2017), which defines the requirements around full PE and post-editors' competences. More specifically, according to this standard, the objectives of the PE process lie in ensuring:

- *"comprehensibility of the post-edited MT output*
- *correspondence of the source language content and target language content*
- *compliance with post-editing requirements and specifications"*

3.2.5 The Post-Editing Process in CAT Tools

Although CAT tools and MT are often discussed and investigated separately, they are inextricably connected, as MT engines can generate a pre-translation used to populate target segments in CAT tools. Moreover, all components of CAT tools are directly applicable to this pre-translation, as linguists can check the MT against the TM, ensure that preferred terminology has been employed by referring to the TB, and use QA checks to prevent errors. It is also common to establish a threshold above which TM fuzzy matches are inserted directly instead of MT (usually around 75-80%). This use of MT in CAT tools is sometimes described as hybrid PE. O'Brien (2020, p.384) argues that the distinction between TM and MT is becoming blurred:

> *"The traditional differentiation between TM and MT technology is being eroded, as is the differentiation between 'translating' with the help of a TM match and 'post-editing' MT output. Cognitively, the tasks are getting closer. For instance, a translator can now be offered a TM match, part of which is then enhanced using MT—this is sub-segment machine translation. A translator may no longer know what part of the text comes from TM and what comes from MT, or even which MT engine generated the proposal".*

Performing PE in a CAT environment thus entails an additional layer of complexity, as there are two primary types of translation suggestions: TM matches and MT output. In certain cases, these two types of suggestions can be combined inside one segment (in so-called repaired matches).
The approach of linguists is presumably different based on the origin of a suggestion. There are indeed different degrees of trustworthiness between a TM match (which is likely to be perceived as a reliable source) and an MT output (which may contain adequacy errors). Since TM and MT segments can come one after the other in the same project, the reliability of translation suggestions is constantly reassessed. In this context, provenance indicators are likely to help linguists in their work. Teixeira (2011) investigated how knowledge of provenance affected speed and quality. The findings showed little evidence of high changes based on the environment (provenance available or not), as quality and speed

did not differ significantly. However, a closer analysis of individual segments revealed that the time spent editing EMs was higher when no origin information was available, which suggests that it is relevant to include provenance data.

Although editing TM matches and post-editing MT outputs are activities carried out with different strategies in mind, they both entail implementing the same type of actions – typically, deleting, inserting, moving and replacing (do Carmo, 2017). Furthermore, Bundgaard, Paulsen Christensen and Schjoldager (2016) identified three types of segment-level decisions which are applied iteratively for each segment: accept, revise and reject. In a TM+MT project, translators repeatedly make this decision for each segment, regardless of the suggestion's origin. The same authors proposed to classify the segment-level decisions as follows:

- Accept: confirm a pre-translated segment without editing it
- Revise: edit a pre-translation before confirming the segment
    - Match-internal revision: revise a pre-translation based only on the content of the current segment itself (autosuggest considered internal)
    - Match-external revision: revise a pre-translation by using resources outside the current segment – more specifically by performing the following actions:
        - "(1) Use of the Copy Source to Target function (where the proposed match is replaced by the source text segment)
        - (2) Use of the Concordance Search
        - (3) Insertion of an MT match instead of a proposed TM match
        - (4) Moving back to a previous segment
        - (5) Use of the reference text (the fully formatted source text)"
- Reject: delete the pre-translation and start translating from scratch

It should be noted that in the context of this study, "match" (in match-internal and match-external revision) refers both to TM and MT suggestions. Although this wording is sometimes used in the industry, talking about an "MT match" is somehow inaccurate, as an automatic translation is generated rather than matched against previous translations (as opposed to suggestions coming from the TM, which are partially or entirely matching previous translations). The operations performed during revision (both internal and external) include, among others, the ones listed hereafter (with the possibility to perform some of these using the keyboard only):

- Moving the cursor
- Moving text
- Writing (typing)
- Accepting autosuggest
- Selecting

- Substituting
- Deleting
- Inserting (letters, symbols, suggestions)
- Copying source to target
- Tag placing
- Formatting
- etc.

It can be remarked here that the authors limited external revision to the use of resources which are external to the segment being revised, but internal to the CAT tool. However, it can take the form of CAT-external revision when translators consult resources outside the tool (such as online concordancers or glossaries). Such an action can even include going back to the project specifications to check instructions. In summary, when faced with challenges, translators can search for information or solutions in:

- Their (biological) memory
- Inside the tool (TB, TM, concordance)
- Outside the tool (project guidelines, web search)

3.3 Theoretical Challenges in Post-Editing

Challenges can be encountered at different levels, from the role and competence of post-editors, their general working situations and the translation or PE process itself – aspects that are sometimes intertwined. It is essential to understand the difficulties faced by translators and post-editors at each level to understand their experience with PE. This section therefore discusses theoretical considerations on the role of the post-editor in the modern translation workflow.

3.3.1 The Post-Editor Profile & Competence

Having clarified the PE process and PE guidelines, some questions arise regarding the profile of post-editors: who should post-edit? What exactly is the role of the post-editor? What competences are required for this role? Apart from the implications on training, attempting to answer these questions will provide further insights on PE itself by describing what the expected skills are. As MT is commonly being integrated into translation workflows, both the industry and academia are working towards a definition of the role and competences of the post-editor (Slator, 2022). In a report published in 2022, Slator (2022, p.21) gave the following definition of the role of the post-editor:

> *"A human post-editor addresses the gap that exists between the quality that MT is able to produce and the quality needed by the enterprise or end-user: identifying and rectifying the shortcomings of MT output against the linguistic requirements of the content with respect to its business purpose".*

In terms of skillsets, several competence models have been proposed for translation, but given the lack of an equivalent for PE, defining the PE competence appears as a more complex task than can be assumed. As a result, there is no consensus on the PE competence model (Ginovart Cid, 2021).

The descriptions of translation and PE in current ISO standards are striking examples of the lack of definition of the PE task. The ISO17100:2015 standard (ISO, 2015) on requirements for translation services indeed identifies six main competences for translators:

- Translation competence
- Linguistic and textual competence in the source language and the target language
- Competence in research, information, acquisition and processing
- Cultural competence
- Technological competence
- Domain competence

The ISO 18587:2017 standard (ISO, 2017) on PE requirements lists the exact same competences and does not mention PE-specific phenomena such as the identification of errors in the MT output as part of the PE competence model. While this is understandable, given that the PE standard was published at the beginning of NMT, it is commonly recognised that a more up-to-date description is now necessary (Ginovart Cid, 2020; GALA[34], 2022).

According to Ginovart Cid (2021), the three main models of translation competence comprise the TransComp (Göpferich, 2009), PACTE (Process of Acquisition of Translation Competence and Evaluation; PACTE Group [2005]) and EMT (European Master's in Translation; EMT Board [2017]) models, represented in Figure 9. While each model has its specificities, common competences can be identified. Among others, each of these models encompasses some form of linguistic, domain and technological competence. O'Brien (2002) also identified a series of PE skills, including knowledge of MT, terminology management, controlled language, programming and text linguistic skills. Similarly, Pym (2013) suggested necessary skills for working with translation technology based on the decision-making processes occurring as a result of using MT and TMs. These were divided into three main components: "learning to learn", "learning to trust and mistrust data", and "learning to revise with enhanced attention to detail".

[34] GALA: Globalisation and Localisation Association

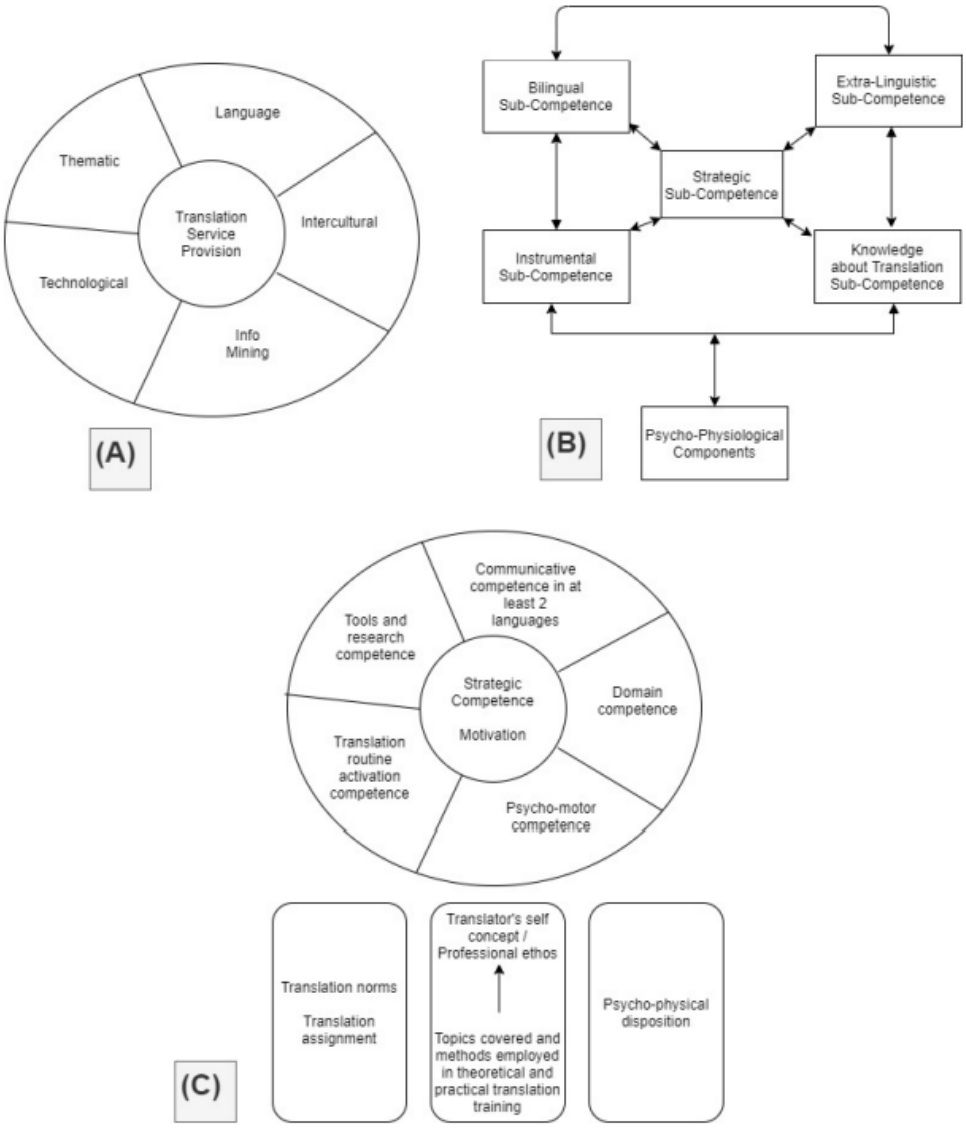


Figure 9: Translation competence models described in Ginovart Cid (2020, p.45). A: EMT; B: PACTE; C: TransComp.

Nitzke and Hansen-Schirra (2021) introduced a house-shaped PE competence model, (Figure 10) consisting of three pillars (error handling, MT engineering and consulting), standing on a foundation block (basic translation competences) and supporting a roof (soft skills). This representation makes it clear that the translation competence (which includes bilingual, extralinguistic and research competences) is paramount in PE. According to Nitzke and Hansen-Schirra (2021), post-editors indeed need the same basic skillset as translators. Then, each pillar comes with further sub-competences. Error handling, for instance, can be divided into error spotting, classification and correction; MT engineering includes training and evaluation of MT systems; and consulting has to do with risk assessment related to the use of MT. The role of each pillar is more or less important depending on the specialisation chosen.

On the one hand, for instance, focusing on practical PE requires solid error-handling skills, which entails a great ability to identify errors and requires that post-editors are aware of the potential weaknesses of specific MT engines. On the other hand, specialising in MT engineering leads to the development of different skills, such as the ability to prepare training data. On top of these pillars, certain soft skills are essential in PE, in particular the ability to concentrate, to manage stress, to think critically and to adhere to a PE briefing, among others. This also confirms the observations made by do Carmo (2017, p.213) according to which the nature of PE (i.e. a "recursive and cumulative process") requires post-editors to be "not only experienced and focused readers, but also efficient technical performers". Nitzke and Hansen-Schirra (2021) also highlighted the importance of being technology-minded, since PE is commonly performed in CAT tools. A key takeaway from this model is that, although the translation competence is the basis of PE, it is not enough to master the PE skillset. In that regard, O'Brien (2002) clearly stated that it cannot be "assume[d] that a qualified translator will be a successful post-editor".

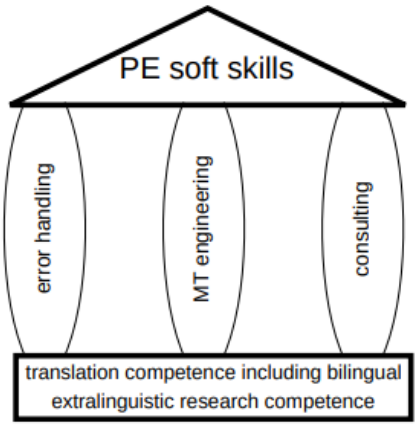


Figure 10: The post-editing competence model presented by Nitzke and Hansen-Schirra (2021, p.70).

Ginovart Cid (2021) identified three core PE skills: error spotting, decision making and application of guidelines. It can be argued that these skills are somewhat encompassed in Nitzke and Hansen-Schirra's model. Error spotting and decision making are both sub-competences of the error handling pillar, and the application of guidelines fits in PE soft skills. However, Ginovart Cid's proposal is clearly more oriented towards practical PE, whereas Nitzke and Hansen-Schirra presented a more versatile model. Such discrepancies are frequently encountered and sometimes lead to confusion. As a matter of fact, do Carmo (2017) analysed a definition of PE presented by TAUS in which providing feedback to the MT team was an integral part of the PE strategy. According to this definition, PE involved not only the practical task but also entailed supporting MT developers with a view to improving MT systems. Therefore, it appears essential to recognise the different profiles of the post-editor depending on the purpose of the role. In that sense, this diversity seems accurately represented in Nitzke and Hansen-

Schirra's model. Recently, Briva-Iglesias and O'Brien (2022) described the emerging profile of the Language Engineer, a role combining deep linguistic knowledge with a solid understanding of language technologies, which also emphasises the need for satellite positions with a versatile expertise in addition to "practical" post-editors.

Besides the PE competence model, different degrees of expertise can also be identified. According to the MTPE protocol, for example, the competences of junior post-editors are centred on the production of target texts, whereas expert post-editors should also be able to provide feedback to improve MT systems (Slator, 2022).

Finally, it should also be noted that since MT is a relatively recent technology, the objectives of PE training are still under definition. It is commonly acknowledged that dedicated PE training is required in translation programmes. In 2002, O'Brien (2002) already emphasised the need to include PE in translator training. Today, it is still rare to find dedicated PE modules in training programmes. However, this transformation is being implemented gradually. The MTPE Training Protocol (GALA, 2022) is a notable example of such an initiative. In the EMT Competence Framework (EMT Board, 2017), PE also appears to be a sub-competence under the translation competence.

3.3.2 Augmented Translation

The gradual automation occurring in translation inevitably raises questions regarding the place and role of the modern translator. Today, the highest level of assistance in the translator's workstation, as described by Melby (1982) represents the basis of a typical PE environment. With the integration of MT in CAT, the diverse sources of technological assistance have become so intertwined that it might be difficult to separate them (Nunes Vieira, 2019). Despite this, several attempts were made to represent the distinct levels of technological assistance in a spectrum. According to Hutchins and Somers (1992), CAT (which, in their representation, includes both HAMT and MAHT) can be found at the centre of the spectrum, which is delimited by two extremes: mechanisation (i.e. Fully Automatic High Quality Translation [FAHQT]) and human involvement (i.e. traditional human translation), as illustrated in Figure 11.

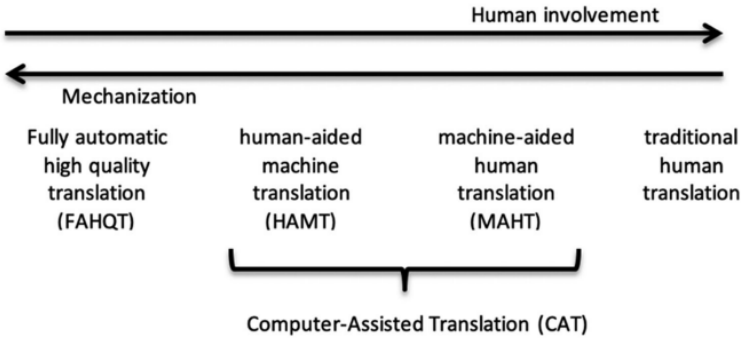


Figure 11: The spectrum of translation methods according to Hutchins and Somers (1992). Source: Paulsen Christensen et al. (2022, p.23).

However, this continuum considers MT and CAT as two separate technologies and fails to represent the integration of MT into CAT (Nunes Vieira, 2019; Paulsen Christensen et al., 2022). Nunes Vieira (2019) presented an alternative spectrum of agency in the PE process, which takes the form of another horizontal axis with two extremes, but accounts for the integration of MT in CAT (see Figure 12). Rather than full automation versus traditional translation, this spectrum is based on the role of humans, and therefore places MT-centred and human-centred paradigms at both ends of the horizontal axis. When the machine is at the centre, the PE process is fully automated, notably via the use of APE models. At the other end of the spectrum, when humans are at the centre, they exercise control over the technology they are using, therefore IMT falls under this category (as in IMT settings, the translation is conditioned by the translator's input). In this model, traditional PE (referred to as "static" PE), whereby a human expert corrects MT output (potentially in a CAT environment), appears in the middle of the spectrum. While in this representation, IMT is considered a human-centred technology, this vision can be questioned. O'Brien (2020, p.385) indeed asked whether IMT could instead contribute to reducing the translator "to an operator that presses 'go' or 'no go' buttons".

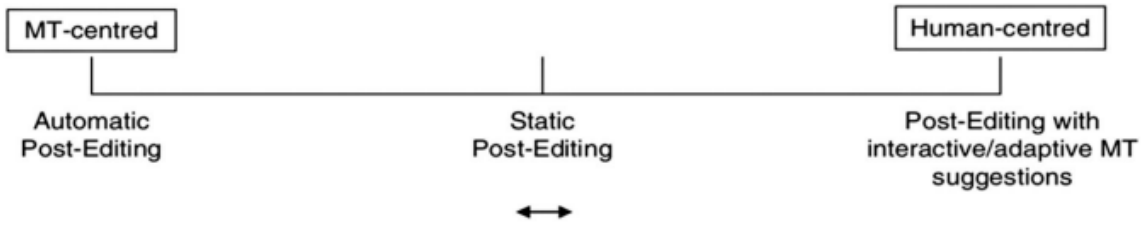


Figure 12: The spectrum of agency in the PE process according to Nunes Vieira (2019). Source: Paulsen Christensen et al. (2022, p.24).

Another classification of the levels of technological assistance was presented by Paulsen Christensen et al. (2022). Taking inspiration from the taxonomy of driving automation levels designed by the Society

of Automotive Engineers (SAE), Paulsen Christensen et al. drew a parallel between driving automation and translation automation. Technological advances point to a future in which cars will no longer need a human driver. Based on this observation, the SAE presented a taxonomy of driving automation, divided into six levels ranging from no automation to full automation. Paulsen Christensen et al. (2022) applied this reasoning to the field of translation and designed a taxonomy consisting of six levels of translation automation, as depicted in Table 18. This classification considers in particular whether the translator or the system performs certain actions, such as controlling the ST and producing the TT as well as detecting and correcting errors.

Level	**Name**	**Dynamic Translation Task (DTT)**		**DTT Fallback**	**Operational Design Domain (ODD)**
		Control of source text analysis and target text production	Error and inadequacy detection and response		
Translator performs all or part of the DTT					
0	**No Translation Automation (TA)**	Translator	Translator	Translator	n/a
1	**Translator Assistance**	Translator and System	Translator	Translator	Limited
2	**Partial TA**	System	Translator	Translator	Limited
"Automated Translation System" (ATS "system") performs the entire DTT					
3	**Conditional TA**	System	System	Fallback-ready user (becomes the translator during fallback)	Limited
4	**High TA**	System	System	System	Limited
5	**Full TA**	System	System	System	Unlimited

Table 18: The levels of translation automation according to Paulsen Christensen et al. (2022, p.28).

In practice, completing a PE task in a CAT tool is likely to imply some back and forth between the different levels of automation, as the post-editor can make use of features corresponding to different degrees of assistance, ranging from looking for occurrences of an expression via the concordance search (which could correspond to a moderate level of assistance) to rejecting an MT suggestion (in which case, the degree of assistance would be relatively low). However, since the post-editor is expected to make PE decisions, PE in a CAT tool would most likely not correspond to the highest levels of automation.

What emerges from these models (i.e., Hutchins and Somers, 1992; Nunes Vieira, 2019; Paulsen Christensen et al., 2022) is that the degree of human involvement can be represented in a continuum expanding between two extremes: total reliance on the human and total reliance on the machine. The

reality of post-editors' work today seems to be located somewhere in the middle of this spectrum, although it might lean towards one end or the other, depending on the type of technology itself. It appears that total reliance on the human agent is unlikely in a context in which TMs and MT are heavily used. Total reliance on the machine is also not desirable today, since MT outputs require revision by human experts, who are responsible for ensuring that target texts comply with certain quality standards.

It is nonetheless evident that technology occupies an increasingly prominent place in the work of translators. Translation therefore became an HCI process (O'Brien, 2012). This symbiosis between humans and machines in translation frequently gives rise to debates.

The most optimistic experts argue that increased automation leads to a form of augmented translation, whereby the capacities of humans are amplified, which allows for being more productive and focusing on more interesting aspects. The idea of an "augmented" intellectual worker was previously formulated in the 1960s by Engelbart (1962) as a conceptual proposal. This is also reminiscent of a similar suggestion by Licklider (1960), who proposed to place routinisable work in the hands of machines while humans focus on setting the goals and performing evaluations. Similarly, Lipmann (1971) considered technology as a way to extend the capabilities of humans. Kay (1980) described what he named the Translator's Amanuensis. According to this concept, machines would gradually take over certain tasks in the translation process. The ideas formulated by Kay proved to be very influential, as they still constitute a good description of modern translation technology. The augmented translation theory continues to attract interest today, especially since CSA Research[35] published an article (DePalma, 2017) showing that this concept contributed to empowering linguists by placing them in the centre of the process, as illustrated in Figure 13.

[35] CSA: Common Sense Advisory. https://csa-research.com/ (consulted: 30/04/2025)

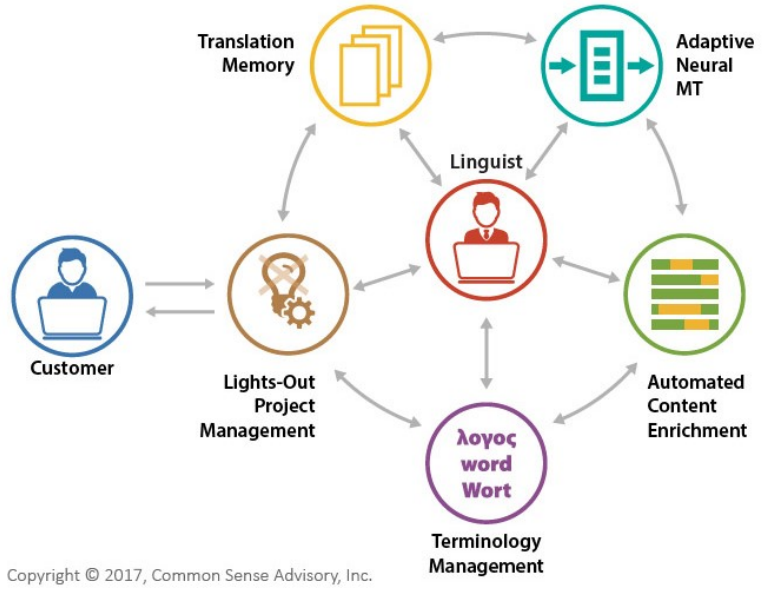


Figure 13: Augmented Translation Puts Linguists in the Centre. Source: DePalma (2017).

The rapid advances in technology entailed a digital transformation in the localisation industry, where LSPs adapt to the changing needs in customer demand by relying on a strategic use of AI, and progressively become Global Content Service Providers. CSA Research (2024) calls this the "post-localisation era". In this paradigm, machine processes constitute a key asset, and different technologies are often intertwined. The place of the translator is inevitably affected, and in the same article, CSA Research introduced a new model, named "human at the core" (see Figure 14).

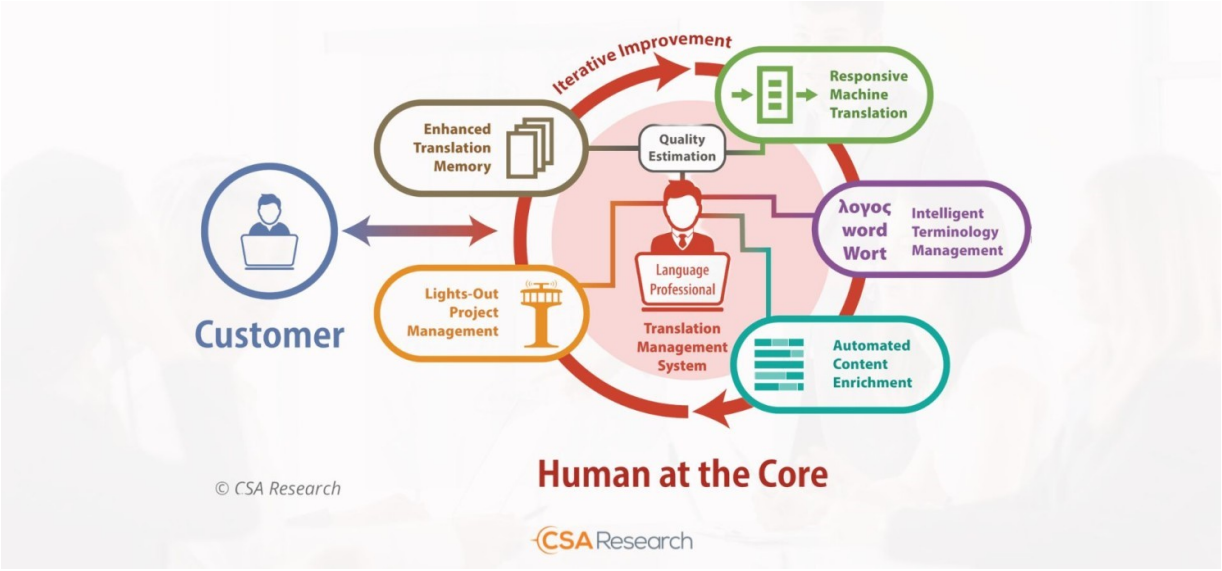


Figure 14: Human at the Core in the Post-Localisation Era. Source: CSA Research (2024).

While this new configuration highly resembles the augmented translation model, several differences should be noted. The "linguist" became a "language professional", which also reflects the tendency in the industry to replace "translator" with more generic terms such as "language expert". Several components surrounding the language professional are now also augmented (TMs becoming "enhanced" and terminology management becoming "intelligent"). A QE element is also incorporated, and all systems are part of an iterative improvement cycle. The role of the language professional is embedded in a complex, auto-improving ecosystem. This may contribute to further augmenting translation; however, language professionals may be faced with challenges, having to evolve in more intricate environments, which highlights the importance of assessing their experience. CSA Research nevertheless adopts a positive perspective on that matter, noting that "unlike 'human in the loop', where humans are relegated to a passive role of correcting output, 'human at the core' places human decision-makers at the forefront, ensuring that technology serves as a powerful augmenting force rather than a directive one.". Similarly, do Carmo, Trigo and Maia (2016) argued that the increasing integration of machine learning in translation technology would eventually lead to a transition from CAT to KAT (Knowledge-Assisted Translation). This focus on human expertise is also the pillar of Intelligence Augmentation (IA; Sadiku et al., 2021), which consists of enhancing human intelligence by letting machines perform tasks which can benefit from automation (such as data analysis and statistical reasoning), while letting human experts make use of the skills they are best at (such as critical thinking and creative tasks). With IA, the aim is not to replace humans (as is often feared in many areas, including translation) but rather to establish a symbiotic relationship with technology and empower them by extending their capabilities, as shown in Figure 15.

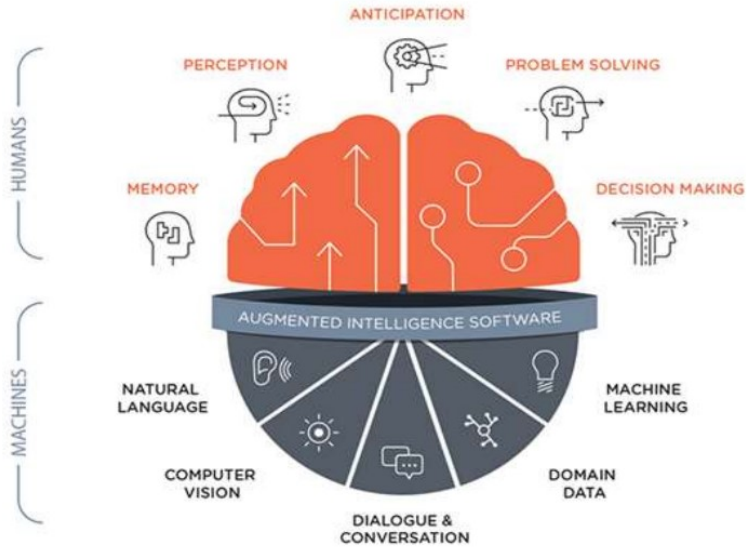


Figure 15: The combination of human and machine intelligences in Intelligence Augmentation. Source: Sadiku et al. (2021, p.775).

Conversely, more pessimistic views claim that automation will eventually reach a level in which humans are completely replaced by machines. van der Meer (2021) even mentions Singularity, a stage in which human translators will not be needed anymore in the translation workflow – and foresees this to happen in 2030. This vision, largely supported over the last few years by the claims of MT reaching human parity, created feelings of uncertainty among translators, who are concerned about the future of their profession. Some post-editors would also argue that working with MT is not as pleasant as translating from scratch, as it entails repetitive actions (including correcting recurrent errors). This perception naturally entails frustrations, since "humans naturally want to feel that they are doing useful, interesting work and that they are using the machine instead of it using them" (Melby, 1982, p.215).

However, several detractors of the Singularity paradigm argue that it is unlikely to happen in the near future (Briva-Iglesias and O'Brien, 2022) and put forward solid arguments against this hypothesis. Melby and Kurz (2021) indeed responded to the publication of van der Meer with counterarguments, defending that translators are "here to stay". Their central argument was the uniqueness of human intelligence, which was illustrated by the Hans-Christian Boos pyramid (see Figure 16). While machine intelligence can indeed perform complex tasks, such as translation, currently it does not have any ability to interpret meaning, nor can it make informed decisions based on principles which cannot be inferred beyond data analysis. Translation is inherently embedded in a sociocultural context and has a *skopos* (i.e. a communicative purpose; Reiss and Vermeer, 2014). These phenomena inevitably orientate translation choices, and despite the progress made in considering context in machine learning models, machines often fail to detect these crucial elements.

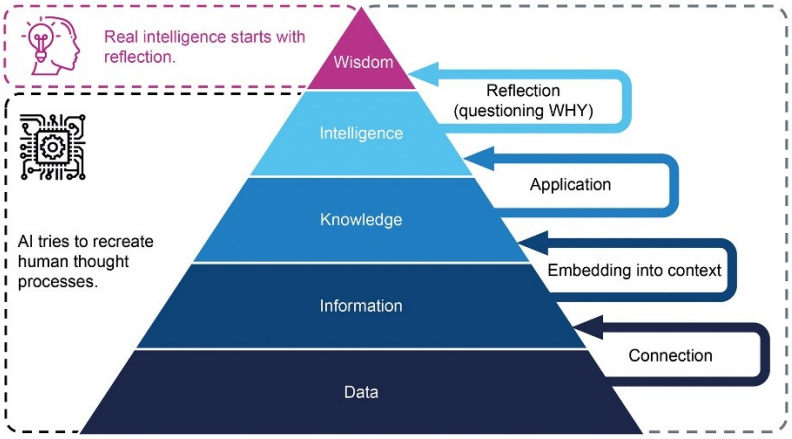


Figure 16: The Hans-Christian Boos pyramid. Source: Melby and Kurz (2021).

In that sense, de Rosnay (2016, pp.66-67) also emphasised that the term "intelligence" should be reserved for human intelligence, which is "rational, symbolic, emotional and capable of projecting itself

into the future or into the past". According to him, systems which are described as intelligent should share three fundamental properties:

> *"They are interactive (they respond immediately to any request), proactive (they anticipate your request by memorising past requests and your search history), and act in real time (the responses arrive before a set deadline)"* [36]

While MT systems appear to adhere to these properties (particularly in an IMT setting), it is evident that they lack the qualities of human intelligence.

3.3.3 Fragmentation & Decontextualisation

The digital transformation has impacted various industries in the tech domain, and translation was no exception. Most freelancers provide on-demand services via digital labour platforms. It has indeed been pointed out that the translation sector was reorganised based on an Uberisation model, which was found to negatively affect the working conditions of translation professionals (Fırat, 2021). This observation is indicative of a broader phenomenon – the emergence of Digital Taylorism in the translation industry (Moorkens, 2020). With the aim of maximising productivity and reducing costs, the translation process is typically fragmented into small units, with freelancers receiving shorter and shorter ad hoc tasks. Although flexibility is a motivating factor for several freelancers, it comes at the cost of instability. A decrease in text length has indeed been observed for translation jobs (Moorkens, 2020). In a Taylorist model, workers can easily be swapped. This idea now applies well to the translation workflow, since a correct use of TMs guarantees consistent wording across documents independently of the translator. It should also be noted that large projects can be distributed among several translators, which can instil a sense of decontextualisation for linguists, as they may find it difficult to identify the purpose of these tasks. In that regard, Mossop (2006) argued that chunking (i.e. dividing a translation task into chunks which are sent to different translators), and the division of labour are likely to affect the mental process of translation, which might result notably in losing the overview of the text as a semantic whole. Consequently, in addition to threatening sustainability, these phenomena have direct impacts on the translation process.

Besides the translation workflow, the textual level is also affected by this fragmentation. O'Brien (2012) indeed stated that the concept of the text itself ("with a beginning, middle and end") underwent radical changes because of how content is produced today. In a CAT environment, texts are typically segmented, which has a number of implications for the translation task. Text segmentation in CAT tools is also

[36] Translated from French.

deeply affecting translation and PE by encouraging translators to process texts on a sentence-by-sentence basis. Taking inspiration in Saussure's definition of the two dimensions of language (de Saussure, 1971), Pym thus argued that segmentation continuously interrupts linearity in such a way that "the paradigmatic is imposed more frequently on the syntagmatic" (Pym, 2011, p.1).

Segmentation is essential to store translated units in the TM; however, this mechanism entails consequences, both for the translation process and its product. Although selecting the sentence as a basic translation unit is convenient from a practical perspective, it does not appear to accurately reflect linguistic phenomena. Candel-Mora (2015a) indeed found that the number of sentences for a comparable corpus of press releases differed from one language to another. This finding throws into question the fixed segmentation mechanism of CAT tools, as mapping sentences across languages is not a realistic strategy. It can be argued that some functionalities are available to translators who would like to merge or split segments (although these can be blocked if deemed relevant during the project setup phase). However, the typical sentence-based layout in most CAT tools creates an "unnaturally strong focus on sentences" (Dragsted, 2006, p.253) and is likely to induce translators into sticking to predefined segments, which can therefore threaten the quality of the final translation in terms of textuality. Besides the number of sentences in a text, textuality is determined by other factors operating beyond the sentence level, such as cohesion and coherence. This seems to confirm Pym's observations regarding the loss of syntagmatic cohesion.

Conversely, it could also be argued that CAT tools provide relevant support to operate at the document level. In a study focusing on the use of technology in literary translation, Daems (2022, p.62) mentioned Termsoup as a tool which claims to offer functionalities allowing translators to "work beyond the sentence level". In a sense, CAT tools come with functionalities allowing users to ensure a certain level of cohesion in their translations. This is the case of concordance search and terminology support. Daems (2022) also pointed out that translators are likely to switch between the CAT environment and a visualisation of the target text. To some extent, the need to see the target text in the context of the final document is catered for in some tools via a document preview feature, as it is the case in Smart Editor[37].

3.3.4 A Cognitive Perspective

Cognition has to do with how translators process and convey meaning. Cognitive approaches in TS are therefore essential to gain a deeper understanding of the mental processes involved in translation, including memory, decision making, and the interpretation of texts. According to O'Brien (2020), the integration of MT in CAT tools, together with other changes (such as IMT) are expected to affect the translation (and by extension the PE) process from a cognitive perspective.

[37] https://languagewire.zendesk.com/hc/en-gb/articles/17195309027613-Preview (consulted: 30/04/2025)

Pym (2011) argues that memory is the process dimension which is the most impacted by translation technology. In a CAT tool, the TM indeed acts as an extension of the translator's memory, by centralising and retrieving knowledge in the form of previous translations. While this can contribute to boosting productivity and ensuring consistency, the suggestions offered by an external memory may entail a demanding decision-making process. When the translator is presented with several alternatives, analysing each of them and selecting the most appropriate one is not always as efficient as translating from scratch. Such a difficulty can be exacerbated when MT is used in addition to TMs.

Dragsted (2006) argued that CAT relies on a form of collaborative cognition in which TMs play a central part, since "rather than having the sole authority to decide on the wording of the TL text, the individual translator 'surrenders' part of the decision to the TM system, and via the TM system, to other translators in the group" (Dragsted, 2006, p.253). It should be noted here that, in this study, "TM system" most likely refers to a CAT tool. Based on this observation, this study proposed to further analyse how the external representation of the translation task (i.e. the design of TM systems) influences the internal representation (i.e. the translators' mental representation). The outcomes of this study revealed significant differences between translation tasks completed with and without a TM. It was found that professional translators were aware of the potential shortcomings of the sentence-based segmentation but deliberately chose to adhere to the content offered by the system. More specifically, translators tended to revise more carefully the text during the translation stage and to dedicate less time to revision at the end when a TM was used. The sentence structures also tended to remain unaltered when working with a TM, which could be explained by the strong sentence-based focus imposed by TMs, the high technical cost of editing sentence structures, and the fact that such modifications would typically occur in a last revision stage (which was not the case when working with a TM). Such behavioural changes therefore corroborate the strong reliance on TM content (i.e. the observation that translators seem to "hand over" the decisions on the translation to the TM system). It can be concluded that the decontextualised, sentence-based structure of TMs affects cognitive processing by modifying the representation of the text (from a coherent whole to a mere sequence of sentences), which also impacts the translation product. Based on these findings, Dragsted (2006) suggested considering paragraph units instead of sentences. The conclusions of this study should, however, be nuanced. It could be argued that the preview functionality encourages translators to consider the document level. Moreover, some CAT tools come with paragraph indicators allowing translators to identify the beginning of a new paragraph. In Trados Studio, for example, a capital letter "P" is displayed to the right of the target side as an indication that a group of segments belongs to the same paragraph.

Since translation and PE mostly occur within a CAT environment, CAT tools should assist translators in their work without interfering with the process. This is, however, not always the case, as interacting with technology is influencing the translator's mental processes (Bundgaard, Paulsen Christensen and

Schjoldager, 2016). Pym (2011, p.1) indeed contended that translation technology is "altering the very nature of the translator's cognitive activity". O'Brien et al. (2017) pointed out that translating entails a certain degree of intrinsic load (i.e. the cognitive effort involved in the task itself), to which is added an extraneous load (i.e. the effort related to extrinsic elements connected to the main task, which can arise from interactions with technology in the case of translation). In the words of Cooper (2004, p.19), cognitive friction is "the resistance encountered by a human intellect when it engages with a complex system of rules that change as the problem changes". This friction can thus emerge when external elements – such as suboptimal CAT functionalities – disrupt the flow of the primary activity. In that regard, Krüger (2016) took the example of QA checks and pointed out that a high amount of false positives can be a hindrance to the human translator process.

The concept of resistance plays a significant role in Andrew Pickering's model of the "mangle of practice", which is defined as a "temporal structuring of practice as a dialectic of resistance and accommodation" (Pickering, 1995, p.xi). This intertwining of human and non-human agents following a continual and mutual adjustment is described as a "dance of agency". Olohan (2011) applied Pickering's framework to model individual and collective interactions between translators and translation technology. She explored how such a "dance of agency" takes shape in translation, that is to say how "the technological […] and the social […] are, in Pickering's terms, reciprocally 'tuned' to achieve an interactive stability" (Olohan, 2011, p.343). Along the same lines, a study by Bundgaard, Paulsen Christensen and Schjoldager (2016) demonstrated that translators experience both resistance and accommodation patterns (Pickering, 1995; Olohan, 2011) when working in a CAT tool with MT. In the experiment conducted as part of this research, the translator did not reject any of the pre-translations, which emphasises the strong influence of this element. This phenomenon was due to the uncertainty perceived by the translator regarding the pre-translation origin (TM/MT), which resulted in sticking closely to the suggested text. In this study, the CAT tool was set up in a way that CMs and EMs were auto-confirmed, and the translator was therefore encouraged to skip those segments. Such a behaviour would contribute to deteriorating the cohesiveness of the resulting translation. However, while Dragsted (2006) found that translators tended to (deliberately) "hand over" the decisions to the CAT tool (which in this case would mean not editing CMs/EMs), Bundgaard, Paulsen Christensen and Schjoldager (2016) observed an opposite phenomenon. On two occasions, the translator indeed accepted EMs but came back to edit them at a later stage in the translation, thus contradicting the sentence-based approach. This resistance pattern suggests that the translator was able to maintain a document-level representation of the text throughout the process. Nonetheless, it should also be noted that certain functionalities were found to have a positive influence, in particular regarding compliance with customers' requirements (such as consistency). More recently, Briva-Iglesias and O'Brien (2022) distinguished three different attitudes towards technology in the language industry: resistance, collaboration, and reinvention. They advocate for the last two, as technology is already here and would not be relevant to fight against it.

3.4 Practical Challenges in Post-Editing

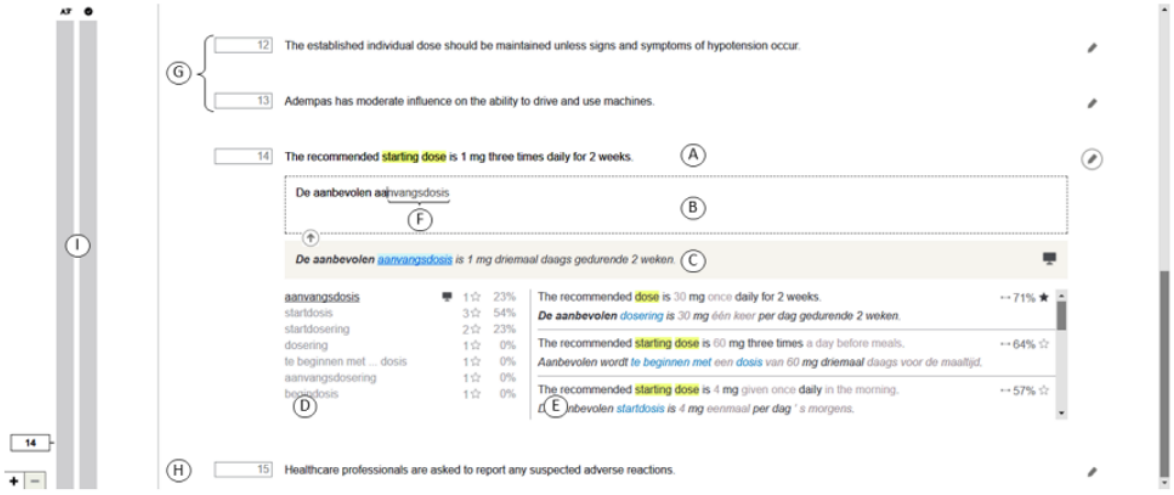
In addition to theoretical considerations, the use of tools has a significant impact on the PE process. Ideally, technological assistance should empower translators and ensure that PE can be performed as smoothly and efficiently as possible. In practice, however, several studies have reported that post-editors encounter difficulties. This suggests that usability of such tools could be improved for PE.

3.4.1 Technological Assistance for Post-Editing

As discussed previously, PE is a complex activity. It is therefore crucial that tools provide relevant assistance in the PE process. Early conceptual proposals clearly envisioned a symbiosis between translators and technology (e.g. the "symbiotic partnership" described by Licklider [1960] and the concept of the Translator's Amanuensis introduced by Kay [1980]). According to Carl, Bangalore and Schaeffer (2016, p.4) the extensive amount of data collected as part of TP research projects now allows for anticipating the type of assistance desired exactly when it is needed. In a way, IMT appears as the perfect concretisation of this hypothesis, as prediction is at the core of this paradigm. Arguably, the same can be said of PE: the analyses of the PE process conducted so far provide insights into how technology can support post-editors. do Carmo (2017, p.199) pointed out that "the current challenge for PE is not so much on improving the quality of the MT output, but on giving translators proper conditions for their job, including PE". Therefore, it appears paramount to ensure that PE environments meet the expectations of post-editors. Candel-Mora (2021) identified three aspects in which technology should help to optimise the performance of translators. Namely, it should make translation faster (i.e. not having to retranslate content), easier (i.e. using a dedicated workstation that allows to focus on translation only), and more efficient (i.e. avoiding repetitive tasks). Arguably, PE environments should have similar objectives, albeit adapted to the PE process. While the PE process was covered in the preceding section, the remainder of this section is dedicated to the study of tools which support PE and their impact on the PE process.

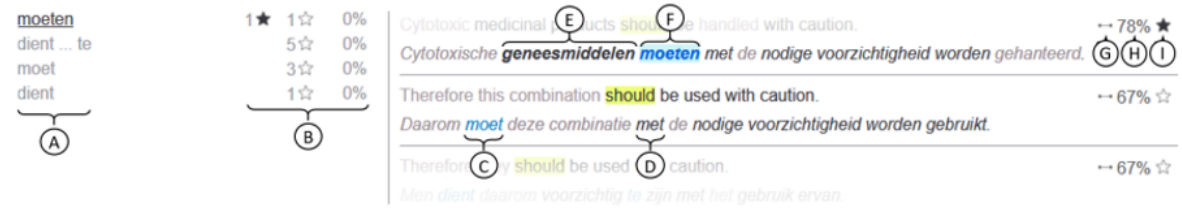
Over the last few years, there has been a tendency to design tools which are more tailored to PE. CATaLog, for example, was developed with the objective of improving productivity and experience (Pal et al., 2016). In particular, this tool offers various translation suggestions (including APE) and a colour code for an easier identification of the similarities between an input sentence and the matched segment from the TM. Coppers et al. (2018) sought to optimise intelligibility in CAT tools and presented Intellingo, a tool in which the visualisation of translation suggestions allows translators to find out where suggestions come from and what their relationship is with other suggestions, as exemplified in Figures

17 and 18. The TM section shows the method used to retrieve fuzzy matches (edit distance or syntax tree). A hybrid MT system combines MT suggestions and TM fuzzy matches and displays the origin of each fragment (the parts of MT which come from the TM are signalled, so that translators can understand better how the MT suggestion was constructed). Translation alternatives also show origin information (MT, TM, TB), and come with usage frequency data (number of occurrences in TM and MT suggestions). Furthermore, when hovering over a suggestion, the corresponding source and target elements in the suggestions are highlighted in yellow and blue, respectively, for an easier visualisation. The results showed that a more intelligible interface could contribute to helping translators in their decision-making process by providing relevant information. However, it did not significantly impact UX or efficiency.



An overview of the Intellingo user interface when translating: (A) the source segment, (B) a text area for the user translation, (C) machine translation, (D) alternative word groups, (E) related earlier translations (translation memory), (F) auto-completion, (G) two preceding segments, (H) the next segment and (I) a progress bar indication of translation and revision. Upon hovering over the translation suggestions, the related phrases are highlighted in yellow (source language) and blue (target language) across all suggestions.

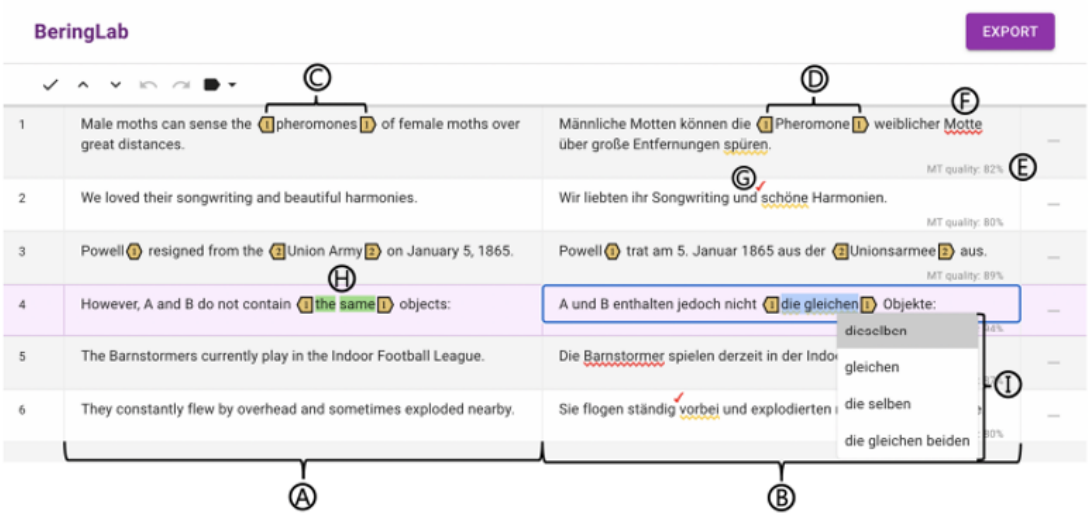
Figure 17: Screenshot of the Intellingo interface. Source: Coppers et al. (2018, p.2).



For the word (group) being translated: (A) a list of alternatives and (B) an overview of its occurrences in machine translation, translation memory and term base to explain the ranking. Words in the fuzzy matches are shown (D) darker when they match with the source segment. Occurrences of all alternatives are highlighted in blue (C), and the corresponding source words are highlighted in yellow. The occurrences of the translation option selected by the user are emphasized (F). When parts of a match were used by the machine translation algorithm, these parts are shown in bold (E) and (I) a filled star is shown. Other intelligible details include (G) the matching metric and (H) the match score.

Figure 18: Screenshot of the Intellingo interface, focusing on alternative translations. Source: Coppers et al. (2018, p.5).

Another adaptation of CAT tools to PE consists of integrating confidence scores based on QE, at the word and/or sentence level, as done in IntelliCAT (Lee et al., 2021; see E, F and G in Figure 19) and as proposed by Wei, Utsuro and Nagata (2022). The MMPE interface (Herbig et al., 2020) was also developed in an attempt to enhance the PE experience. Despite these initiatives, it seems that there is still no consensus regarding how CAT tools can be further adapted to PE.



The IntelliCAT Interface. After a document (i.e., an MS Word file) is uploaded, (A) sentences from the original document (source) and (B) the initial MT output for each sentence (target) are shown side-by-side. (C) Formatting tags indicate where a specific style (identified by an integer style id) is applied and (D) are automatically inserted at the proper position of the MT output based on word alignments. (E) The interface shows the quality of each machine-translated sentence based on sentence-level QE. (F) Potentially incorrect words and (G) locations of missing words are highlighted based on word-level QE. When the user selects a sequence of words in the MT output, (H) the corresponding words in the source sentence are highlighted with a heat map, and (I) up to five alternative translations are recommended.

Figure 19: Screenshot of IntelliCAT. Source: Lee et al. (2021, p.12).

More recently, Landwehr, Steinmann and Mascarell (2023) introduced OpenTIPE, an editing environment in which APE suggestions are displayed in the UI, and incrementally improved based on user corrections (see Figure 20).

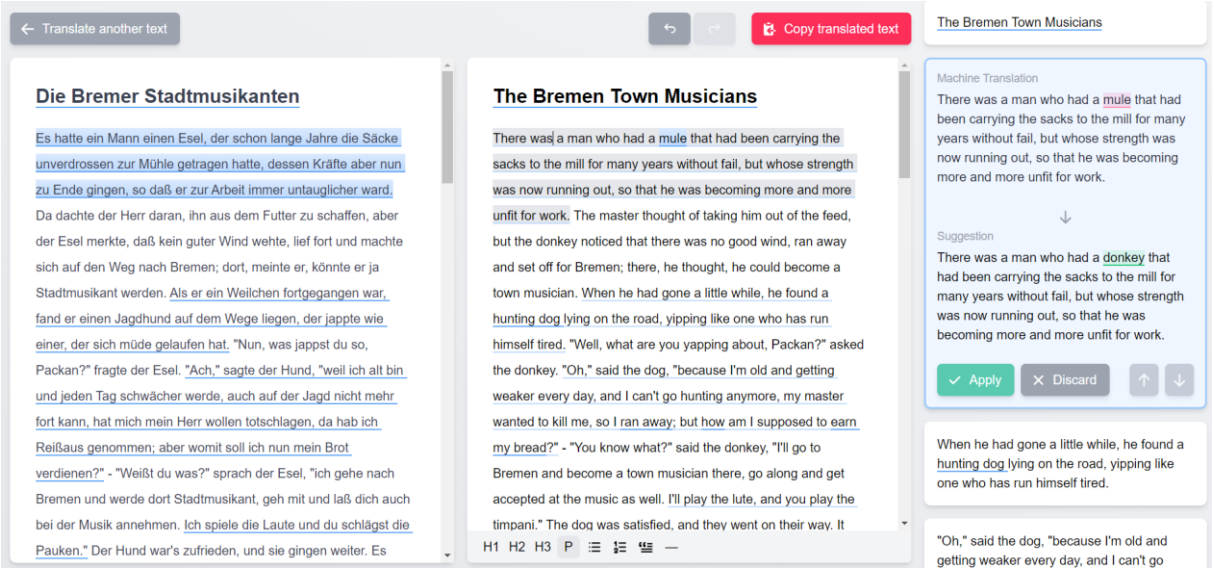
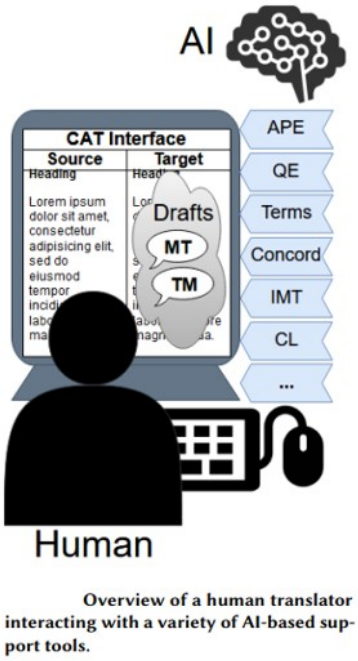


Figure 20: Screenshot of OpenTIPE. Accessed from: Landwehr, Steinmann and Mascarell (2023).

Herbig et al. (2019) discussed possible research avenues and challenges deriving from the combination of human and artificial intelligence for translation. Recent developments have resulted in the creation or improvement of tools and resources translators are interacting with, leading to an increasing level of technological assistance, as depicted in Figure 21. Therefore, it is paramount to investigate the best ways to integrate and use this intertwining of technologies in the PE environment.



Overview of a human translator interacting with a variety of AI-based support tools.

Figure 21: Overview of human-AI interactions in a translation tool according to Herbig et al. (2019, p.1).

To that end, Herbig et al. (2019) listed essential directions for optimal leveraging of synergies between humans and technology. In particular, they pointed out that, despite the advances in technology and the integration of MT, the interface of translation tools did not evolve. MT outputs are indeed used either to directly populate target segments or can be made available in a translation suggestions section. Apart from this integration, no further adaptation to the task of PE has been developed in most CAT tools. However, PE entails different user actions compared to translation from scratch, such as a reduced number of mouse and keyboard events, which calls for exploring other interaction modalities, like MMPE (Herbig et al., 2019). Error detection and correction is a core task in PE, which suggests that PE environments should ideally provide adequate support to allow post-editors to identify errors and make the necessary adjustments as efficiently as possible. To that end, Herbig et al. (2019) recommended three types of assistance, namely QE, alignments between the source text and the MT output for a fast comparison, and colour coding to identify similarity between source segments and their matches retrieved from a TM. The authors recognised that certain components of CAT tools (e.g. concordance search and TBs) can provide valuable assistance but also encouraged researching whether other support tools could be well suited to PE. They also suggested providing tailored assistance during challenging parts of the workflow (which could be detected via measures of the cognitive load). In addition, they emphasised the importance of finding an optimal balance in the number of translation suggestions presented to the post-editor, and mentioned other forms of assistance, such as dynamic suggestions based on user input, incremental adaptation of the output based on APE models to avoid repetitive corrections, and support in detecting bias in MT outputs.

3.4.2 Analysis of Translation Tools from the PE Perspective

While translation technology undoubtedly constitutes valuable assistance for translation professionals, particularly in terms of productivity and consistency, it does not come without certain shortcomings. In a survey, O'Brien et al. (2017) indeed found that more than half of respondents found certain CAT tool features irritating. This finding only reinforces the need for understanding how CAT technology can be improved based on user needs.

Nunes Vieira and Specia (2011) compared various tools for PE with a special focus on certain elements of the interface. They distinguished a baseline comprising features that were common to the majority of tools analysed (including, among others, colour codes assigned to the different fuzzy match bands, the presentation of both source and target texts, shortcuts, concordance search, integration of glossaries), and additional features, which were less common (such as spelling / grammar / style checkers, multiple TM+MT integrations and QA checks). Furthermore, this study also introduced a list of desirable features, which at the time were absent in translation editors but were deemed relevant for supporting PE. These included notably enhancements to the TM+MT integrations, change tracking and MT confidence scores.

Moorkens and O'Brien, 2013 investigated user attitudes towards PE interfaces. Their research stemmed from the observation that CAT tools were created to effectively support editing of translations produced by humans (TM matches) but may not be optimised for editing translations generated automatically (MT outputs). They found that users required the current CAT interfaces to be optimised for regular translation tasks, before adapting these to PE. The improvements to current interfaces included the following items:

- simplicity (clean and uncluttered User Interface [UI])
- customisation, rather than using default set up (including layout, tag visibility, font type, colours)
- plug-ins for dictionary and Internet search
- improved concordance search
- global find-and-replace function
- dictionary plug-ins
- reliable concordance features
- QA interoperability (particularly with ApSIC Xbench)

PE-specific needs were also identified, in particular:

- MT confidence scores (in a percentage format or via colour codes)
- presentation of matches (most participants wanted to see MT+TM)
- subsegment integration of MT+TM (however, participants shared concerns about origin display; they appeared to be in favour of marking with colours but also marking as "MTM" match)
- dynamic MT improvements (IMT)

Further needs which are relevant for both translation and PE were also found. For example, shortcuts (e.g. dictionary search, reject MT suggestion, change capitalisation, etc.) appear relevant as they allow for increasing productivity in both cases. However, language-specific shortcuts (e.g. for changing gender) proved not to be a popular feature, as respondents expressed doubts as to how such a feature could work and argued that it was likely to be difficult to remember a high number of shortcuts. What emerged from this study is that it appears challenging to find a balance between the information to be made available and the need to have an uncluttered UI. Respondents indeed seemed to require as much information as possible (e.g. confidence scores, marking origin in repaired matches), which can result in conflicting needs at the interface level (e.g. it is likely to be difficult to find a clear way of combining terms highlighting, spellcheck errors display and marking the origin of a repaired match inside the same segment). Four years later, Moorkens and O'Brien (2017) extended this study and confirmed the findings of the previous round, in particular:

- Translation tools would need to be improved, even before considering PE
- Customisability and shortcuts were found to be essential elements (although with the same reservations regarding language-specific shortcuts)
- Most participants would like to see confidence scores

- Combining sub-segments (MTM) was deemed useful by a majority (60%+) of respondents
- Dynamic adaptation of MT was also an item of the UI wish list

Certain PE-specific features were also mentioned (including notably word reordering, terminology and grammar). When asked if they favoured simplicity of the UI or feature richness, the participants expressed that what mattered was not necessarily supporting a large number of features but rather having the "right" ones. As acknowledged by the authors, it might be questionable to ask for feedback on features which users could not test already (not available yet), as it can be difficult to imagine how these would work in practice. However, the authors emphasised that gathering opinions is a fundamental step in the pre-design phase, and that user needs should be refined in subsequent testing and evaluation phases.

Zaretskaya (2017) also investigated user needs related to translation technology via a survey. The results allowed for an identification of the most popular features but also shed light on possible improvements. The concordance search functionality appeared to be the favourite feature for the majority of respondents, which confirms the relevance of retrieving contextual examples from the TM and might suggest that incorporating additional corpora could also be useful. Features related to terminology management were also deemed essential by the participants, although they did not seem to meet the expectations of users in their current implementation. Moreover, the most problematic aspect of CAT tools was found to be multifunctionality and the resulting lack of customisation. Translators expressed their desire to use different versions of the tool depending on the context (e.g. "light" vs more advanced version). In addition, she found that the use of translation tools varies according to the user profile. For example, education and training are likely to have an impact on the use of technology.

When discussing how PE can be supported by technology, it also seems relevant to assess the assistance provided by IMT systems. In such settings, the MT output is constantly updated based on the inputted translation (as explained in 1.3.4). On the one hand, suggestions are dynamically adapted, and in the event that they coincide with the translation that the linguist has in mind, they can directly accept the adapted suggestion. On the other hand, reading the constantly changing suggestions might entail a high cognitive load if they do not coincide with the translation being formed in the mind of the translator. In both cases, IMT alters the PE process, since translators do not focus on spotting errors in a static output but rather reassess the quality of each suggestion on the fly. In that regard, do Carmo (2017, p.180) argued that PE is not correctly modelled by predictive typing as the translators' mental processes "are divided between [their] mental representation of what the translation should be and the metamorphosed suggestions [they see] ahead of [their] writing".

Furthermore, it seems that there is no clear consensus regarding the potential benefits of effort indicators. Moorkens et al. (2015) studied the impact of confidence estimation scores on PE. In this project, the PE effort was first estimated by raters, which resulted in the production of Post-Editing Effort Estimation Indicators (PEEIs). It should be pointed out that contrarily to confidence scores, which are generated

automatically from QE, PEEIs are set by humans. The findings showed that the presentation of PE effort indicators at the interface level did not impact the actual PE effort, which throws into question the usefulness of confidence scores for PE.

The findings of Moorkens and O'Brien (2013, 2017) regarding the relevance of customisability were reinforced by O'Brien (2021), who emphasised the importance of personalisation. She envisaged relatively advanced levels of personalisation by calling the reader to

> *"Imagine a technological aid that learns about the specialised domain a translator is most interested in and finds only appropriate resources, discarding ones that are less relevant or less trustworthy. Imagine an aid that understands the context in which a translation is being produced and tailors its features accordingly. […] Or imagine one that learns about the uncertainty tolerances of each individual translator and only presents suggestions based on those tolerances. Imagine a technology that can detect, based on gaze information from eye trackers embedded in our personal devices, when a translator needs assistance and when she or he is in a cognitive 'flow' and should not be interrupted with prompts"* (O'Brien, 2020, p.385).

She, however, acknowledged that achieving such a level of personalisation is far from a trivial task, which only reinforces the need for further studies on this topic. Detecting when translators need assistance certainly constitutes a challenging task. This idea has been explored by Porto Veloso (2013), who presented Escriba, a web-based CAT tool with an adaptive user interface. This tool builds on principles of personalised learning (Liu, Wong and Hui, 2003), which records interaction events and attempts to predict the next user action based on identified patterns, as depicted in Figure 22.

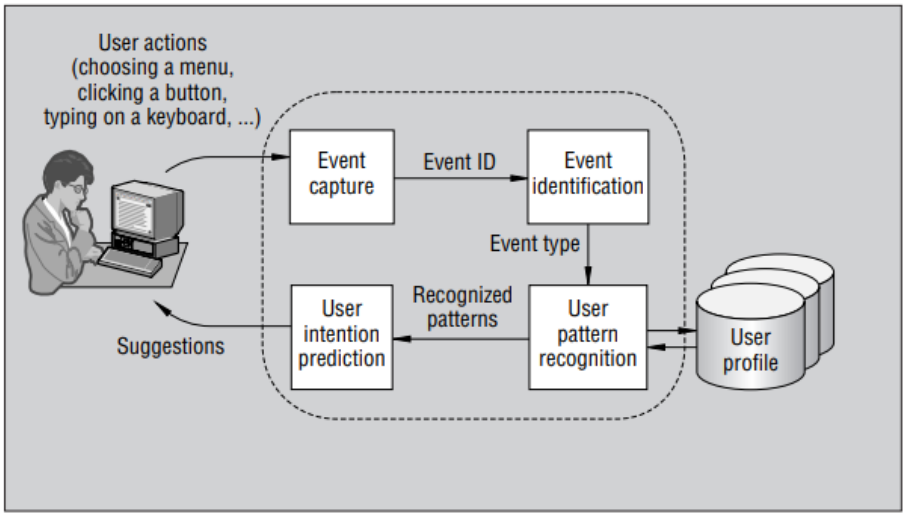


Figure 22: Architecture of an adaptive user interface. Source: Liu, Wong and Hui (2003, p.54).

The user study of Escriba, however, suggested that such interfaces do not always have a positive impact on usability. Arguably, the field would benefit from further research on adaptive interfaces in translation environments.

3.4.3 The Central Place of User Feedback

Beyond the impact of technology on the PE process, it also appears essential to study the overall perception of post-editors working with MT and CAT tools. This is particularly relevant given that perception can be conditioned by past experiences. Briva-Iglesias and O'Brien (2022) indeed found that the perception of post-editors who are dissatisfied with MT correlates with the adequacy of the final translation. The authors stated that this observation does not apply for CAT tools, as they are already well accepted by translators. It appears reasonable to say that translators are more comfortable with CAT tools than with MT because they have been around for a longer time. Nonetheless, one could hypothesise that this higher acceptance of CAT tools can be justified precisely by the same argument: a longer exposure, resulting in translators becoming more familiar with CAT tools. This raises another concern, as it can thus be argued that translators got used to the limitations of CAT tools and adapted to them over time.

As discussed earlier, although technology is becoming increasingly present and important for translators, there appears to be a dichotomy regarding how translation tools both empower humans (augmented translation) and affect the translation process (cognitive friction). A possible root cause for this may stem from the lack of participation of the most impacted players (i.e. translators themselves) in the process of developing these tools. This observation has been emphasised by various scholars. It has been put forward that the main interest of companies which develop translation tools has always lied in increasing translators' throughput and in adapting to the needs of customers. As a matter of fact, when discussing TM matches, O'Brien, O'Hagan and Flanagan (2010, p.190) stated that

> *"UI design has […] been driven by the needs of the translation client and not by the needs of the translator, the ultimate user of the software in question". do Carmo (2017, p.6) also observed that "one of the reasons why the development of translation aids has never been central to TS may have been the fact that the tool manufacturers' main objective has been to increase translators' output, leaving language quality to the responsibility of translator trainers, which happens to be in line with the main concern of TS scholars."*

Along the same lines, Lagoudaki (2006, p.2) pointed out that "many of the existing commercial TM systems are technology-driven applications (e.g. with an abundance of useless features and a complex,

impractical and difficult to learn UI), rather than user-driven applications". She also drew attention to the fact that, for the most part, CAT tool users are only invited to share feedback on a tool at a stage in which the product is almost fully developed and ready to be launched. This implies limitations as to the changes that can be implemented before making the tool available on the market. Involving translators at earlier stages of design therefore seems a desirable practice. Likewise, O'Brien (2012, p.115) argued that "there is little evidence to suggest that tools that are proposed as aids to the translation process have been designed from the point of view of the humans who have to use them". According to Ehrensberger-Dow and O'Brien (2015, p.122),

> *"if we take seriously the notion of translation being a situated activity, then we have to provide opportunities for translators to explain what their personal preferences are and to have their voices heard by designers of translation workplaces and language technology".*

When discussing the degradation of working conditions of translation professionals, Fırat (2021, p.66) asserted: "translation workers have not been included in key decision-making processes". On a similar note, Moorkens (2020, p.26) contended that

> *"responsible and ethical use of Taylor-influenced methods will require a balancing of efficiency with the need to move away from extractivism for all resources and to consider long-term value. This necessitates translation employers bearing some responsibility for the satisfaction and motivation of workers, as a sustainable industry will ultimately benefit all stakeholders".*

Zaretskaya (2017) argued that software developers tend to prioritise functionality at the cost of usability. While she acknowledged that usability is a more abstract concept, she strongly recommended considering it when evaluating CAT tools. Candel-Mora (2021) also commented on the tendency of developers of translation tools to consider the translator as a single end user, which results in failing to account for the different translator profiles and translation scenarios (type of tools in use, language pair, text and content type). This observation is in line with O'Brien (2020), who also highlighted that the predominance of the "one-size-fits-all" approach. In addition, Candel-Mora (2021) deplored the lack of communication between developers and end users (i.e. translators) and considered this as the explanation for the absence of a complete and effective translator workstation. O'Brien (2020, p.382) shared the same concerns, as she raises the following questions:

> *"What proportion of the programmers who have designed TM or terminology management tools have ever translated content? What proportion of MT system developers are translators? What do they know about the cognitive process involved in translation?"*

Therefore, the striking lack of communication between developers and users of translation technology appears to be the root cause of the dissatisfaction that translators experience with translation tools. This phenomenon emphasises the need to follow user-centred design principles in the field of translation technology. In particular, Gould and Lewis (1985) presented three system design principles which appear highly relevant here. Namely, these principles are:

- An "early focus on user and tasks" (it is paramount for designers to gain an in-depth understanding of the users who will be working with the system, and the work they will be accomplishing)
- "Empirical measurements" (analysing the performance of users working on real projects using prototypes)
- And "iterative design" (entering a cycle of iterations in which the design is tested and adjusted until reaching a satisfactory version).

According to Gould and Lewis (1985, p.302), users should be able to "instill their knowledge and concern into the design process from the very beginning", which appears as an optimal solution to the issues described above. This suggestion is reminiscent of agile software development. Since software developments are often occurring in a rapidly changing environment, defining a complete set of system requirements from the start may not always be feasible or desirable, as users' behaviour may not be easily predictable in a new environment. For that reason, agile methods aim at involving users from the start in the development process and delivering frequent releases, rather than focusing on requirements engineering (Sommerville, 2011). Agile development therefore appears as a relevant methodology to gain insights from users from the very beginning. Given the sometimes unpredictable nature of users' requirements, especially in a constantly evolving technological landscape, Pickering's model also seems to apply here. Marick (2008) even described agile methods as a "manglish way of working", as the evolution of a system cannot always be known based on the initial conditions, which implies an interplay of resistance and accommodation.

Such recommendations would be highly beneficial for CAT tool development, as they would contribute to building more user-centred applications and to making translators feel included in the development of the tools they are using on a daily basis. Arguably, these methods have not been adopted (or at least not to a satisfactory extent) by most CAT tool developers, which could explain the translators' complaints emphasised above. Nevertheless, some exceptions are worth mentioning. For example, COACH was based on translators' feedback from the design phase, thus "creating a sense of stakeholdership" among involved linguists (Bota, Schneider and Way, 2013, p.313). Similarly, the development of Intellingo strongly relied on a user-centred approach in which professional translators were heavily involved (Coppers et al., 2018). Consequently, these approaches should serve as relevant exemplifications of ideal processes to follow for the development and improvement of CAT tools.

3.5 Summary & Outcomes of this Section

This section allowed for gaining knowledge about the PE process as well as the existing tools to help post-editors. Defining the PE process is not an easy task. As the integration of MT in CAT tools is relatively recent, further studies would help reach a more comprehensive description of the PE process. In that regard, Mossop (2006) deplored the lack of observation of the translator's workplace in real situations. The same can be said of PE today.

In view of the close relation between PE and translation, it was considered relevant to explore first the translation process before defining the PE process. It was observed that both processes are divided into three main phases. However, there are significant differences between them: in the case of PE, the target text is already present when starting the task, which leads to differences in the way the text is processed. The bulk of the work no longer consists of generating the translation, but of identifying and correcting errors in an MT output. This naturally affects the mental processes of translators, since instead of producing a translation from scratch, they perform mostly editing actions (i.e. deleting, inserting, moving and replacing chunks of text). The definition of the PE process was complemented by a study of the guidelines and skills needed in PE. This examination showed that identifying errors is a central part of PE. In addition, decision making (e.g. avoiding preferential changes) and the application of guidelines are also essential.

As with the study of the PE process, the study of the PE environment was first conducted from the perspective of translation, based on CAT tools. This analysis identified certain deficiencies in these tools (such as decontextualisation and lack of customisation) which also apply in the case of PE, as this task is performed in the same tools. Numerous studies, based on translation or PE, have focused on users' perception and identification of their needs, and have revealed that there is substantial room for improvement. Apart from a few exceptions (e.g. IMT and QE), the integration of MT into CAT tools seems to have occurred without any specific adaptation to the needs of PE (error correction appearing to be the main one). It therefore seems essential to study the needs of post-editors in more depth in order to adapt CAT tools accordingly.

Overall, there is a discrepancy between the idea that augmented translation increases translators' capabilities by making them more efficient, and the reality of the profession, whereby MT is often imposed on translators, who must also work in tools that do not fully satisfy them. Therefore, rather than perpetuating the pattern of resistance and adaptation, it would appear appropriate to assess the PE environment with a view to identifying ways of improving it. The feeling of not being heard appears to be a recurring concern for translators. Therefore, taking their feedback into consideration also seems essential.

Finally, in addition to modifying the cognitive processes of translators, the integration of MT could have repercussions on the quality of the final product. The study of the guidelines and standards in the industry

has shown that, in the case of full PE, it is generally expected that the quality of the final text should be equal to or similar to that of a human translation. However, one of the difficulties of PE lies precisely in the identification of errors that are not obvious at first sight, which constitutes a quality risk.

It therefore seems essential to better understand the concept of quality in translation and PE and to examine how it can be measured.

4. Quality in Translation

This section discusses the concept of quality in translation. Before delving into the various definitions proposed, it is crucial to clarify the focus adopted here. Quality management is a complex process, and it can become easy to confuse some of its steps (e.g. "quality assurance" and "quality assessment"). Ensuring high translation quality involves a series of processes at different stages of the translation workflow: before, during and after production. These processes form part of a quality management strategy. Vela-Valido (2021) provided an overview of the different steps of the quality management workflow based on several industry standards (ISO 9000 and 17100; ASTM F2575 and WK46369), as summarised in Table 19.

Quality Management		
Before production	**During production**	**After production**
Quality planning: establishing quality objectives and processes to complete to achieve these objectives ***Quality estimation***: predicting the quality of an MT output without access to a reference	***Quality assurance***: procedures implemented to ensure quality, including: • ***Revision*** (bilingual examination of the target text) • ***Review*** (monolingual examination of the target text) • ***Formatting and compilation*** • ***Proofreading and verification*** (checking for spelling, formatting and typographical errors) ***Quality control***: checking of either a sample or the entire deliverable against specified quality requirements	***Translation Quality Assessment (TQA) / Translation Quality Evaluation (TQE)***: evaluation procedure undertaken to assess the extent to which the requirements have been fulfilled, normally before delivery ***MT evaluation***: assessment of the raw MT output, which can be completed via human or automatic evaluation, and which can involve a comparison between the initial output and the final delivery

Table 19: Processes involved at the three stages of the quality management workflow, adapted from Vela-Valido (2021).

This classification into three phases makes it evident that quality-related procedures are expected to occur throughout the entire production process. Namely, quality planning and estimation happen before production; quality assurance and control occur during production; quality assessment (also called evaluation) is conducted after production. In this section, although quality assurance (i.e. processes in place to ensure quality before delivery) is briefly discussed, the concept of quality is discussed from the perspective of quality assessment (i.e. evaluation of the final text). Consequently, this section examines the views of both the industry and academic community regarding quality. After starting with a discussion of the quality of human translation, the focus shifts towards MT and presents quality

assessment processes to evaluate this technology. In addition, the impact of using MT as a starting point is also explored, which includes a discussion around the translation quality of post-edited texts.

4.1 Defining Translation Quality

When attempting to define translation quality, one inevitably faces difficulties. This is understandable, as translation is a complex phenomenon. do Carmo (2017) even described quality as an "elusive" concept. Any analysis of a translated text involves considering a wide range of factors, going from cultural appropriateness to technical accuracy. The expected level of quality is also highly dependent on the context. Furthermore, any form of assessment, particularly when applied to language, implies a certain degree of subjectivity. However, defining and assessing translation quality remains a central concern, and these ambiguities often result in debates (Castilho et al., 2018). This section explores the definitions of translation quality in both research and industry settings, and how these have evolved over the years.

4.1.1 Quality Assurance: An Industry Perspective

The industry focuses largely on the production phase itself, which is understandable given the need to deliver the best translations possible. A typical objective for LSPs lies in the implementation of standardised processes (Lommel, Uszkoreit, and Burchardt, 2014; Castilho et al., 2018). To that end, several standards have been introduced, in particular by the International Organisation for Standardisation (ISO) which contributes to establishing standard practises. The ISO 11669 standard on "Translation projects — General guidance" (ISO, 2024) emphasises the role of specifications in regard to quality expectations:

> *"Efficient communication entails that requirements are explicitly defined and agreed-upon as translation project specifications. Successful translation projects are a result of:*
>
> - *developing and following these translation project specifications,*
> - *involving people with the appropriate competences and qualifications, and*
> - *assuring smooth communication flows throughout the projects."*

Similarly, the scope of the ISO 17100 standard ("Translation services — Requirements for translation services"; ISO, 2015a) is defined as providing "requirements for the core processes, resources, and other aspects necessary for the delivery of a quality translation service that meets applicable specifications",

which also highlights the importance of complying with specifications. The ISO 9000 standard ("Quality management systems — Fundamentals and vocabulary"; ISO, 2015b) is no exception, as it defines quality as the "degree to which a set of inherent characteristics […] of an object […] fulfils requirements". Another organisation producing relevant standards is the American Society for Testing and Materials (ASTM) International. Recently, the ASTM F2575-14 standard ("Standard Guide for Quality Assurance in Translation") was replaced by ASTM F2575-23 ("Standard Practice for Language Translation"; ASTM, 2023). This standard specifies essential components for each phase of a translation project, in which specifications play a vital role. Specifications should be identified during preproduction, followed during production, and projects should be evaluated to ensure that they conform to them during postproduction.

It is therefore apparent that quality in industrial settings is entirely dependent on compliance with specifications. Consequently, the guidelines and expectations of customers represent essential factors in this case. Furthermore, various local standards also provide guidance for the provision of translation services, and most mention measures to be undertaken to ensure high quality. Relevant examples of translation quality standards in Europe, as summarised by Corpas Pastor (2006), include the following:

- The Italian standard UNI[38] 10547 ("*Definizione dei servizi e delle attività delle imprese di traduzione ed interpretariato"*), which specifies monitoring and control as service requirements.
- The German standard DIN[39] 2345 ("*Übersetzungsaufträge"*), which is centred on contracts and procedures.
- The European "Quality Standard for Translation Companies" published by the European Union of Associations of Translation Companies (EUATC), which emphasises the crucial role of quality assurance and indicates measures to be taken in order to ensure the customers' satisfaction.
- CEN prEN[40] 15038 ("Translation Services"), a European standard introduced by the *Centre Européen de Normalisation* (CEN) which mentions quality management measures.

The TAUS MT Post-Editing Guidelines (Massardo et al., 2016) constitute another good example of an established set of recommendations in the industry. A principal component in this document is the expected end quality of the final text, which can be either "human translation" quality (full PE) or "good enough" quality (light PE). "Human translation" quality calls for a thorough correction of any type of error in the MT output regardless of the error severity, whereas the aim for "good enough" quality lies in ensuring accuracy and comprehensibility without focusing on the details (e.g. style). The desired MT quality therefore revolves around the purpose of the translation. Texts which are translated for

[38] UNI: Ente Nazionale Italiano di Unificazione

[39] DIN: Deutsches Institut für Normung

[40] prEN: preliminary European Norm

dissemination thus have higher quality requirements compared to those which are translated for gisting purposes (Candel-Mora, 2015b).

4.1.2 Quality Assessment: The Focus of Translation Studies

In contrast with industry players, the academic field – in this case, Translation Studies – does not experience any pressure for delivery to customers and has thus always been more concerned with post-production examinations of quality. Early theories around translation quality were heavily centred on the ST. Nida and Taber (1969, p.12) provided the following definition of translation: "Translating consists in reproducing in the receptor language the closest natural equivalent of the source language message, first in terms of meaning and secondly in terms of style". As the ST and the TT should have an equivalent value, several scholars have introduced theories revolving around the concept of equivalence, and establishing distinctions between translations skewed towards the ST, and those which are more focused on the target context. For instance, Nida and Taber (1969) distinguished formal equivalence, which aims at preserving as much of the source as possible (both in terms of language and culture) and dynamic equivalence, which is more considerate of the audience's perception. Newmark (2003) also proposed two types of equivalence: semantic translation, which attempts to remain as close as possible to the original text, and communicative translation, which seeks to produce on the target audience the closest effect to that obtained on the source audience. Similarly, Toury (2012) formulated a difference between adequacy and acceptability: adequate translations are biased towards the source, whereas acceptable translations are biased towards the target. However, Toury recognised that "the normal state of affairs is a degree of incompatibility between adequacy and acceptability" (ibid., p.70). The concept of equivalence served as the basis for the foundational model of TQA introduced by Katharina Reiss in 1971, described as "one of the first systematic models for TQA" (Castilho et al., 2018, p.12). This model recommends translation strategies based on text type. Several text types were therefore identified, together with their corresponding strategy. Consequently, the translation of informative texts should be centred on the content, that of expressive texts should be more focused on the form, and operative texts should be translated based on an adaptive strategy in order to produce the desired effect on the target readership (Reiss, 2000).

While these approaches remained rather theoretical, House (2015) presented a thorough methodology for TQA, mostly based on functional equivalence, with a first version introduced in 1977. This model was deeply rooted in Halliday's model of discourse analysis, which revolved around the study of the relationship between linguistic choices and the communicative situation (Halliday and Hasan, 1985). In Halliday's model, one of the central elements was register, which was in turn composed of three variables: field (i.e. the subject matter), tenor (i.e. the participants involved in the communication) and mode (i.e. the communication channel). House's TQA model took this definition of register and

incorporated it into a methodology for comparing ST and TT. The aim of this comparison was to identify register mismatches that occur between an original text and its translation. More specifically, this methodology involved producing a profile of the ST register, genre and function. While register was mostly taken from Halliday, genre was added to account for the cultural context in which the communication is embedded. In addition, a statement of function of the ST clarified the function of the text at the ideational and interpersonal levels (i.e. stating the information being conveyed and the relationship between the addresser and the addressees). The same analysis was produced for the TT. Afterwards, the ST and TT profiles were compared in order to identify mismatches, which were categorised as either covertly erroneous errors (i.e., mismatches in the dimensions of field, tenor and mode) or overtly erroneous errors (i.e., mismatches in meaning or TL system). Based on this analysis, a statement of quality could be made, and the translation could be categorised as overt or covert translation. Overt translation was centred on the source audience and context, and therefore did not aim at being received as an original text. On the contrary, covert translation was focused on the target culture and aimed at being perceived as an original, and therefore relied on a cultural filter whereby cultural elements were adapted to the target context. This model was criticised, notably by Gutt (2014), who questioned the usefulness of register analysis when it came to determining the function of a text. Moreover, considering mismatches as "errors" also fails to account for conscious divergence from the ST when implementing certain translation strategies. For example, compensation aims to maintain the overall tone by restoring to another place in the statement the nuance which could not be conveyed in the same place as in the original, and explicitation consists of introducing into the TL details which are implicit in the source language, but which can be understood from the context or situation (Vinay and Darbelnet, 1972). Such strategies are clear digressions from the ST but are sometimes necessary, and it is therefore unfair to consider them as errors. Although the traditional focus on the source started to be questioned, notably by Toury (2012), who suggested to consider translation as a target-oriented framework, in the view of House (2015), the ST continues to play a crucial part and not placing it at the centre results in "fundamentally misguided" approaches (ibid., p.69).

In MT evaluation, quality is also generally assessed based on source and target factors. White, O'Connell and O'Mara (1994) defined three dimensions to consider for MT quality: adequacy (degree to which the content of the translation matches that of the source), fluency (degree to which the translation is perceived as natural sounding) and informativeness (or "modified comprehension", i.e. degree to which source information is preserved in the target language). However, due to the blurred boundaries between adequacy and informativeness, it has become common to consider only adequacy and fluency factors in MT evaluation.

4.2 Approaches to Translation Quality Assessment

Translation quality evaluation is paramount for several reasons. For human translation or post-edited texts, evaluation is essential to gain insights into the quality of the final texts and provide feedback to linguists. LSPs typically incorporate TQA into their vendor evaluation processes, which allows for easily finding the highest-performing translators. As far as MT is concerned, evaluation is equally important for keeping track of progress, comparing different models and identifying weaknesses. Due to these needs, several translation quality standards have been developed and currently co-exist in the field (Corpas Pastor, 2006). Furthermore, the types of translation errors vary based on the translation scenario, and specific errors are typically found in Human Translation (HT), MT and post-edited texts. For example, MT can be deceptively fluent and may hide adequacy errors behind the apparent naturalness of the output. Post-edited texts tend to show features of post-editese (further discussed in 4.5 on Quality of Post-Edited Texts) which is characterised notably by a high interference of the source language (in terms of sentence length and structure) and simplification (decreased lexical variety and density), as observed by Toral (2019). Despite this, several aspects of quality remain equally important in HT, MT and PE, therefore TQA methodologies seem to converge (Castilho et al., 2018). This section is dedicated to the presentation of the two main approaches to TQA: human evaluation, semi-automatic and automatic metrics. The challenges of quality assessment in the AI era are also discussed. It should be noted that the primary focus is on quality assessment of MT. However, various methods, particularly those belonging to the category of human evaluation, can also be applied in the context of HT evaluation.

4.2.1 Human Evaluation

Human evaluation involves relying on a human expert to provide an assessment of a given translation. Different methods are conceivable for this. The two most common are rating (i.e. assigning a score) and ranking (i.e. ordering various translation options based on their quality). Although human evaluation is highly beneficial and can bring valuable insights into translation quality, it also has several shortcomings. Inter-annotator agreement is one of the primary challenges: certain aspects of quality remain subjective, and evaluators may have different opinions. Moreover, human evaluation is not only time-consuming and expensive, but is also challenging to reuse (Castilho et al., 2018).

In an attempt to standardise methodologies and increase replicability and comparison between different studies, several error typologies became commonly used in the localisation industry in the late 1990s (Castilho et al., 2018). These typologies divide translation errors into distinct categories, and some also come with a scoring system, in which weights are assigned to each error based on its severity; these weights are then either summed up or deducted from an initial benchmark. Two of the most popular

typologies, the Dynamic Quality Framework (DQF; Görög, 2014) and the Multidimensional Quality Metrics (MQM; Lommel, Uszkoreit, and Burchardt, 2014), have both been developed simultaneously as part of the Quality Translation 21 project (QT21, Lommel et al., 2015), which aimed at providing a standard for translation quality evaluation. Both DQF and MQM comprise a typology of error types divided into dimensions (main categories) and error types or issues (subcategories). While these typologies are compatible (after a harmonisation process, DQF was defined as a subset of MQM), they operate at various levels of granularity. More specifically, MQM is more precise and includes a higher number of issues, whereas DQF has a more limited error coverage, as illustrated in Figures 23 and 24.

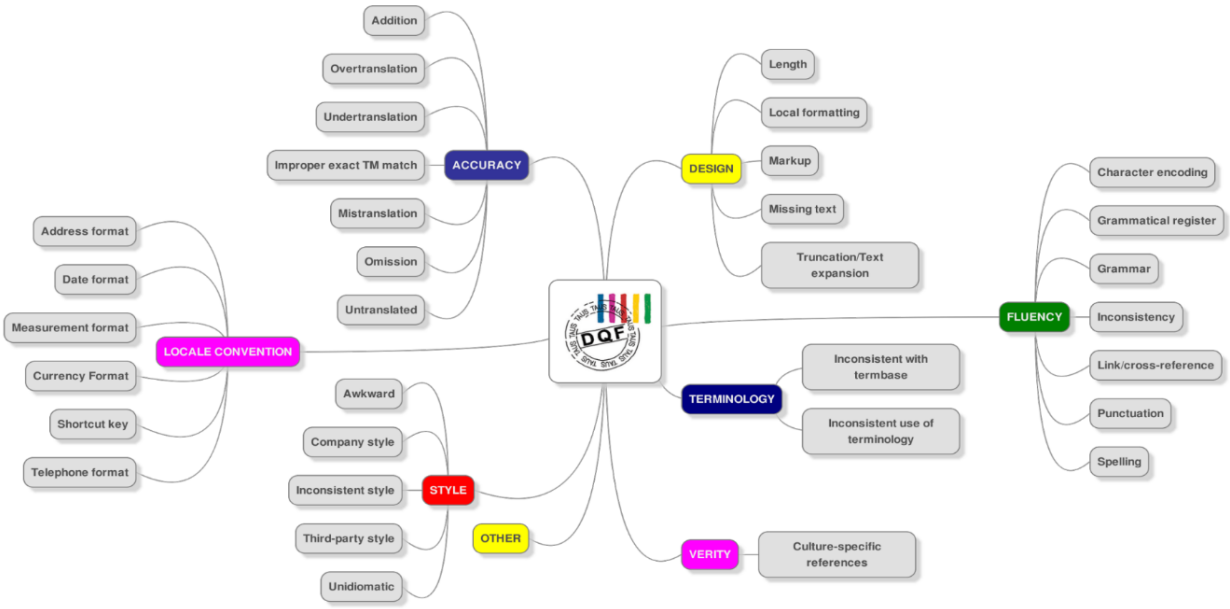


Figure 23: DQF error typology (TAUS, 2016).

Figure 24: Pre-harmonised MQM typology (Lommel et al., 2015, p.25).

Given that DQF and MQM are rather generic typologies, the specific requirements of certain fields resulted in the creation of domain-specific error taxonomies. As a matter of fact, the Localisation Industry Standards Association (LISA) and the Society of Automotive Engineers (SAE) developed their own models, targeted at the software localisation and automotive industries, respectively. The typology introduced by LISA (called the LISA QA Model) continued to be widely used even after its deprecation in 2011 (Castilho et al., 2018). As for J2450, the model designed by the SAE (SAE, 2016; typology in Table 20 below), it is still widely used in the industry.

	Category Name	Sub-Classification	Weight (serious/minor)
1	Wrong term	Serious	5/2
2	Syntactic error	Minor	4/2
3	Omission		4/2
4	Word structure or agreement error		4/2
5	Misspelling		3/1
6	Punctuation error		2/1
7	Miscellaneous error		3/1

Table 20: SAE J2450 typology (SAE, 2016).

As can be seen from the figures above, the error classifications differ, although several common errors can be identified across typologies. The same goes for the severity levels: MQM and DQF have four severity levels (neutral, minor, major and critical), which all come with an associated number of penalty points (0, 1, 5 and 10 for DQF, and 0, 1, 5 and 25 for MQM). SAE J2450, on the other hand, only has two levels (minor and serious), and the associated penalty points vary based on the error type. Regarding the threshold of tolerated errors, although some recommendations are available in the literature, both the number of penalty points and the threshold can be customised according to the needs of a specific translation context – hence the "dynamic" dimension. Today, conducting quality assessment is a generic activity for a number of LSPs, and such a process is made easy by the most common CAT tools (e.g. Trados Studio, memoQ, Matecat), which support error annotation in their interface and automatic score calculation. Several tools come with the aforementioned standard models and also allow for customisation.

In research circles, standardisation is less predominant, and typologies are frequently designed for specific studies. For example, Vilar et al. (2006) aimed to classify errors by comparing translations from an SMT system to a reference HT, working with several corpora extracted from European Parliament Plenary Sessions (English to Spanish and Spanish to English) and from broadcast news (Chinese to English). Even though automatic metrics (for more details, see 4.2.2 on Automatic Metrics) were used in this study, the main framework is based on a manual classification of errors. Thus, the authors have defined five main classes of errors ("missing words", "word order", "incorrect words", etc.), as illustrated in Figure 25.

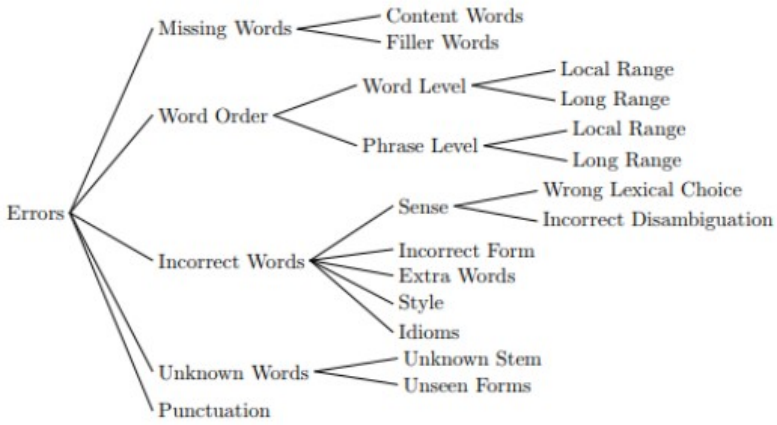


Figure 25: Error typology introduced by Vilar et al. (2006, p.699).

The classification used by Vilar et al. (2006) is based on a previous study by Font Llitjós et al. (2005), who proposed a framework for automatic refinement in rule-based transfer MT systems based on feedback from bilingual speakers. They created an MT error typology comprising five classes of errors (see Figure 26) for English-Spanish translation, with the aim of extending this process to resource-poor languages (Mapudungun and Quechua).

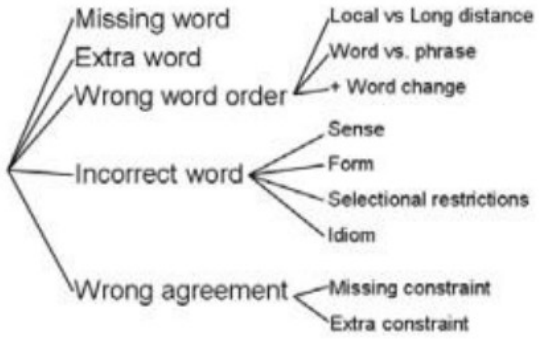


Figure 26: Error typology introduced by Font Llitjós et al. (2005, p.89).

Ahrenberg (2017) also proposed a framework for MT evaluation (Table 21). His study aimed to compare an HT of a press article (English to Swedish) with a translation produced by Google Translate. Ahrenberg provided a list of procedures that MT systems cannot implement (such as sentence splitting): such findings require a good command of translation theory and could not have been obtained with automatic metrics (see Table 21).

Type of edit	Description
Major edit	substantial edit of garbled output requiring close reading of the source text
Order edit	a word or phrase needs reordering
Word edit	a content word or phrase must be replaced to achieve accuracy
Form edit	a form word must be replaced or a content word changed morphologically
Char edit	change, addition or deletion of a single character incl. punctuation marks
Missing	a source word that should have been translated has not been

Table 21: Error typology introduced by Ahrenberg (2017), adapted.

Several ad hoc typologies created for specific projects are proposed regularly. For example, Costa et al. (2015, p.139) contended that common taxonomies are centred around English errors and proposed an extended one, centred on key areas of Linguistics, to account for the specificities of Romance languages, which have a richer morphology (Figure 27).

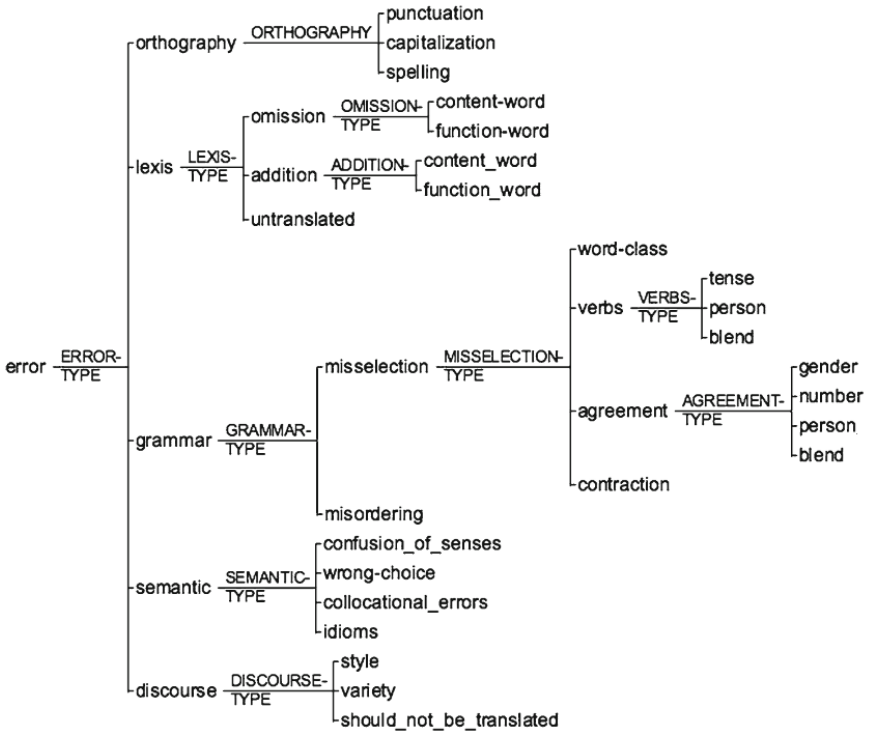


Figure 27: Error typology introduced by Costa et al. (2015, p.139).

Similarly, Hassan et al. (2018) adapted the typology introduced by Vilar et al. (2006), which resulted in focusing on nine error categories (namely, incorrect words, ungrammatical, missing words, named entity, word order, factoid, word repetition, collocation and unknow words). In addition, Popović (2018) emphasised the difficulty of establishing a unified methodology which can be applicable to various scenarios and suggested using generalised categories, which can be further specified at different levels, as described in Table 22.

Level 1	**Level 2**	**Level 3**
Lexis	Mistranslation	Terminology
	Addition	
	Omission	
	Untranslated	
	Should not be translated	
Morphology	Inflection	Tense, number, person
		Case, number, gender
	Derivation	Part of speech
		Verb aspect
	Composition	
Syntax	Word order	Range
	Phrase order	Range
Semantic	Multi-word expressions	
	Collocations	
	Disambiguation	
Orthography	Capitalisation	
	Punctuation	
	Spelling	
Too many errors		

Table 22: Error typology introduced by Popović (2018, p.141).

Despite the creation of numerous typologies, several error types are recurrent in both academic and industry resources. Candel-Mora and Borja-Tormo (2017) identified terminological errors, lexical ambiguity, syntax, concordance, omission and repetition, as well as punctuation errors, as being the most common error types in both academic and industry resources. However, the granularity varies significantly from one model to another. Certain differences can also be observed due to unique features of some languages: for example, verb inflection errors are only relevant in certain languages (Popović, 2018). Furthermore, while precise categories provide a more in-depth analysis, it should be borne in

mind that a high number of error types increases the cognitive load for the annotator, which may result in more annotation errors (ibid.).

Overall, the industry tends to have a preference for "one-size-fits-all" approaches, which is relevant when it comes to standardising processes, but may lack flexibility (Lommel, Uszkoreit, and Burchardt, 2014). On the other hand, more flexibility can be observed in research, which allows for a closer analysis of more diverse errors but entails difficulties in terms of replicability.

4.2.2 Automatic Metrics

Due to the weaknesses of human evaluation, notably in terms of time, cost, replicability and subjectivity, automatic metrics have been introduced, and some have been extensively used in MT research. Automatic metrics are mostly relevant in MT evaluation, as they typically rely on a reference translation and measure the difference between this reference and the MT output (the assumption being that the more similar the output is to the reference, the better). These metrics can be divided into three main types, based on the technique they are based on: error rate, precision, and machine learning.

Error-rate-based metrics are built around the Levenshtein distance (Levenshtein, 1966) as a way to measure the minimum number of editions separating a candidate translation from the reference. The Word Error Rate (WER, Niessen et al., 2000), which was initially conceptualised for automatic speech recognition, is based on substitutions, insertions and deletions. Several variations of this metric exist. For example, the Translation Edit Rate (TER, Snover et al., 2006) allows for shifting sequences of words, the multi-reference WER (mWER, Niessen et al., 2000) considers several references, and the Human-targeted TER (HTER, Snover et al., 2006) relies on real corrections produced by linguists.

In the following category, precision-based metrics, the BiLingual Evaluation Understudy (BLEU, Papineni et al., 2002) is the most popular. It calculates the geometric mean of the n-gram overlap (i.e. n-gram precision) multiplied by a brevity penalty. Some metrics are based on recall instead of precision, as is the case with the Metric for Evaluation of Translation with Explicit ORdering (METEOR, Lavie and Agarwal, 2007). Automatic metrics are also characterised by various degrees of flexibility in terms of matching criteria (e.g. METEOR accounts for paraphrases, stems and synonyms) and matching level (e.g. chrF 3 [Popović, 2015] operates at the character level).

More recently, and given the significant advances in this field, machine learning approaches have been explored in automatic metrics. The main difference between machine learning techniques and the aforementioned methods lies in the fact that traditional metrics rely on a distance calculation, whereas machine learning metrics are trainable. Relevant examples include metrics which are based on embeddings, such as Regressor Using Sentence Embeddings (RUSE, Shimanaka, Kajiwara and Komachi, 2018), a metric based on sentence embeddings capable of considering segment-level information. It is also relevant to mention that automatic evaluation is becoming easier to perform even

for researchers who are not familiar with programming, thanks to the introduction of the MAchine Translation Evaluation Online tool (MATEO, Vanroy, Tezcan and Macken, 2023), which allows for easily getting scores after uploading the data to be evaluated.

Although automatic metrics are fast and inexpensive (Papineni et al., 2002), they have been heavily criticised. Common reproaches typically revolve around the lack of interpretability (Vilar et al., 2006) and the low correlation with human judgement (Specia and Wilks, 2016; Koehn, 2010). BLEU remains the most widely used metric in MT evaluation (Lo et al., 2023), and while this is understandable in terms of comparability, this has several implications, as BLEU operates at the n-gram level and does not measure grammaticality, semantics or pragmatics (Ulitkin, 2013).

4.2.3 Semi-Automatic Assessment

Given the limitations of automatic metrics, semi-automatic techniques have been introduced in an attempt to achieve more realistic evaluation procedures. For example, the Post-Editing Tool (PET, Aziz, Castilho and Specia, 2012) was designed around the assumption that the translation quality increases when the editing effort decreases. This tool therefore keeps track of a series of metadata, including keystrokes and edit distance, to calculate the PE effort. Postediting Calculeffort (Candel-Mora and Borja-Tormo, 2017) was also developed based on a similar assumption, as it allows for estimating time and effort based on error annotation. Nevertheless, it is worth mentioning that such methods come with the shortcomings of human evaluation, notably regarding time and cost.

4.2.4 State-of-the-Art Methods: Quality Assessment in the AI Era

It appears evident from the discussion above that a plethora of evaluation methods is now available. However, given the unprecedented quality of NMT, one could question the relevance of most models to assess the outputs of more recent and sophisticated MT systems. BLEU, the most commonly used metric in MT evaluation, was created more than 20 years ago, when more obvious errors could be found in MT outputs. The same can be said of error typologies, like DQF and MQM, which are sometimes used for MT quality evaluation, but were also designed before NMT and LLM-based MT. This calls for reconsidering the methods used to evaluate current MT systems.

In that sense, some attempts for moving towards machine learning approaches can be observed, particularly at the WMT[41] shared tasks. For example, in the 2023 edition of the QE shared task, the aim was not only to predict the quality of an MT output, but also to identify critical errors. This can be achieved by training a model on annotated data. A successful example of such a model is CometKiwi

[41] WMT: Conference on Machine Translation (previously called “Workshop on Machine Translation”)

(Rei et al., 2023), which is based in particular on the MQM framework. This model is also the foundation of Unbabel Qi[42], a tool which provides automatic TQA scoring, including error identification.

Figure 28: Overview of a report generated by Unbabel Qi.

As can be seen in Figure 28, this tool provides more information compared to previous automatic methods, notably by specifying sentence-level quality ranges and by locating problematic fragments directly in the text. Nonetheless, without clearly labelling the error types found, this information is of limited usefulness. The model behind this tool was also trained using "pre-NMT" methods, which may result in failing to account for specific NMT errors. More generally, evaluation methods would also benefit from some adaptations. The first typologies, which can be applied to HT and MT, were designed a long time before the emergence of neural technology. Arguably, fine-tuning these typologies to reflect the advances in terms of MT quality would allow for yielding more accurate evaluations.

In addition, the use of MT has become ubiquitous today, particularly with social media, as MT is commonly embedded in popular platforms and is thus instantly available for users. However, User-Generated Content (UGC) differs from texts which are traditionally used for training MT engines, and is more likely to contain errors, which often results in a lower MT quality. One of the objectives of the WMT 2021 QE shared task was to detect critical MT errors (Specia et al., 2021). In this case, critical

[42] https://qi.unbabel.com/ (consulted: 30/04/2025)

errors were defined as “translations that deviate in meaning as compared to the source sentence in such a way that they are misleading and may carry health, safety, legal, reputation, religious or financial implications” (ibid., p.686). These deviations in meaning can take three forms (ibid.):

- Mistranslation: “critical content is translated incorrectly into a different meaning, or not translated (i.e. it remains in the source language) or translated into gibberish”
- Hallucination: “critical content that is not in the source is introduced in the translation, for example, profanity words are introduced that were not in the source”
- Deletion: “critical content that is in the source sentence is not present in the translation. For example, the source sentence may contain a negation or hateful word that is removed in the translation”.

Furthermore, a set of five critical error categories was defined as follows for this shared task:

- Deviation in toxicity (TOX)
- Deviation in health/safety risks (SAF)
- Deviation in named entities (NAM)
- Deviation in sentiment or negation (SEN)
- Deviation in numbers, time, units, or date (NUM)

Al Sharou and Specia (2022) extended this taxonomy by adding:

- Deviation in instructions (INS)
- Other critical meaning deviation (OTH)

The same authors also found that critical errors are commonly encountered in MT outputs, particularly in the case of UGC, which is more likely to contain source elements that may affect the MT model performance (e.g. misuse of punctuation, use of symbols and special characters, abbreviations, grammatical mistakes). While this classification provides relevant insights into the kind of errors that are particularly prone to leading to critical errors, the study conducted by Al Sharou and Specia (2022) also revealed challenges related to the annotation process, as annotators sometimes faced difficulties following the guidelines and found it difficult to focus solely on critical errors. Furthermore, deciding whether an error is critical is not always straightforward, and may require a certain extent of subjective judgement, which may in turn contribute to lowering the inter-annotator agreement. Similarly, Candel-Mora (2023) proposed to assess the automatic translation of user reviews from the perspective of sentiment analysis. This study introduced an evaluation model based on sentiment transfer composed of five categories (polarisation, intensity, misalignment, identification and cultural adaptation). The findings revealed that a quarter of the reviews examined presented sentiment transfer errors. This suggests that sentiment transfer is a weakness even for recent MT systems and that incorporating a sentiment identification mechanism could be a possible avenue for improving MT, notably in the case of UGC. Furthermore, this study also emphasises the need for adapting MT evaluation methods by including relevant genre-specific dimensions which are commonly overlooked in traditional typologies.

4.3 Impact of MT Quality on Post-Editing

It can be said from the above-mentioned evaluation metrics that the field of MT evaluation relies on various methodologies. Nevertheless, the typical metrics used to evaluate MT quality usually consider MT as a final product and therefore do not provide an accurate representation of the utility of the MT output as a starting point for PE (Denkowski and Lavie, 2012). BLEU is a salient example of this phenomenon, as BLEU scores cannot provide much insight into the extent to which a raw MT output is "usable" for PE. On the one hand, a low BLEU score only indicates a high divergence from the reference translation, which may occur even when the MT output is of high quality. On the other hand, a high BLEU score indicates that the MT is similar to the reference, but a candidate may be close to the reference and still be of low quality in terms of grammar or style, which are not encompassed by this metric.

Analysing the differences (i.e. edit rate) between a raw MT output and the corresponding post-edited version is a good starting point to study MT utility, but it is still insufficient. As mentioned in 3.2 (The Post-Editing Process), the final version does not necessarily convey all dimensions of the PE effort. While it can shed some light on the technical effort (HTER), it does not reflect the thinking process and intermediate steps (e.g. temporary versions of a translation which were not kept in the final version) which occurred before reaching the final translation. For this reason, when examining the PE effort, it is crucial to consider both product-related measures, such as HTER, in combination with process-related measures, such as the temporal and cognitive effort (Daems et al., 2017) to gain a holistic view of the PE effort.

The impact of MT quality on the different dimensions of PE effort thus appears to be a relevant object of study. Several experiments have examined this phenomenon. For example, pauses can be interpreted as an indicator of cognitive processing. O'Brien (2006) investigated whether pauses can serve as a cognitive effort indicator for PE based on the expected MT quality. The underlying hypothesis was that for source sentences which contained negative translatability indicators (i.e. linguistic features which are known to have a detrimental consequence on MT), the MT output would be of a lower quality, thus involving a higher cognitive effort to fix the errors, resulting in turn in more frequent and longer pauses. The outcomes of this study were not conclusive and suggested that although pauses may be useful as an effort indicator, it is essential to triangulate different effort measures to achieve a reliable analysis.

Doherty and O'Brien (2009) tested the use of eye tracking data as an MT evaluation method by examining the correlation between human evaluations of MT quality and different types of eye tracking indicators. The findings showed significantly high values in terms of average gaze time and fixation count for the sentences of low quality, which demonstrates that these values correlate well with human

judgement of MT quality. However, no significant difference was found regarding the average fixation duration.

Temnikova (2010) investigated the impact of using a controlled language on the cognitive effort. The outcomes of this study demonstrated that using a controlled language contributed to reducing the cognitive difficulty. While this is a relevant finding, what is more pertinent in terms of PE effort here is that a classification of MT errors based on the cognitive effort was established in order to conduct this analysis. This categorisation was based on the taxonomy of errors introduced by Vilar et al. (2006), and was adapted to reflect the cognitive effort, based on theoretical models around reading, working memory and error detection. The resulting classification consists of a list of error types ranked according to their estimated cognitive difficulty (from easiest to hardest), as shown in Table 23 below.

Morphological level	1. Correct word, incorrect form
Lexical level	2. Incorrect style synonym
	3. Incorrect word
/	4. Extra word
	5. Missing word
	6. Idiomatic expression
Syntactic level	7. Wrong punctuation
	8. Missing punctuation
	9. Word order at word level
	10. Word order at phrase level

Table 23: Classification of MT errors based on cognitive effort (adapted from Temnikova, 2010, p.3488).

Koponen (2012) studied the relationship between the technical effort (i.e. edit rate) and the cognitive effort (i.e. the perceived effort). To that end, the cognitive effort was assessed by post-editors and compared against the actual PE actions performed. A more fine-grained analysis also shed light on the impact of the types of MT errors on the post-editor's perception, based on the typology introduced by Temnikova (2010). Overall, the more salient factor proved to be sentence length, as long segments were perceived to involve more effort, even when only minor edits were performed. For sentences which were considered the easiest, the edits performed involved mostly changes in word form. In the case of sentences judged difficult to edit, reordering was found to be the most common change implemented. This suggests that changing a word form is a straightforward editing action, whereas reordering is more cognitively demanding. The observation that the effort depends on the type of editing actions was also

confirmed by Popović et al. (2014), who examined the temporal and cognitive efforts for predefined editing actions (namely, correcting word form, correcting word order, adding omission, deleting addition, and correcting lexical choice). According to the findings, removing additions did not have a strong effect on the effort, whereas a high temporal effort was associated with lexical and word order edits. The temporal effort was also high for long sentences.

Zaretskaya et al. (2016b) examined how different MT errors affect PE time and technical effort. Overall, the study demonstrated that the correlation between time and technical effort was weak. This does not constitute a surprising finding, as some errors may be challenging to fix, even when the solution does not involve a high number of editing actions. One specific error category (overly literal) resulted in high values, both in terms of time and technical effort. More diversity was observed for the other error types. Certain errors involved a long editing time (mistranslation, overly literal, spelling and entity errors), whereas other errors were fixed more rapidly (word form, function words, locale convention, omission, word order, and typography errors). As far as the technical effort is concerned, this indicator was the highest for overly literal errors. For other error types, the patterns were not that clear due to a low agreement in technical effort (i.e. high variety of values among participants). Similarly, Daems et al. (2017) examined the impact of different MT error types on the PE effort, considering both product and process effort indicators. The product effort indicator used was HTER. Regarding process effort, the indicators were obtained via eye tracking and keystroke logging, and comprised fixation duration, number of fixations, pause ratio and average pause ratio, PE duration and production units. A correlation was found between a decrease in MT quality and an increase in several effort indicators (namely, duration, number of fixation and production units, HTER score) as well as a decrease in the average pause ratio. However, the pause ratio and the average fixation duration did not seem to be affected by MT quality. This study also revealed that each process effort indicator was sensitive to certain error types. For instance, coherence issues were found to affect duration, whereas other meaning shifts appeared to influence fixation duration. The error types which were found to affect more than one indicator were coherence issues, other meaning shifts, grammar and structural issues. Therefore, these errors can serve as a basis for predicting the PE effort. This analysis also provides guidance regarding which dimensions of MT quality should be regarded as a priority for improvements based on the type of effort which one wishes to focus on. To exemplify this recommendation, the authors suggested that fixing coherence issues (which affect duration) should be a priority for LSPs which are concerned with speed. A study by Vardaro, Schaeffer and Hansen-Schirra (2019) focused on how errors are recognised and corrected in PE. The experiments conducted involved an annotation process of errors found in NMT based on the MQM typology. Then, the processing of three specific error types (mistranslations, terminology and stylistic errors) was examined using eye tracking and key logging. While the authors hypothesised that the behaviour would vary based on the error type in terms of eye movements, no

specific pattern was found. Instead, similar error recognition patterns were observed independently of the error category.

Consequently, it seems undeniable that the quality of MT outputs affects the PE effort. Nevertheless, the extent of this phenomenon appears to be difficult to measure. While the aforementioned studies shed some light on the impact of MT errors on the PE process, several limitations should be acknowledged. First, the fact that each study uses a different methodology entails negative repercussions in terms of replicability. In each experiment, the measures are different, and one can find several interpretations of some dimensions of PE effort. For example, in Koponen (2012), the cognitive effort is a perception rating, while in Popović et al. (2014), this indicator is more related to quality. Different measures are also used depending on the study: Koponen (2012) relied on both temporal and cognitive indicators, Zaretskaya et al. (2016b) looked into time and PE effort (provided by Matecat), while Daems et al. (2017) used a combination of HTER, eye tracking and keystroke logging. Regarding error annotation, Doherty and O'Brien (2009) used a simple categorisation (excellent vs bad), Koponen (2012) employed the typology proposed by Temnikova (2010), and the MQM was used in two cases (Zaretskaya et al., 2016b; Vardaro, Schaeffer and Hansen-Schirra, 2019). Moreover, error annotation can be influenced by individual perceptions around quality, which leads to variations from one person to another (i.e. low inter-annotator agreement). Individual preferences and behaviours may also influence the data. For instance, Zaretskaya et al. (2016b) found low agreements in terms of temporal and technical effort. In addition, in several cases, participants were translation students (Doherty and O'Brien, 2009; Zaretskaya et al., 2016b). Such experiments are likely to yield different results compared to those which involved professional translators (Koponen, 2012; Vardaro, Schaeffer and Hansen-Schirra, 2019). Another important limitation lies in the fact that most of these studies, with the exception of one (Vardaro, Schaeffer and Hansen-Schirra, 2019), did not use NMT outputs, but rather older MT technology. While the relationship between the error types and the PE effort are likely to remain consistent in the current models, further examination with more recent systems would appear relevant, as the quality gains in NMT may have an impact on the PE task.

4.4 Towards New MT Quality Evaluation Methods

The notable advancements in training methodologies and model architectures in the last decade resulted in the advent of NMT. The substantial improvements in translation quality led to a wide adoption of this technology. Today, translation models continue to be enhanced. For example, leveraging LLM augmentation and incorporating personalisation techniques via the use of linguistic assets have become common practises, aiming to tailor translations to specific domains or customer preferences.

Despite these advancements, the methodologies for evaluating translation quality have not evolved at the same pace. Traditional evaluation methods, such as the DQF or MQM, remain common standards

in practice. Even the latest ISO 5060 standard, on "Translation services — Evaluation of translation output — General guidance" (ISO, 2024) relies on an error typology similar to MQM, without considering the specificities of NMT outputs.

Automatic evaluation metrics, such as BLEU (Papineni et al., 2002), which were sufficient in the preliminary stages of MT research, have proven inadequate in capturing the nuances of NMT (Lee et al., 2023). More recently, embedding-based metrics have been introduced as alternatives. However, as with previous automatic metrics, they also fail to provide comprehensive insights into NMT performance because of their lack of explainability (Leiter et al., 2024), though this situation seems to evolve given the recent use of error annotations to train QE models (Rei et al., 2023).

These observations emphasise an urgent need for a comprehensive framework tailored specifically for the evaluation of NMT output. Such a framework should address the unique characteristics and challenges posed by NMT systems while ensuring transparency and interpretability in the evaluation process.

4.4.1 Quality of State-of-the-Art MT Systems

Adapting current quality models requires identifying NMT-specific phenomena. While several dimensions remain equally relevant (e.g. terminology, style, etc.), the latest MT models tend to make errors which are considerably less often observed in older systems (like SMT) or in HT.

The differences in quality between SMT and NMT outputs can be explained by their distinct architectures and training processes. NMT systems exhibit a robust generalisation capability as they encode units as numeric vectors representing concepts, as opposed to the string-based encoding process occurring in SMT. This capability also enables NMT systems to identify long-distance phenomena (Toral and Sánchez-Cartagena, 2017). Consequently, NMT outputs demonstrate a higher inter-system variability, meaning that translations produced by different pairs of NMT systems exhibit greater divergence than those generated by SMT systems (Toral and Sánchez-Cartagena, 2017). Moreover, the learning curve of NMT systems is highly dependent on the amount of training data, and while NMT models achieve a high performance in high-resource scenarios, their outputs are generally of low quality when the size of the training dataset is limited (Koehn and Knowles, 2017). These key differences in model architecture and training process also make NMT outputs less interpretable compared to SMT (Koehn and Knowles, 2017).

These differences have led to higher quality levels, resulting in turn in reduced PE effort. Bentivogli et al. (2016) demonstrated that NMT outperforms other approaches, based on BLEU and TER scores. Speerstra (2018) observed that NMT tends to be more workable compared to SMT, based on linguists' judgement – a finding that justified the adoption of NMT over SMT at Infor. Moreover, Mutal et al.

(2019) found that NMT outputs require fewer edits compared to SMT. In this study, the most frequent types of edits for both systems were substitutions, deletions, insertions, and word order adjustments. However, NMT exhibited a higher proportion of substitutions and a lower proportion of deletions.

The differences in quality have been the object of numerous studies. Based on findings reported in the literature, the remainder of this section summarises the strengths of NMT, by listing concrete improvements achieved over SMT.

- Fluency: Toral and Sánchez-Cartagena (2017) found that NMT outputs exhibit greater fluency. Similarly, Van Brussel, Tezcan, and Macken (2018) observed substantial improvements in fluency for English-Dutch translations with NMT, including all categories with the exception of lexicon. Stasimioti et al. (2020) also reported higher fluency scores for NMT and tailored-NMT outputs compared to SMT.
- Lexical errors: Bentivogli et al. (2016) found that NMT systems produce at least 17% fewer lexical errors compared to SMT.
- Grammar and syntax: Castilho et al. (2017) noted a significant reduction in sentence structure errors in NMT translations compared to SMT, particularly in patent translation. This finding is in line with the study of Stasimioti et al. (2020), in which fewer grammatical errors were observed in NMT outputs compared to SMT.
- Word order: Bentivogli et al. (2016) reported substantially fewer word order errors in NMT outputs. Castilho et al. (2017) observed fewer word order errors in NMT translations for the Massive Open Online Course (MOOC) domain. Toral and Sánchez-Cartagena (2017) also highlighted that while NMT introduces more changes in word order than SMT systems, it performs better in terms of reordering in all tested language directions.
- Morphology: Studies by Bentivogli et al. (2016) and Castilho et al. (2017) indicate that NMT generates translations with a higher morphological correctness compared to other systems. Toral and Sánchez-Cartagena (2017) also emphasised NMT's proficiency in producing correct inflections.
- PE effort: Bentivogli et al. (2016) observed a significant reduction in PE effort in the case of NMT.

Conversely, despite the overall quality enhancements, certain errors appear to occur more often in the case of NMT.

- Accuracy: NMT systems may be prone to accuracy issues, particularly in handling omissions, additions, mistranslations, and instances of "Do Not Translate" (DNT) terminology. For example, Castilho et al. (2017) found inconsistencies in adequacy scores for NMT output, with higher mean scores observed for German SMT. In the case of patent translation, Castilho et al. (2017) noted a higher prevalence of omission errors in NMT outputs compared to SMT. In the

MOOC domain, SMT systems exhibited fewer omission, addition, or mistranslation errors for specific language pairs compared to NMT outputs. Van Brussel, Tezcan, and Macken (2018) found that SMT outputs contained the fewest omissions, while NMT exhibited fewer omission errors compared to SMT. They also observed a higher ratio of omitted words per omission error in NMT, thus suggesting a higher impact per error. Additionally, NMT systems were found to drop more content words, which exacerbates this issue. What is more, omission errors in NMT are less evident to spot since they may not be anticipated by fluency errors, which has a detrimental impact on the PE effort. Ustaszewski (2019) also highlighted that omissions and additions of content are particularly difficult to identify given the high fluency of NMT outputs. Moreover, in a study by Van Brussel, Tezcan, and Macken (2018), although NMT tended to achieve higher accuracy scores, a closer examination revealed that SMT handled DNT better than NMT.

- Morphology: NMT systems may encounter challenges related to inflection errors, particularly in specific language pairs and domains. Castilho et al. (2017) indeed observed a higher proportion of inflectional morphology errors in NMT outputs for German patent translation.
- Domain mismatch: NMT systems tend to exhibit poorer performance in out-of-domain scenarios, as noted by Koehn and Knowles (2017). However, it should be noted that this study was performed at the beginning of NMT, and as anticipated by the authors, several domain adaptation techniques have been developed since.
- Punctuation: Stasimioti et al. (2020) observed that NMT outputs, particularly tailored NMT, may contain more punctuation errors compared to SMT. However, this finding was attributed to the omission of certain punctuation marks (em dashes).
- Lexicon: NMT systems may be prone to lexical choice errors. Van Brussel, Tezcan, and Macken (2018) reported that while NMT achieved a higher performance for function words, it made more lexical choice errors for content words.
- PE effort and human perception: human evaluators may prefer SMT output over NMT in certain cases. Castilho et al. (2017) conducted human evaluations across different domains, revealing nuanced preferences for NMT and SMT systems. For e-commerce product listings, NMT models were preferred over SMT. On the contrary, for patent translation, SMT was ranked best more frequently than NMT. For the MOOC domain, however, NMT was perceived as more fluent than SMT. Mutal et al. (2019) highlighted challenges faced by translators when working with NMT. Translators reported greater difficulty in determining whether a sentence required editing when working with NMT compared to SMT. Additionally, a lower agreement was found among translators regarding the corrections needed in NMT outputs, which indicates that errors may be easier to spot in the case of SMT.

In addition to specific NMT errors, other factors should be acknowledged.

- Sentence length: NMT often faces difficulties when it comes to maintaining translation quality for very long sentences. Castilho et al. (2017) found that SMT systems outperformed NMT systems in translating short sentences. Toral and Sánchez-Cartagena (2017) found that while NMT generally outperformed SMT for sentences of average length, PBMT showed better performance for very long sentences. This clashes with the findings of Bentivogli et al. (2016), yet the difference can be explained by the shorter average sentence length in this study due to the nature of the data (TED talks). Likewise, Koehn and Knowles (2017) noted lower translation quality for very long sentences with NMT. However, this observation could be nuanced, as Van Brussel, Tezcan, and Macken (2018) reported that the expected degradation of NMT performance in long sentences may no longer hold true.
- Hallucination: NMT outputs are also more likely to contain hallucination errors, characterised by the generation of erroneous or fictional content, particularly when noise is present in the input sequence (Lee et al., 2018).
- Sentiment deviation: after investigating MT outputs from the perspective of sentiment analysis, Candel-Mora (2023) found that a quarter of the reviews analysed presented sentiment transfer errors, which is concerning for online platforms where MT is used as a final product (i.e. without PE).
- Lack of a consensus: It is essential to point out that while several studies corroborate other findings reported in the literature, some results seem to diverge for certain MT errors. For example, Bentivogli et al. (2016), Castilho et al. (2017) and Toral and Sánchez-Cartagena (2017) found that NMT performed better in terms of morphology. However, in the same study, Castilho et al. (2017) reported that NMT did not perform well for a specific language (German) and use case (patent translation). The same can be said of the claims around PE effort, since Bentivogli et al. (2016) noted a significant effort reduction for NMT, whereas in the study by Mutal et al. (2019), translators found it more difficult to decide whether a sentence required to be edited in NMT outputs.

It also be acknowledged here that with the advent of LLMs, these models could soon replace NMT. If this occurs, further studies will be necessary to investigate output quality differences between NMT and LLM-based MT.

4.4.2 Post-Editing-Based MT Quality Evaluation

Existing evaluation models, primarily designed for HT assessment, such as DQF, are often employed for assessing MT outputs. However, the majority of commonly used evaluation models was developed

before the widespread adoption of NMT systems. Consequently, they may fail to adequately capture and address NMT-specific issues and nuances in translation quality.

While a plethora of automatic metrics has been developed, it should be pointed out that certain phenomena cannot be accurately captured by these metrics and thus call for human evaluation. For example, in terms of accuracy, Ustaszewski (2019) tested cross-language aligned word embeddings as a way to account for semantic similarity (measured with cosine similarity), but the results obtained in this study were not encouraging, as the proposed model was not successful in detecting missing content and did not correlate with human judgement, which highlights the importance of human evaluation.

According to Ovchinnikova (2020), errors which are not recognisable by automatic evaluation systems include:

- Transferring the author's opinion (e.g. simplifying or complicating the narrative)
- Strategy of translation (e.g. inaccurate transfer of the objective behind the message)
- Pragmatics and discourse (e.g. errors in coherence or genre)
- Text-external errors in information processing (errors in metric systems or locations)

Ovchinnikova (2020) also distinguished three categories of errors, namely:

- Fluency errors ("which damage the TT readability")
- Accuracy errors ("which misrepresent the content")
- Functional errors ("which distort the objective of the ST translation and obstruct the perception of the TT in the target culture")

According to her, the error typologies commonly used in the industry do not include functional errors because MT developers have not been able to solve those.

When determining the appropriate evaluation model for MT, considering the specific context and usage scenario of the output is paramount. Therefore, it is suggested here to apply different criteria and methodologies depending on whether the MT output serves as the final product or as an intermediate step before PE. The present work focuses on the quality of MT from the perspective of PE. In scenarios where the MT output serves as an intermediate step before PE, the evaluation model should be based on the MT errors that have a direct impact on the PE process. Here, the evaluation criteria may emphasise aspects such as fluency, adequacy, consistency, and adherence to specific style guides or terminology.

Furthermore, MT models are constantly evolving: LLM-based MT may be prone to different types of errors, and so may future types of models. Therefore, establishing a framework that is not entirely dependent on the specificities of each technology appears as a more sustainable solution. In the context of evaluating MT for PE, assessing MT quality based on the PE effort thus seems a relevant approach.

As discussed above, a significant challenge in NMT lies in the difficulty of identifying errors, especially when adequacy errors are masked by high fluency. The participants in the study of Castilho et al. (2017)

reported finding NMT errors more challenging to identify, while SMT output errors, including fluency issues, were detected faster. Moreover, adequacy errors, as opposed to fluency errors, are particularly challenging. Van Brussel et al. (2018) and Ustaszewski (2019) indeed noted the deficiencies of automatic metrics when it comes to capturing adequacy errors. These errors involve inaccuracies in conveying the intended meaning or content of the source text. In addition, they are often more subtle and context-dependent, which makes them harder to detect and address, even for human subjects.
Error categorisation could be further tailored to reflect the types of errors commonly encountered when working with NMT outputs. An effective approach to designing an evaluation model here could be to weight errors based on the PE effort they involve. Following this logic, errors that require a higher PE effort should incur a higher penalty. Given the information presented earlier, adequacy errors, which are more challenging to manage due to the high fluency and involve a higher PE effort, could be assigned a higher severity in the evaluation model. Other MT errors appear to involve a high PE effort. In that regard, the classification provided by Temnikova (2010, p.3488; Table 23) ranked MT errors based on the cognitive effort (easiest to hardest).

The PE effort associated with different MT errors has already been explored; therefore, the remainder of this section revolves around a classification of MT errors based on their associated PE effort according to the consulted literature. MT errors involving a high PE effort comprise:

- Reordering: in the study of Koponen (2012), in cases where sentences were perceived as difficult to edit, reordering was frequently identified as the most common change implemented. In Popović et al. (2014), high temporal effort was also associated with word order edits. However, it should be noted that these findings are not aligned with the study by Zaretskaya et al. (2016) where word order errors were fixed rapidly.
- Lexical edits: Popović et al. (2014) highlighted that high temporal effort was often related to lexical edits.
- Source language interference: Zaretskaya et al. (2016) found that one specific error category, overly literal translations, resulted in high values of both time and technical effort.
- Mistranslation: Zaretskaya et al. (2016) observed that mistranslations led to long editing times. However, this clashes with another finding in the same paper whereby omissions were fixed rapidly. Additionally, Daems et al. (2017) indicated that meaning shifts also influenced the fixation duration.
- Grammatical and structural issues: Daems et al. (2017) noted that grammar and structural issues impacted more than one indicator of PE effort.
- Document level and coherence issues: Daems et al. (2017) found that coherence issues affected duration. This is particularly relevant, given that document-level phenomena (also encompassing cohesion, cross-references like anaphora, etc.) constitute a common challenge for

NMT – the document level was the main argument of Läubli, Sennrich, and Volk (2018) against the claims of MT achieving human parity (Hassan et al., 2018).

- Sentence length: Koponen (2012) noted that segments perceived as involving more effort, even with minor edits only, were often characterised by longer sentences. Similarly, Popović et al. (2014) found high temporal effort associated with long sentences.

Conversely, MT errors involving a low PE effort comprise:

- Word form edits: Koponen (2012) observed that for sentences perceived as the easiest to edit, the majority of edits involved changes in word form. Similarly, Zaretskaya et al. (2016) found that word form errors were fixed rapidly.
- Removing: Popović et al. (2014) noted that removing additions from the MT output did not significantly impact effort.
- Function words: Zaretskaya et al. (2016) observed that errors related to function words were fixed rapidly, suggesting that these errors posed minimal difficulty.
- Locale conventions: errors related to locale conventions were also fixed rapidly (Zaretskaya et al., 2016).
- Typography: errors in typography were also rectified quickly (Zaretskaya et al., 2016).

4.5 Quality of Post-Edited Texts

Beyond the implications in terms of PE effort, working with MT also entails consequences in terms of quality of the final product. Certain textual features of texts translated in CAT tools, in particular regarding segmentation, have already been discussed in 3.3.3 on Fragmentation & Decontextualisation. Since PE can occur in such environments, post-edited texts can also be affected by segmentation, which can affect cohesion and coherence. In addition to this, examining the features of post-edited texts constitutes another pertinent research avenue related to translation quality. According to Baker (1993), translated texts exhibit certain features (translation universals) which distinguish them from original texts. These features comprise simplification (replacing complex elements by simpler ones), normalisation (standardisation) and explicitation (making implicit information more explicit). Moreover, translated texts are also prone to being affected by interference from the source language (Toury, 2012). When it is particularly noticeable that a text is the result of a translation, this phenomenon is called translationese – a term coined in 1986 by Gellerstam (Daems, De Clercq and Macken, 2017). Similarly, MT outputs can be characterised by several features, and sometimes lead to an equivalent phenomenon, machine-translationese, which is characterised by some traces of translationese but does not completely emulate features of HT (Bizzoni et al., 2020). A study conducted by Vanmassenhove, Shterionov and Way (2019) demonstrated a loss of lexical diversity and richness in MT outputs. This study also

examined algorithmic bias and found that the most frequent words in the input (training data) tended to occur more frequently in the output, whereas less frequent words in the input were more prone to being "lost in translation". The results reported by Recski and Kádár (2023) go in line with these observations, as they showed that MT tended to generate lexically simpler texts, compared with human translations. Following the same logic, post-edited texts tend to exhibit similar features, which results in post-editese. Toral (2019) examined the presence of translation universals (simplification, normalisation and interference) in post-edited texts compared with HT. The measures used in this study included lexical variety (type-token ratio), lexical density (content words-total number of words ratio), length ratio (length difference between the source and target texts) and part-of-speech (POS) sequences. The results showed differences for each of these metrics. Simplification was observed in terms of lexical variety and density, which were lower in the case of post-edited texts (which also suggests that simplification features are exacerbated in post-editese compared to translationese). Interference was found in the analysis of sentence length, which proved to remain closer to the source texts in post-edited versions (this feature is also characteristic of normalisation) and POS sequences, which also remained more similar to the source in the case of PE. As an explanation for these results, the author suggested that post-editors were primed by the MT output, which is why MT features remained in post-edited texts.

However, these phenomena are frequently subject to debate. The existence of translation universals has indeed been questioned, both in terms of the type of universals and the extent to which universal features appear in translated texts (Corpas Pastor et al., 2008). The same goes for post-editese. Daems, De Clercq and Macken (2017), for example, attempted to find distinctive features of post-editese via both human perception and computational analysis, but their experiments did not support any proof of the existence of this phenomenon. Similarly, Volkart and Bouillon (2023) observed four candidate features of post-editese (lexical richness, lexical density, length ratio and sentence structure similarity between source and target) and were able to find evidence of post-editese only for one indicator (sentence length), thus suggesting that there are no universal features characterising post-edited texts. Moreover, post-editese can also become less apparent as MT quality increases. Farrell (2023) replicated an experiment originally conducted in 2018 and found that the lexical impoverishment typically observed in PE was lower when using a more recent version of DeepL.

4.6 Summary & Outcomes of this Section

This section provided insights into translation quality from both academic and industrial perspectives, and discussed several evaluation methodologies as well as quality analyses of HT, MT and PE.

Defining translation quality is an arduous task, but all suggested definitions appear to revolve around the concepts of adequacy (i.e. source factors) and fluency (i.e. target factors).

In the language services industry, standardisation and customer satisfaction are paramount, which explains the prevalence of "one-size-fits-all" approaches and the strong focus on compliance with specifications. In academic settings, the definition of quality may stretch to accommodate specific use cases; hence, new TQA methodologies are regularly introduced, including different typologies of translation errors. Furthermore, TQA can be conducted as a human evaluation or can rely on automatic or semi-automatic metrics.

While each approach comes with its strengths and weaknesses, most were introduced before state-of-the-art MT systems, which tend to output high-quality results and therefore require more fine-grained analyses. This situation calls for the elaboration of adapted TQA processes able to encompass errors produced by the latest MT models. This highlights the need for new evaluation models to assess MT quality. In that sense, this section posited that establishing a PE effort-based approach could be a way forward and would have the advantage of being adaptable to future MT technology, since it would be based on the PE experience. Nevertheless, designing such a model calls for further research on the difficulty associated with certain types of PE actions.

The use of technology, comprising both CAT tools and MT, also proved to affect translation quality. While certain features can be found in translated texts, certain characteristics seem to be typical of MT outputs and of post-edited texts – although this phenomenon remains subject to debate. One could therefore question whether technological assistance should endeavour to be completely seamless and allow linguists to produce texts without traces of post-editese, or if, on the contrary, machine translationese and post-editese represent inevitable evolutions of human language and should thus be accepted as such.

III. METHODOLOGY

The preceding chapters have highlighted a significant shift in the translation landscape, driven by advancements in technology and the growing integration of MT into professional workflows. This new paradigm, characterised by augmented translation tools and PE practises, demands a new evaluation approach tailored to the evolving needs of users. While the theoretical background revealed several methodologies for assessing translation technology, the absence of a consensus on how to effectively evaluate these tools in practice calls for research on that topic.

Building on the insights from the theoretical framework, this chapter transitions from exploration to application, proposing a methodology designed to address the gaps in current evaluation practises. The framework is structured around three core pillars – productivity, UX and quality – and aims to provide a comprehensive, user-centred approach to assessing PE environments.

This chapter begins by outlining the motivation and rationale for this research, followed by a presentation of the research questions, the objectives, and the research context. It then introduces the benchmarking methodology and presents the design of the framework.

5. Rationale

To understand the necessity of this new evaluation approach, it appears necessary to summarise key learnings of the theoretical background.

A. The Translator's Workstation

From their conceptualisation in the 1980s to more recent developments, translation tools have evolved significantly. While the first proposals, especially from Martin Kay and Alan Melby, were centred around empowering human translators, MT has become the standard today, and PE is the reality for many linguists. As technology continues to evolve constantly, new additions to translation tools have emerged, including IMT, multi-modality, QE and AI assistants.

B. Assessing the Post-Editing User Experience in Translation Tools

Given these advancements, user studies are paramount. However, despite several attempts at assessing the usability of translation tools, including applying UX methodologies, no consensus exists on a standard framework to assess translation tools, especially for PE. It is all the more concerning knowing that linguists tend to feel excluded from the development of the tools they are using.

C. From Translation to Post-Editing: Understanding the Post-Editing Process

Beyond UX, defining the PE process itself is also crucial to understanding the needs and expectations of linguists as well as the challenges they are faced with. Although the translation competence is necessary to perform PE, it should be noted that PE requires a different skillset. Unlike translation from scratch, in PE, the effort is concentrated on spotting errors and deciding which corrections are appropriate, which results in a different thinking process. These differences also suggest that CAT tools, in which PE occurs, could be better adapted to cater to PE needs.

D. Quality in Translation

Apart from the tools in which PE is performed, the impact of MT on translation quality should also be raised. Translation quality has tended to be defined as an equilibrium between adequacy and fluency, both for human and automatic translation. A plethora of assessment methodologies has been designed, including a wide range of error typologies as well as automatic metrics. However, the shift to NMT entailed drastic quality improvements which are not reflected in the standard evaluation models.

In light of the above, it is evident that a paradigm shift has occurred and continues to occur in translation technology today. In this rapidly evolving technological landscape, models and tools undergo frequent changes. The use of MT and CAT has become standard practice in the industry, and customers have become increasingly informed and exigent. In this context, the use of state-of-the-art technology is essential to thrive and remain ahead of the competition, which often leads to fast adoption of recent developments. As a result, linguists have no choice but to adapt to new ways of working.

Nevertheless, a paradigm shift cannot be successful without a solid assessment. New models and features should not be implemented only because they are technically feasible, but because they bring substantial value to the translation process. For that reason, evaluation frameworks are necessary to identify the strengths and weaknesses of existing technology and provide directions for future implementations. However, assessment practises do not seem to have followed the pace of development, which resulted in the lack of a standard methodology for the evaluation of translation tools, particularly when PE is involved. Designing a benchmarking framework, in order to establish a continuous evaluation process, therefore appears as a pertinent avenue (for further details on benchmarking, see 8.1 Introduction to Benchmarking).

6. Primary Research Questions & Objectives

This section outlines the research questions and objectives that guide the development and application of a benchmarking framework for PE environments. As the translation industry continues to evolve, driven by advancements in MT and the integration of AI assistance, the need for a systematic approach to evaluate and improve the PE environment continues to grow.

In this context, the present research seeks to address this need by answering the following research questions:

- How can the PE environment be benchmarked?
    - What elements should be included when benchmarking translation tools for PE?
- How can translation technology, and more specifically, the PE environment, be improved to provide an enhanced experience while meeting key user needs?

The primary objectives of this research are as follows:

- Design a benchmarking method for the PE environment.
    - Understand the PE process and its unique challenges.
    - Identify user requirements and expectations through feedback and analysis.
    - Select the main pillars that form the core of the framework.
- Apply the benchmarking framework in an industrial setting.
    - Evaluate a proprietary CAT tool (Smart Editor) in a PE context to assess its strengths and weaknesses in a real-world scenario, focusing on the selected pillars.
- Apply the benchmarking framework to evaluate emerging technologies.
    - Explore the impact of recent developments, such as AI assistance, on the PE environment.
- Identify opportunities for optimisation.
    - Based on the applications of the framework, pinpoint areas for improvement in PE environments, with the goal of enhancing efficiency, user satisfaction and quality.

This research is expected to make several key contributions:

- A novel benchmarking framework tailored to the unique demands of PE environments, integrating productivity, UX and quality.
- Practical insights from the application of the framework in both industrial and experimental settings, providing actionable recommendations for tool developers.
- A user-centred approach to evaluating translation technology, ensuring that tools are designed to meet the needs of post-editors.

- A foundation for future research on the optimisation of PE environments, which could be extended to further use cases in the context of emerging technologies.

By addressing these research questions and objectives, the present research aims to advance the understanding of PE needs and contribute to the development of tools and practises that better support post-editors in their work.

7. Research Context

The research described herein is conducted in the context of an Industrial PhD, which is concretised by a collaboration agreement between the Universitat Politècnica de València and LanguageWire. This section provides an introduction to the company, followed by a description of its internal structure during the course of pursuing the present research.

7.1 LanguageWire

LanguageWire is a Language Service and Technology Provider founded in 2000 with headquarters in Copenhagen, Denmark. The company has a global presence, with a focus on specific regions such as the Nordics, DACH[43], UK and US. Its largest office is located in Valencia, Spain, and gathers an important part of the LanguageWire's operational and tech departments. LanguageWire's mission is to make global communications smarter and more efficient by providing access to cutting-edge language technology and expert human resources. The company has positioned itself as an innovative leader in the industry, offering a comprehensive language management ecosystem that combines AI technology with human expertise. The company's strategy focuses on adapting solutions to customer needs, streamlining workflows, and ensuring data security within a protected infrastructure so that customer data never leaves the LanguageWire ecosystem.

LanguageWire offers a wide range of multilingual services, such as translation and localisation of both textual and multimedia content, editing and proofreading, desktop publishing and content creation. It also offers various proprietary products, including in particular the LanguageWire Platform (used to manage projects), Smart Editor (a CAT tool), linguistic assets management tools (Term Base and Translation Memory Management Systems), LanguageWire Translate (the company's MT product) and LanguageWire Generate (a Generative AI content creation tool).

With over 400 employees and a network of more than 7,000 industry-specific language experts, LanguageWire has established itself as a significant player in the language services industry. The company's commitment to innovation, security, and quality is evident through its numerous certifications, including ISO 27001, ISO 17100, ISO 9001, ISO 18587, ISO 13485, and TISAX[44]. In addition, LanguageWire reached the top position for post-edited MT in the 2022 CSA report[45].

---

[43] DACH is an acronym that represents three German-speaking countries: Germany (D), Austria (A) and Switzerland (CH).

[44] TISAX: Trusted Information Security Assessment Exchange. For the complete list of LanguageWire's certifications, see https://www.languagewire.com/en/technology/information-security (consulted: 30/04/2025)

[45] Source: https://www.languagewire.com/en/blog/languagewire-named-top-in-pemt (consulted: 30/04/2025)

7.2 Internal Structure

The research objectives correspond, at least partly, to activities which fall within the scope of the Linguistic Engineer position held by the author at LanguageWire from February 2022 to June 2024. This role involved guiding the engineering team in developing and enhancing Smart Editor, the company's proprietary CAT tool. Among other tasks, it entailed conducting comparative studies of translation tools, collecting feedback from linguists and analysing suggestions for improvements and new functionalities. The primary objective was therefore to help build a product line which is continuously improving based on user needs. Key responsibilities, as stated in the work contract, included the following:

- Helping the team understand the detailed and subtle requirements of professional translators regarding language tools.
- Engaging with the translator community with a view to improve language tools.
- Guiding the product management with the right scoping and priorities during feature discovery and planning.
- Collaborating with technical writers to create interesting and helpful tutorials and manuals for language tools.

The Linguistic Engineer thus constitutes a strategic position, involving strong contact between users and builders of the tools, as illustrated in Figure 29.

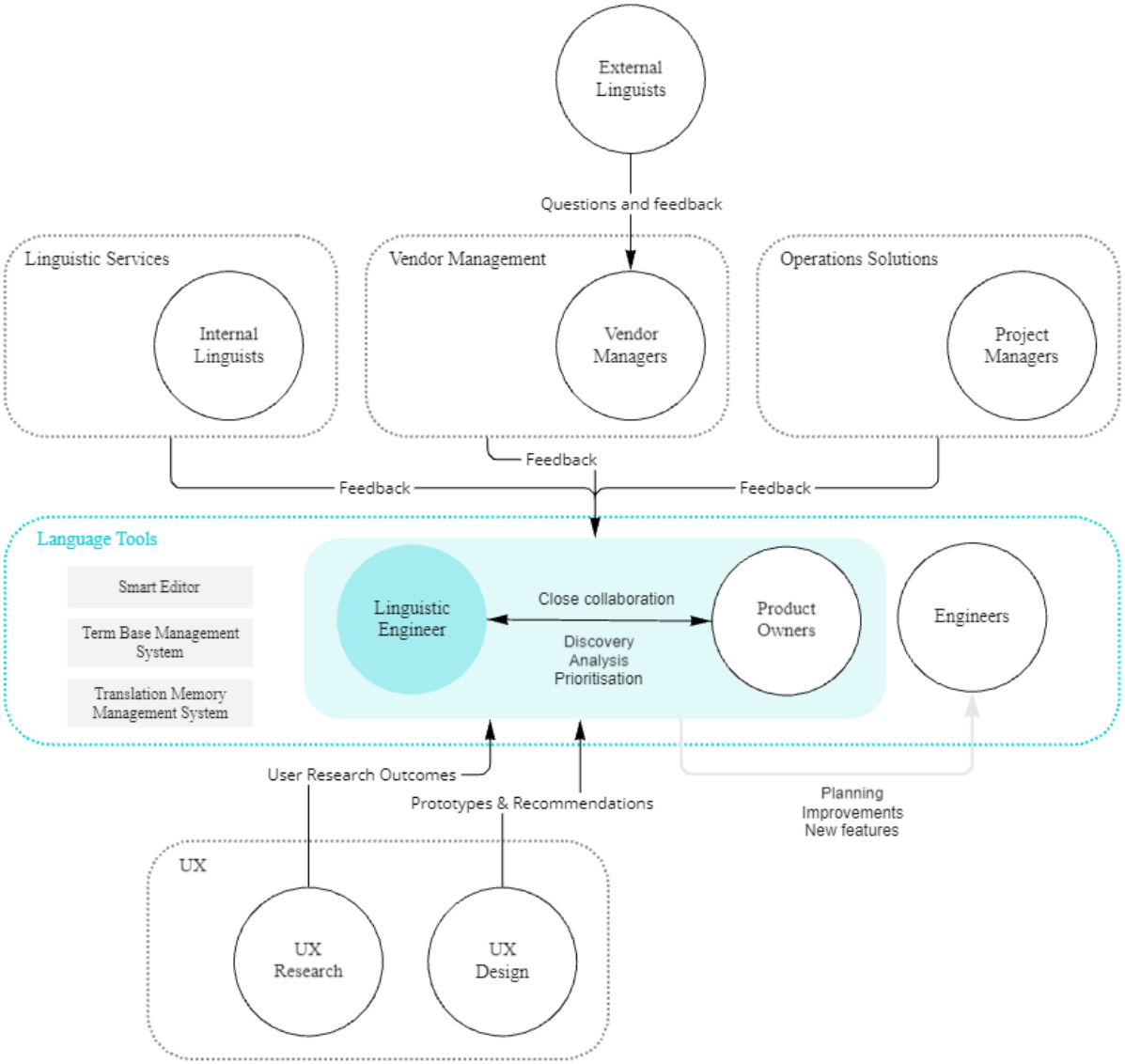


Figure 29: The Linguistic Engineer in the LanguageWire organisational chart.

The Linguistic Engineer position is highly reminiscent of the concept of language engineering introduced by Sager (1994, p.21) as follows:

> *"Language engineering is concerned with the design and use of tools for activities involving languages. Tools are designed to be purpose-specific, to facilitate activities and to solve problems arising from activities. They are applied to the substance of language in general and to manifestations of languages in documents selected for particular applications"*

This idea was reemphasised later by Briva-Iglesias and O'Brien (2022), who described the role of the Language Engineer as a job profile combining language and technology expertise, with the ability to understand how AI is impacting the industry and therefore to advise on the best use of technology.
Furthermore, participating in decisions regarding feature prioritisation and planning of future developments is a particularly rewarding activity, since these decisions directly impact the daily work of translators. Referencing the HCI perspective, O'Brien (2012) emphasised the importance of deeply understanding interaction processes between human subjects and technological solutions in the field of translation. In this sense, the Linguistic Engineer position is an excellent concretisation of such a recommendation. Shaping software products by placing users at the centre goes beyond analysing user needs: it is about humanising technology. Such an ideology is reminiscent of the purpose of Interaction Designers, who, in the words of Kolko (2010, p.14), "exist to support experiences through the continual dialogue between people and products".

Starting in June 2024 and following a restructuring of the Product Management team, the Linguistic Engineer position evolved into that of the User Researcher. This change represented a natural progression, since both roles are deeply rooted in understanding human behaviour, communication, and interaction. More specifically, the key responsibilities of the User Researcher encompass:

- Creating a user insights framework and process for the products and technology LanguageWire offers.
- Using the user insights and data to influence the products and features being developed.
- Putting the user needs at the centre of development decisions.
- Collaborating with internal teams and stakeholders, and ensuring the problems are understood and the solutions are based on the relevant insights.
- Assisting the development teams with user flows, wireframes and prototypes.
- Clearly communicating ideas, thoughts and concepts to the development teams.

While the Linguistic Engineer was focused primarily on Language Tools, in particular Smart Editor, the User Researcher's work includes most LanguageWire products, in particular those overseen by the Machine Learning and LanguageWire Platform teams, as illustrated in Figure 30.

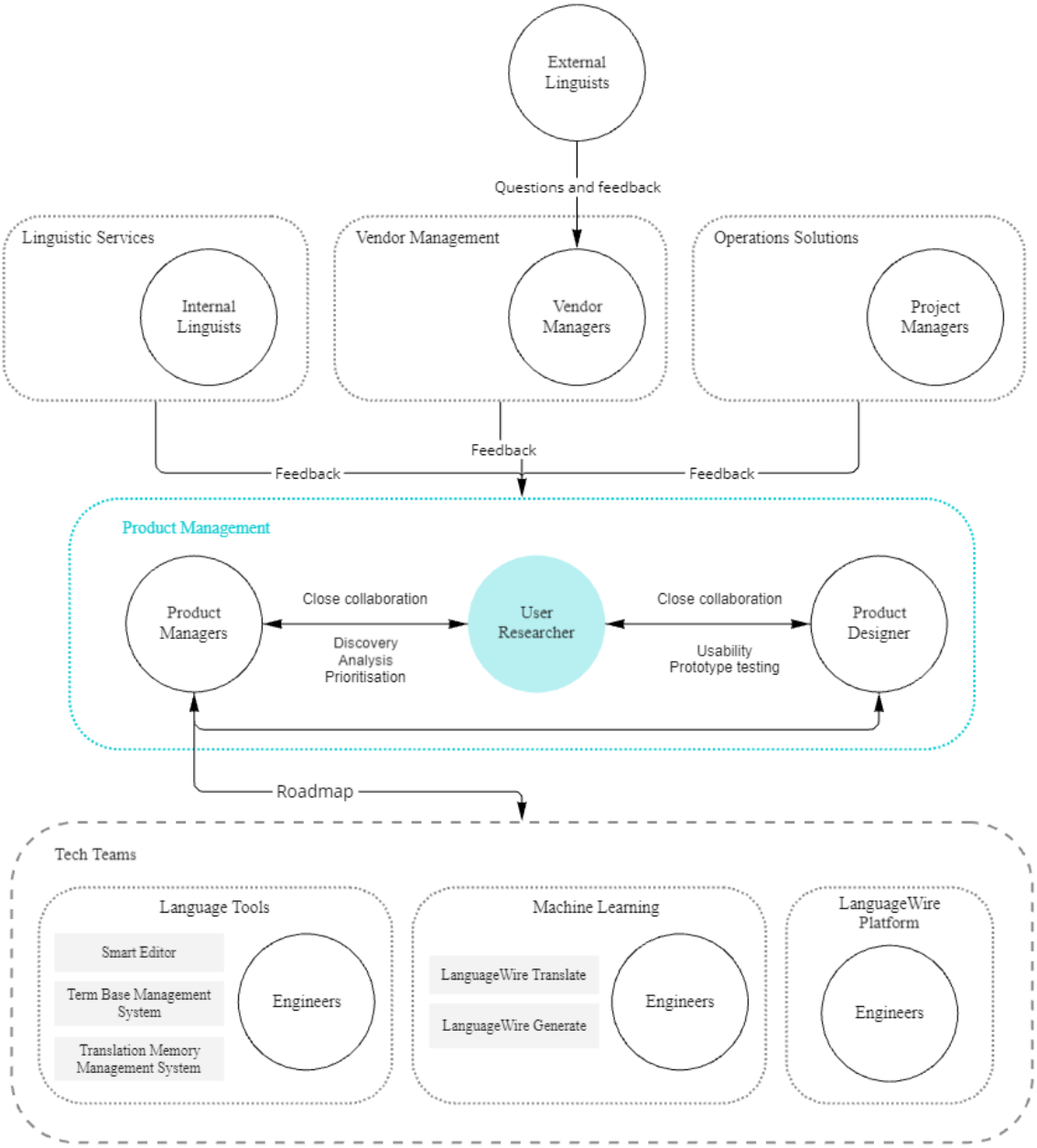


Figure 30: The User Researcher in the LanguageWire organisational chart.

These roles are therefore ideal for conducting research and establishing a collaboration between relevant stakeholders. An industrial PhD creates a stimulating research context and contributes to bridging the gap between the industry and academia. Such research is highly beneficial for both parties. From the LanguageWire perspective, the company ensures that a process is in place for keeping up to date with state-of-the-art research and is represented as an active member in academia. As for the University, such collaboration links academic research to real-life applications and direct users, which constitutes a unique opportunity, given that investigation with industry partners is often more complex to arrange.

8. Methodology: A Benchmarking Approach

The previous chapters provided evidence of the lack of a benchmarking framework to assess the PE environment. To address this gap, the present work proposes to design such a framework. This section is devoted to the introduction of benchmarking and to the conceptualisation of the framework based on the key factors identified in the literature.

8.1 Introduction to Benchmarking

Before delving into the design of the benchmarking framework, it is essential to lay the foundations of benchmarking theory in order to apply it as effectively as possible. To this end, this section aims at defining benchmarking, identifying different methods, and finally discussing the benchmarking of translation tools.

8.1.1 Definition and Types of Benchmarking

Continuous (re)assessment is paramount to identify strengths and weaknesses of a product and to determine which improvements are necessary to achieve a certain target. This is precisely the aim behind *benchmarking*. Etymologically, the term "benchmark" finds its origins in land surveying. In this field, it refers to a point of reference with a known altitude which the surveyor uses as a standard for measuring the elevation of other elements. Today, it has become common to use the word *benchmarking* in a more figurative sense to describe studies which aim at comparing a product to a standard (point of reference). The ISO 25010 standard (ISO/IEC, 2011)[46] provided the following definition of a benchmark: "standard against which results can be measured or assessed".

In computer science, benchmarking typically involves testing a computer programme based on a set of criteria to assess its performance. However, benchmarking can be applied to a wide range of domains

[46] Note that the 2011 version has been withdrawn and is now replaced with the following standards:

- ISO/IEC 25002:2024 on "Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model overview and usage"
- ISO/IEC 25010:2023 on "Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Product quality model"
- ISO/IEC 25019:2023 on "Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality-in-use model"

to assess various elements, going from software products to company processes. More generally, benchmarking can be understood as the

> *"process of identifying the highest standards of excellence for products, services, or processes, and then making the improvements necessary to reach those standards – commonly called 'best practices' "* (Bhutta and Huq, 1999, p.254).

The first application of benchmarking outside land surveying has started in the field of business management. Xerox, a leading company in the printing industry, is known as the pioneer in the history of industry benchmarking. In the late 1970s, this company observed that its Japanese competitors were selling printers at significantly lower prices, which threatened their sales. Robert Camp (1989) therefore launched Xerox's benchmarking programme, which relied on a detailed analysis of the company's direct competitors. More specifically, the components of Japanese printers were analysed carefully with the aim of identifying processes applied to generate outstanding results. Following the success of this programme, benchmarking became a common process for the entire company and was progressively adopted by other industries.

The ISO 14598 standard (ISO/IEC, 2001) on product evaluation states that "software product evaluation depends on a set of evaluation techniques and metrics that provide information about the quality characteristics of the software". Nevertheless, although benchmarking involves a critical evaluation and comparison of various systems or products, it has a broader scope, especially because beyond the mere comparative analyses, it aims to yield further benefits, including discerning best practises and reaching continuous improvements. This distinction is summarised in Figure 31 based on García-Castro (2008).

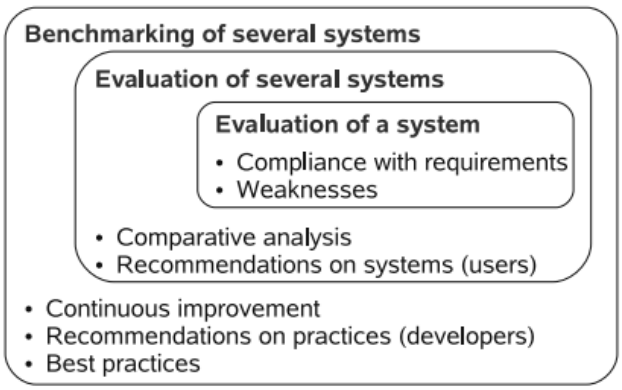


Figure 31: Benchmarking benefits. Source: García-Castro (2008, p.24).

The quality of a software product can be evaluated based on different attributes. The three ISO standards which today replace ISO ISO/IEC 25010:2011 all revolve around " Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE)", but from two perspectives:

ISO/IEC 25002:2024 presents a “Quality model overview and usage”, which is further detailed in ISO/IEC 25010:2023 on the “Product quality model” and in ISO/IEC 25019:2023 on the “Quality-in-use model”.

The product quality model encompasses nine main characteristics, further divided into sub-characteristics, as represented in Figure 32. These provide “a reference model for the quality of the products to be specified, measured and evaluated” (ISO/IEC, 2023).

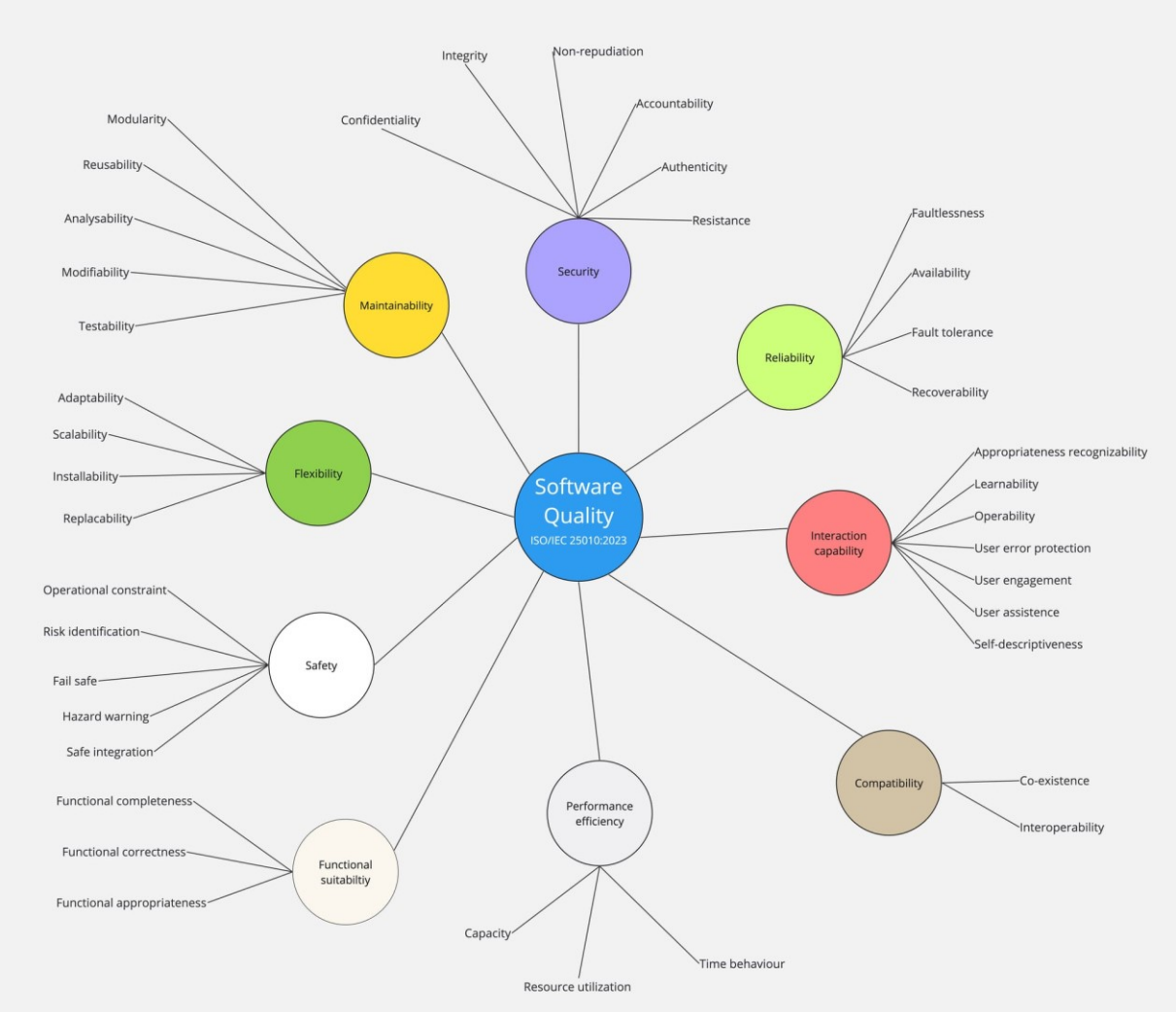


Figure 32: Product quality characteristics presented in ISO/IEC 25010:2023[47].

[47] Source: https://www.workingsoftware.dev/the-ultimate-guide-to-write-non-functional-requirements/ (consulted: 30/04/2025)

On the other hand, the quality-in-use model includes "three characteristics (which are further subdivided into sub-characteristics) that can influence stakeholders when products or systems are used in a specified context of use" (ISO/IEC, 2023).

It should be pointed out that different types of benchmarking exist. Camp (1989) distinguished four categories of benchmarking based on the kind of participants involved (García-Castro, 2008):

- Internal benchmarking involves comparing processes within the same organisation.
- Competitive benchmarking involves a comparison of direct competitors to identify their strengths and weaknesses.
- Functional benchmarking involves comparing processes between organisations belonging to the same industry (the scope is broader than in competitive benchmarking).
- Generic benchmarking involves comparing processes which are common across various organisations, regardless of their field of specialisation.

8.1.2 Benchmarking Methods

Furthermore, benchmarking should be conducted in a specific environment. An ethical framework is essential, as benchmarking is not about encouraging industrial espionage or simply copying competitors. Instead, it is a tool used to identify and understand best practises. Therefore, in the case of competitive benchmarking, it is paramount to establish partnerships with interested stakeholders.
There is not a single model (or a "recipe") to perform a benchmarking study. Several methodologies are conceivable, and the method to be applied can depend on the requirements to be met and the resources available. Xerox developed a 10-step benchmarking process, composed of five stages (Ou and Kleiner, 2015, p.23; Figure 33):

1. Planning: "identifying the process to be benchmarked, finding the competitor or internal department that holds the best practice for the particular process, determining data indicators and method for data collection, and collecting data"
2. Analysis: "identifying performance gaps between the organization and the benchmarking target and forecasting future performance levels for both companies"
3. Integration: "during this phase the benchmark team should present and communicate the findings to management to commit resources to improvement projects and set goals for these projects"
4. Action: "developing action steps for the improvement projects, carrying out the action items, reviewing results, and adjusting the improvement process if goals are not met"
5. Maturity: "the benchmarking team reviews the results from the process that underwent benchmarking and re-evaluates the company's position versus its competitors. If the desired

industry leader position is acquired, benchmarking efforts can be started on a different process. Otherwise, the team should go back to the action phase and adjust the improvement program".

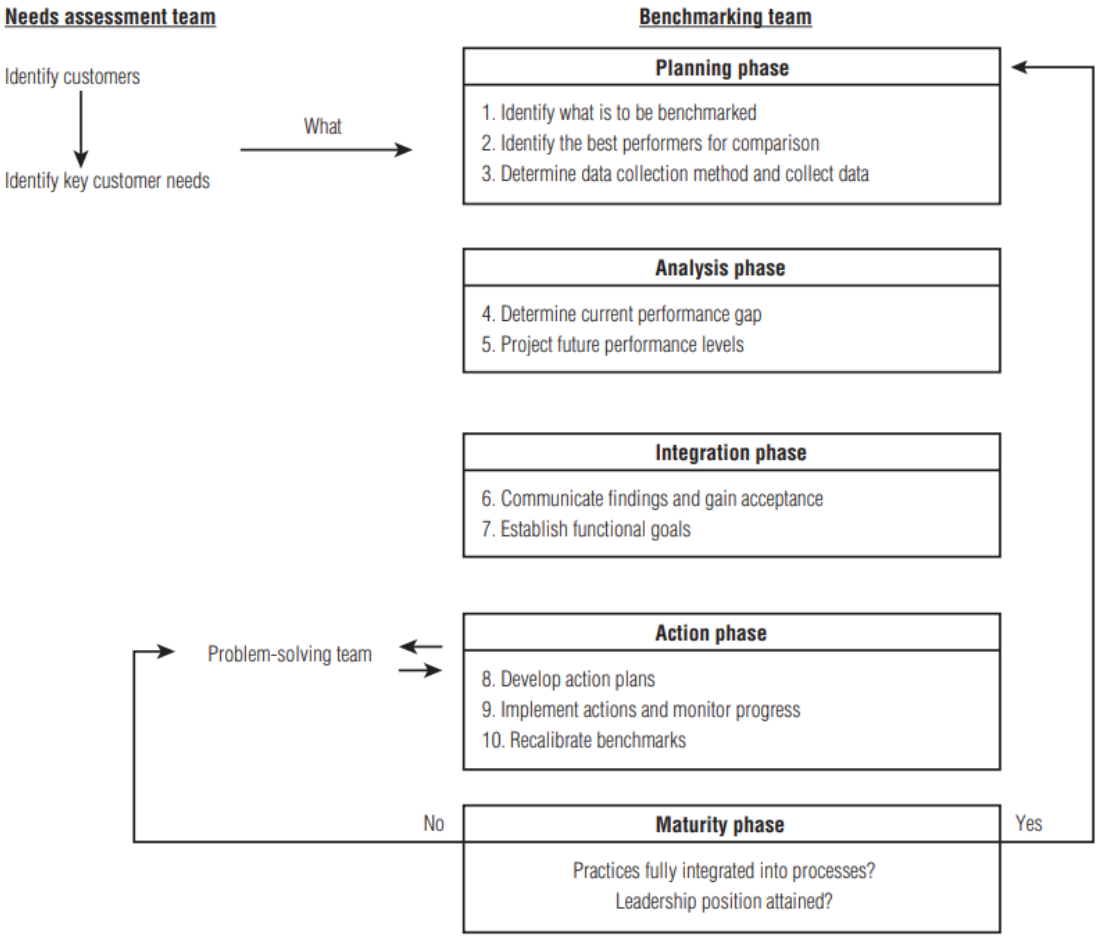


Figure 33: Xerox's benchmarking process (Ou and Kleiner, 2015, p.23).

Maxwell and Forselius (2000) proposed a process which can be applied in various situations and starts by defining the study as follows:

TO <activity>

FROM THE VIEWPOINT OF <kind of software personnel>

TARGET OBJECT <object to be studied>

SO AS TO <objective>

Once the definition and goals are set, the object of study can be characterised by a subset of its attributes, to which different weights can be assigned. These attributes will depend on the object to be studied, and according to the ISO 14598 standard (ISO/IEC, 2001), it might be "necessary to develop new metrics for specific use".

Bhutta and Huq (1999) also state that the process can take various forms and can be implemented in different numbers of steps, but it remains the same in essence. However, general recommendations can be applied in most situations, and the majority of benchmarking processes follow a similar cycle. Relevant modelisations of this cycle include PDCA (Plan, Do, Check, Act; Bhutta and Huq, 1999) and the benchmarking wheel introduced by Andersen and Pettersen (1995).

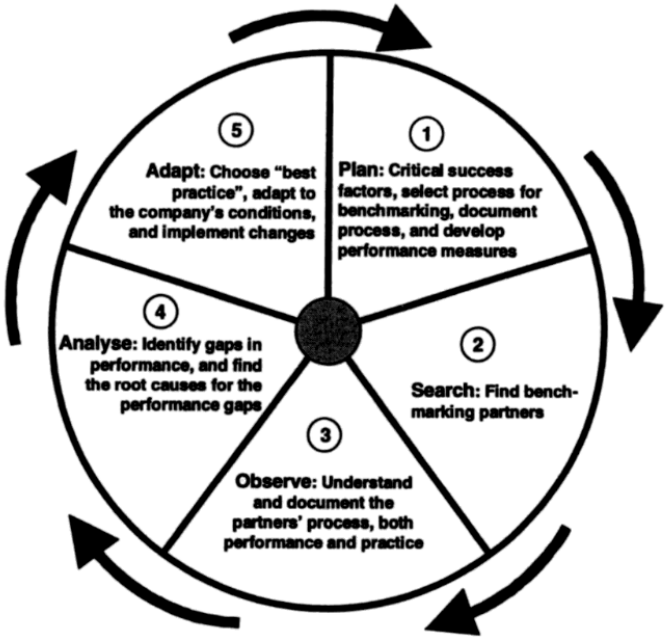


Figure 34: The benchmarking wheel. Source: Andersen and Pettersen (1995, p.14).

As can be seen in Figure 34, this wheel defines benchmarking as an iterative process in which the five major components (plan, search, observe, analyse and adapt) are strongly connected. Despite a few discrepancies from one benchmarking model to the other, particularly in the number of steps, these processes share the same overall structure. After comparing several models, García-Castro (2008) reached the same conclusion and summarised common tasks across benchmarking methodologies (see Table 24).

Phase	Task
Plan	Identify the process to be benchmarked
	Form and train the benchmarking team
	Analyse and document the current process
Data collection	Define criteria for benchmarking partners
	Identify potential partners
	Contact the potential partners and enlist
	Define method and tools for data collection
	Collect data
Analysis	Determine the current performance gaps
	Identify the causes for the gaps
Change/adapt	Communicate findings and gain acceptance
	Formulate strategy to close the gaps
	Develop the implementation plan
	Implement the plan
	Monitor and report progress
	Document the study

Table 24: Common tasks in benchmarking methodologies (García-Castro, 2008, p.31).

In each case, the process starts with a planning phase in which the benchmarking objectives are defined. Data is collected and studied based on the objectives. Conclusions are derived from this data analysis in the form of best practises, which are in turn implemented in the company's processes.

8.1.3 Benchmarking Translation Technology

In a software development context, benchmarking also plays a significant role and should be performed on a regular basis. In today's fast-paced technological environment, agile methods are often employed in order to ensure rapid development and frequent releases, while maintaining a functional product throughout time (Sommerville, 2011). To achieve this, it is necessary to find the right balance between user satisfaction, incremental delivery and flexibility for future developments. This point of equilibrium is somewhat reminiscent of the concept of the dance of agency. Agile software development therefore relies on the involvement of users at early stages of the development process and on continual assessment in order to identify key areas for improvement.

Benchmarking thus appears as a powerful approach and has been successfully implemented in various industries. In the field of translation technology, numerous studies have examined the UX of translation tools (as described in 2. on Assessing the Post-Editing User Experience in Translation Tools) and even conducted comparative analyses (e.g. study of Makoushina [2007] on QA checks). Krüger (2016, p.115)

pointed out that the literature tends to focus on these types of studies (i.e., "product reviews, product comparisons or aspects of workflow management"), and although usability is indirectly encompassed in such topics, no explicit UX model had been proposed.

The same goes for benchmarking. The majority of previous studies are typically rooted in ad hoc methodologies, which are admittedly well documented and have a certain degree of replicability but are not based on a unified method – i.e., a benchmarking process.

One could argue that the 7-step recipe elaborated as part of the EAGLES-II project (see section 2.1.5 on Standardised Evaluation Models) is somewhat close to a benchmarking methodology, as it details fundamental steps for evaluating language technology – which are reminiscent of the stages of the benchmarking methods described above. However, although it provides insightful guidance, the EAGLES model remains relatively generic and does not specify the dimensions to be investigated.

Therefore, to the best of our knowledge, no benchmarking framework has been developed to assess translation tools – especially for PE.

8.2 Applying Benchmarking to the Post-Editing Environment

Having established the theoretical foundations of benchmarking, this section focuses on its application to the PE environment. The aim is to develop a Benchmarking Framework (BF) that is both generic enough to be widely applicable and specialised enough to address the unique demands of translation technology, particularly the user requirements associated with PE. To achieve this, the section first clarifies the scope and purpose of the benchmarking effort, situating it within a broader classification of benchmarking methodologies. Key benchmarking elements are then extracted from the theoretical background to define the framework's pillars, including insights from the translator's workstation, UX, PE process, and translation quality. These elements are distilled into three core pillars which form the foundation of the BF. By integrating these dimensions, the framework aims to provide a comprehensive and user-centred approach for the evaluation of PE environments.

8.2.1 Defining the Benchmarking Context

Before diving into the specific elements that will compose the BF, it is paramount to acknowledge that these elements are deeply rooted in a certain context. The definition of "context of use" provided by Krüger (2016, p.132) in the CAT tool usability model is highly relevant here:

> *"The general model of CAT tool usability is embedded in a specific context of use. This is intended to highlight the in vivo nature of usability, i.e. the fact that usability can only be established relative to specific users pursuing specific goals in a specific context […]. For example, the suitability of the standard configuration of a given CAT tool cannot be determined in isolation but only if we specify which task the*

configuration is to be suitable for and in which situation the CAT tool is to be used. This emphasis on the context of use or the situation in which a given CAT tool is employed is also in line with the basic tenets of Situated Translation, which stresses the strict situation-dependence of the translator's cognitive performance".

For the present work, the context is defined as a PE task performed by a linguist in a CAT tool. Consequently, it should be acknowledged that PE is a situated activity, which entails an interplay between various factors, from the characteristics of the text at hand (e.g. complexity, language pair) to cognitive processes, which are in turn influenced by the individual experience of the linguist (e.g. training, prior experience with technology).

Taking the process of Maxwell and Forselius (2000), the design of the BF could be formulated as follows:

TO <activity> evaluate
FROM THE VIEWPOINT OF <kind of software personnel> post-editors
TARGET OBJECT <object to be studied> the PE environment (CAT tool)
SO AS TO <objective> optimise the PE environment

Since several approaches are conceivable when it comes to assessing software products, it is necessary to further clarify the scope. More specifically, a tool can be evaluated in terms of functional and non-functional requirements (Sommerville, 2011), which are two distinct yet complementary aspects of software quality. Functional requirements specify what the software must do (i.e. the functions it should perform). They define the core actions that enable the software to fulfil its intended purpose and are thus essential to ensure that the software meets user needs and achieves its intended objectives. On the other hand, non-functional requirements have to do with how well the software performs these functions and the overall experience it provides. They address quality attributes like performance, usability, reliability, and security, which are not directly related to specific functions but impact how effectively the system operates.

In the case of a CAT tool, typical functional requirements could include XLIFF support, TM and TB retrieval, and MT integration, whereas non-functional requirements could encompass elements such as technical performance (e.g. response time), security, and usability (e.g. intuitiveness of the interface).

The present work focuses on certain aspects of software quality that affect the PE experience, which includes both functional requirements (i.e. the set of features supported by the tool) and non-functional requirements (i.e. the usability of the tool). Other dimensions which are purely technical, such as performance, security or reliability, are out of scope of the evaluation model herein presented, as it is assumed that optimal performance is always a must-have. However, it is acknowledged that technical factors may have an impact on some BF dimensions (e.g. a slow response time will affect both productivity and UX). Furthermore, it should be noted that this work does not focus solely on software

quality from a UX perspective but also revolves around further dimensions which are relevant in a PE context, as further detailed in the following section.

8.2.2 Identification of the Benchmarking Pillars

This section provides a summary of the key elements extracted from the theoretical background and how they materialise in the BF. To facilitate comprehension, these elements are presented in table format (Table 25).

Outcomes of literature review	Coverage in the BF
Translator's Workstation	
Following ALPAC, the focus has been on tools to assist the human translation process, going from the first conceptualisations (such as the Translator's Amanuensis) to modern CAT tools and Augmented Translation. Technology should assist language professionals by augmenting their capabilities.	The present work adopts a user-centred approach and also considers that the PE environment should allow for extending the capabilities of language professionals.
Placing the translator in the driver's seat is crucial (e.g. allowing the post-editor to select the desired level of assistance [Melby 1982]).	The importance of control is further investigated in the UX section of the BF.
CAT tools were originally conceptualised for working with TBs and TMs. Some adaptations have been developed to better support PE (e.g. QE, IMT, MMPE, AI assistance), however these technologies would need to be further evaluated.	The evaluation of adaptations of CAT tools to better suit PE represents the fundamental reason behind the design of the BF. Additionally, the BF is tested to evaluate one of these adaptations.
Today, MT is replacing HT for productivity reasons – the underlying assumption being that MT should make the process easier, faster and ultimately more efficient.	The PE environment should allow post-editors to be more productive. Therefore, Productivity is considered as a pillar for the BF.
UX	
As software products are continuously evolving, agile software development is a relevant method to keep users involved throughout the development cycle.	The BF is envisioned to be a continuous process.
Understanding user needs is essential in software development, and in the field of HCI, a number of usability and UX assessment methodologies have been proposed. User studies in translation technology have	There seems to be no consensus on a unified evaluation framework on a UX method for the assessment of translation technology. Since usability and UX models were developed for more generic

typically been conducted using a wide range of methodologies, sometimes adopting UX approaches (e.g., Briva-Iglesias, O'Brien and Cowan, 2023) or creating new models (e.g., Krüger, 2016).	software products, it appears necessary to find out which UX dimensions are the most important to post-editors to design an assessment method that truly reflects their needs.
PE Process	
The PE process was defined as composed of three phases (orientation, adjustment, self-revision) and consisting of four possible editing actions (deleting, inserting, moving, replacing).	Gaining a thorough understanding of the activity assessed was considered a prerequisite for the design of the BF.
Specific MT errors (Temnikova, 2010) or editing operations (Popović et al., 2014) are more demanding than others (e.g.), and after exceeding a given editing threshold, the post-editor is retranslating more than performing PE (do Carmo, 2017).	MT quality can be envisaged as a predictor of the PE effort (to complement the Productivity component of the BF).
The PE competence includes: • Translation competence (linguistic, cultural, research, tools) • Error handling (spotting, classification, correction), which relies on MT knowledge, and plays a significant role in terms of PE effort • Decision making (accept, revise, reject, correction to apply) • Guidelines application	The PE environment should provide relevant support for MT error identification and correction as well as decision making. This observation should guide the development of technology tailored to the PE process.
The PE effort can be measured based on three dimensions (temporal, technical, cognitive). Cognitive friction should be kept minimal: when distinguishing the intrinsic load from the extraneous load, the latter should be minimised. In other words, the focus should be on the text (intrinsic load), while interaction with the tool should be smooth and should not interfere with the main task (extraneous load).	The PE environment should not interfere with the post-editors' thinking process, instead it should allow them to focus on the task at hand. This observation is to be considered when defining the UX pillar.
Direct users of translation technology have tended to feel excluded from the development of the tools they are using (Ehrensberger-Dow and O'Brien, 2015; Fırat, 2021).	UX is an integral part of the BF, and it is designed with post-editors' feedback in mind.

Quality	
Quality is paramount in the translation industry, notably because it is a key factor to ensure customer satisfaction.	Given its central place, Quality is considered as a BF pillar.
Quality assessment is an arduous ask, particularly because quality is an elusive concept (inter-annotator agreement issues are frequent in human evaluation due to subjectivity).	When defining the Quality pillar, examining various perspectives, notably via analysis of inter-annotator agreement, is deemed essential.
A substantial number of quality evaluation models have been conceptualised before NMT and may fail to capture NMT-specific issues.	It appears relevant to adapt quality models for the assessment of more recent technology (NMT).
The expected quality of post-edited texts is often the same as for HT.	Since the expectations are the same, using conventional assessment models seems acceptable to evaluate post-edited texts.

Table 25: Summary of outcomes of the literature review.

The synthesis of key elements from the theoretical background has led to the identification of three core pillars for the BF: productivity, UX and quality. These pillars reflect dimensions that are not only theoretically significant but also practically essential for improving PE environments. Their selection is supported by empirical evidence, such as the COACH study (Bota, Schneider and Way, 2013), which demonstrated that improvements in translation tools are most effective when addressing speed, quality, usability, and utility. Similarly, Briva-Iglesias (2024) also focused on UX, productivity and quality to assess interactive PE.

It is worth acknowledging here that these three pillars are deeply interconnected, and their interplay is likely to affect the BF results. For instance, if the UX is suboptimal, it can negatively impact the mental processes of post-editors, leading to reduced productivity and, in turn, lower translation quality. Conversely, a well-designed UX can enhance efficiency and performance, creating a positive feedback loop that benefits all three pillars. This interplay is further supported by theoretical insights. As do Carmo (2017, p.7) noted, "efficiency and quality must go together, since investing solely in one of them is not an excuse for producing translations that fail in the other". Similarly, the balance between productivity and quality is a critical dimension of professional translation practice, especially when PE is involved. In that regard, do Carmo (2017, p.187) observed that "translation practice has always been a demonstration that productivity and quality do not have to be incompatible". These insights highlight the importance of addressing all three pillars from a holistic perspective, ensuring that improvements in

one area do not come at the expense of another. Therefore, the relationship between the BF pillars should also be examined.

The selection of these pillars is not only theoretically grounded but also aligns with the practical needs of post-editors. By focusing on productivity, quality, and UX, the BF provides a comprehensive model for evaluating and improving PE environments. It ensures that tools are designed to meet the linguistic and ergonomic needs of users, while also supporting efficient and high-quality workflows. This approach not only enhances the immediate experience of post-editors but also contributes to the long-term development of translation technology.

Besides the three core pillars, it is also acknowledged that external factors are likely to impact productivity, UX and quality – though they are not studied directly within the present research. Notably, the different dimensions of the PE effort are not considered. The temporal effort is directly encompassed within the BF via the *Productivity* pillar. The technical effort is considered indirectly in an *MT quality* satellite parameter, which is intricately linked to *Productivity*. The cognitive effort, which is challenging to quantify directly, may be reflected via the *UX* pillar. Although this topic is considered out of scope for the present research, exploratory analyses of the cognitive effort could investigate whether the *UX* dimension indeed correlates with more objective cognitive measures like pause frequency or revision patterns. Other practical elements are also likely to influence the benchmarking results. The purpose of PE (e.g., rapid turnaround vs. high-quality output) may skew productivity measures. For instance, a focus on speed may increase PE productivity at the expense of quality or UX. Similarly, the experience of post-editors, from the perspective of both MT quality and UX of the tool, may vary based on the language pair due to resource availability (e.g., low-resource languages) or linguistic complexity (e.g., morphologically rich languages). Specialised domains or highly creative texts may also require more extensive corrections, thus also affecting productivity and quality outcomes. Post-editors' proficiency with translation technology and expertise in the domain also influence their interaction with the PE environment. For example, a linguist with advanced CAT tool skills may achieve higher productivity scores regardless of UX limitations. In certain situations, financial limitations may restrict access to high-quality MT systems or advanced tool features, therefore also impacting benchmarking outcomes. The BF accounts for these factors by contextualising results within project goals, ensuring that the results are interpreted based on intended outcomes; however, it does not aim at measuring the impact of these factors on the results.

8.2.3 Presentation of the BF Design and Application Processes

Having identified the three core pillars of the BF, this section details their definition process and integration into a cohesive assessment method (Figure 35). Each pillar is indeed defined following specific methodologies:

- *Productivity* is quantified as the number of words post-edited per hour.
    - *MT quality* is considered a satellite parameter to contextualise the *Productivity* pillar by accounting for the effort required to edit the raw MT output.
- *UX* is assessed through a PE-specific UX questionnaire, designed specifically to evaluate the PE environment based on post-editors' requirements.
- *Quality* assesses the final post-edited text. Since quality expectations for post-edited content remain consistent with quality levels expected from HT, quality can be evaluated using standardised methodologies.

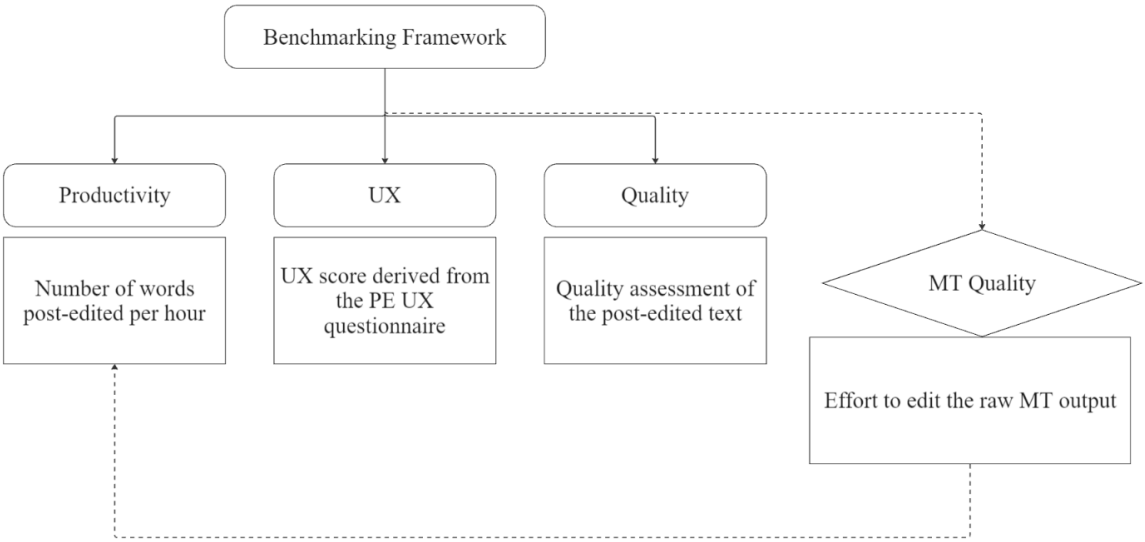


Figure 35: Structure of the BF.

To further illustrate how the present work is structured, a schema of the approach for both the design and application of the BF is included and shall serve as a map for the research conducted herein (Figure 36). It is based on the Xerox process (adapted from Ou and Kleiner, 2015), structuring the research into five primary phases.

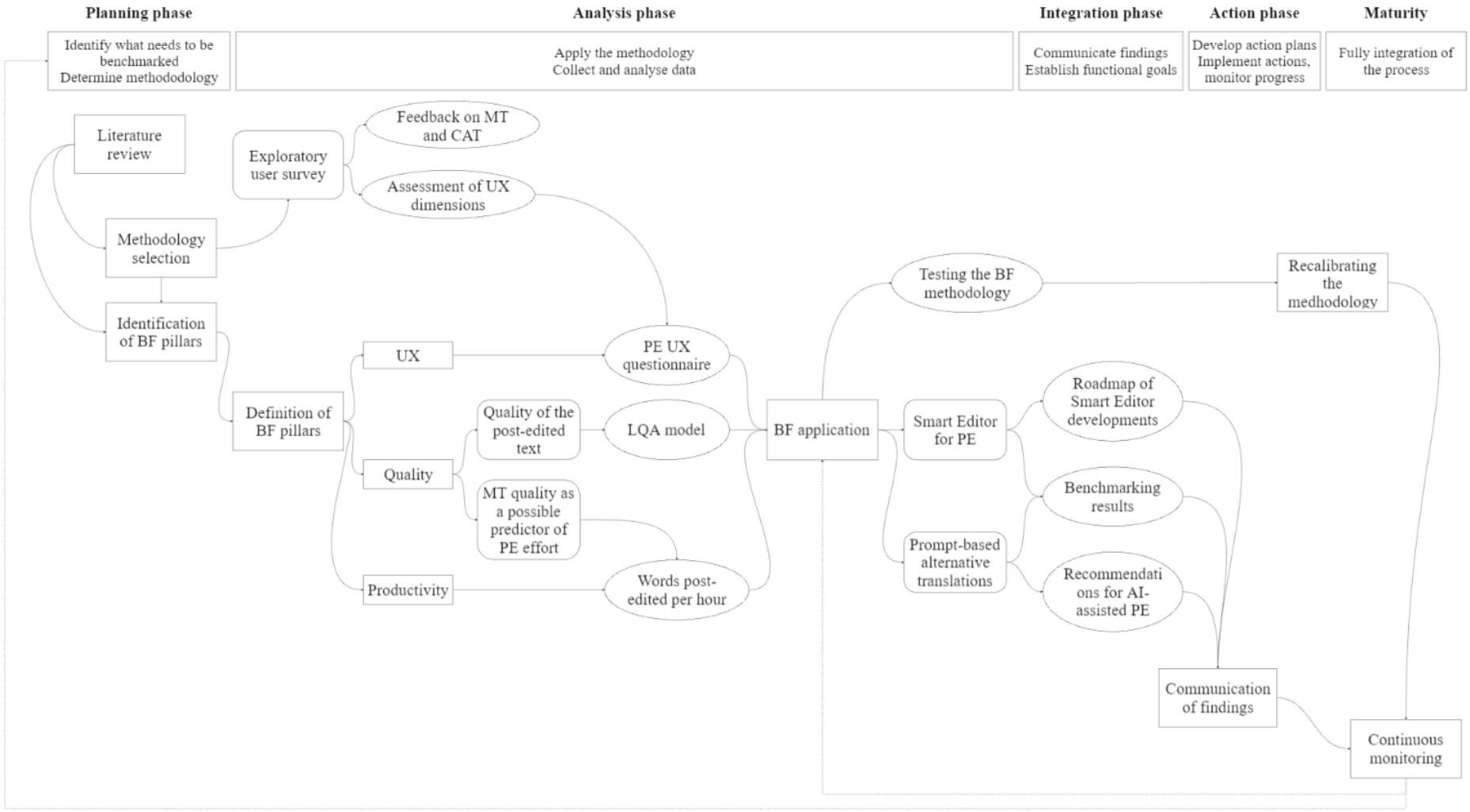


Figure 36: Approach to BF design.

The planning phase established the theoretical and methodological foundations of the BF. A comprehensive literature review identified benchmarking as the most suitable approach for evaluating PE environments holistically, while also revealing gaps in existing assessment. This phase allowed for determining the framework's core pillars (productivity, UX and quality) as well as a satellite parameter to account for MT quality to contextualise productivity measurements.

During the analysis phase, these theoretical pillars were operationalised through empirical data collection and refinement. An exploratory survey of post-editors provided essential input on typical PE challenges and clarified the UX dimensions that are deemed essential to post-editors. These insights directly informed the development of a tailored PE UX questionnaire which reflects the specific needs of post-editors. Simultaneously, the quality pillar was structured to evaluate both the final post-edited text (using traditional LQA models) and to explore the relationship between MT quality and PE effort. Productivity was defined as the number of words post-edited per hour, with possible variations based on MT quality. The framework was then tested via two applications. In an industrial context, the BF evaluated LanguageWire's proprietary CAT tool, Smart Editor, over a one-year period (2024), which resulted in the formulation of actionable improvements. A second application assessed prompt-based alternative translations using an AI assistant in Trados Studio, which shed light on the benefits and challenges of AI integrations in CAT tools for PE. These applications not only validated the BF's practicality but also allowed for identifying specific recommendations for the PE environment in each scenario.

In the integration phase, findings were synthesised with a view to communicating them and proposing functional goals based on the BF results.

The action phase translated these insights into a concrete planning of future objectives. For Smart Editor, the BF results directly informed the 2025 development roadmap, prioritising enhancements that appeared crucial to post-editors. For AI-assisted PE, the evaluation of prompt-based suggestions led to concrete guidelines for optimising this feature based on direct feedback.

Finally, the maturity phase ensures the BF's relevance and potential evolution as a dynamic evaluation tool. The framework is designed for recalibration, allowing adjustments to pillar definitions based on specific needs and available data (e.g., measuring customer feedback if LQA cannot be completed for the entire dataset under evaluation) or the incorporation of new metrics (e.g., cognitive load measurements). This phase should also prepare subsequent benchmarking cycles, enabling continuous assessment of next-generation tools.

By adopting this structured yet adaptable process, the BF bridges academic rigour and industrial applicability. Its phased approach – from theory-driven planning to data-informed action – provides a replicable model for evaluating PE environments while remaining responsive to the rapid evolution of translation technologies.

IV. ANALYSIS & DISCUSSION

Having established the objectives and methodology, this chapter presents the design of the framework and demonstrates its application to two real-world scenarios.

9. Design of the Benchmarking Framework

Building on the theoretical and contextual foundations established in the previous chapters, this section focuses on the design of the BF, following the methodology previously described. The aim of the following sections is to define how the BF operates in practice, detailing the research conducted to establish an assessment model for each pillar.

9.1 Exploratory User Survey of Post-Editing Requirements

Since the present work adopts a user-centred approach, the first step involved conducting a user survey to gather feedback from post-editors and shape the BF directly based on what matters the most to them (Escribe and Candel-Mora, 2024). This section therefore describes the questions that this survey sought to answer and presents the analysis of responses. The fundamental areas explored here consist of the current satisfaction of post-editors with CAT tools and MT quality, the difficulty of various PE-related tasks, the importance of UX dimensions, and the relevance of state-of-the-art PE features.

9.1.1 Rationale

The landscape of CAT tools has undergone a significant evolution in recent years, transitioning from the early days of TMs and TBs to the contemporary era of hybrid approaches integrating TMs with PE of MT output. PE has become a widespread practice in the translation workflow, and this evolution has entailed changing needs and expectations for language professionals. As a result, translators engaging in PE tasks had to navigate this change, adapt to new ways of working and new guidelines, and learn to find a balance between efficiency and quality.

While CAT technology did not seem to adjust to PE initially, the industry is witnessing a progressive incorporation of state-of-the-art features, such as QE and AI assistants based on LLMs. In today's rapidly evolving technological landscape, collecting user opinions is paramount in software development. Knowing that several studies have deplored the sentiment of exclusion that language professionals tended to experience in the past (Lagoudaki, 2006; O'Brien, O'Hagan and Flanagan, 2010; do Carmo, 2017; O'Brien, 2020), elucidating user needs is paramount to ensure valuable developments.

Conducting a user survey to find out more about the experience of post-editors thus appeared a natural first step.

9.1.2 Objectives and Research Questions

The primary objective behind this survey hence lay in gathering and analysing post-editors' feedback regarding their experience with MT in CAT tools. More specifically, it sought to answer the following research questions:

- How satisfied are post-editors with the tools they are currently using to perform PE?
- How satisfied are post-editors with MT quality?
- What PE tasks are the most difficult? Specifically,
    - What MT errors are the most difficult to handle for post-editors?
    - What editing actions are the most difficult to perform for post-editors?
    - What decisions are the most difficult to make for post-editors?
- What UX dimensions matter the most to post-editors? Note that the data collected to answer this question is analysed in 9.1, as an integral part of the UX pillar definition.
- Which state-of-the-art PE feature appears the most attractive to post-editors?

9.1.3 Experimental Setup

This section describes the methodology used to design the survey. It consisted of three main parts: demographic data, PE experience and a feature wish list.

The first section, Demographics, consisted of nine elements to clarify the background of respondents, including age, technical literacy level, employment type, experience in translation and in PE, number of PE jobs usually completed per month, domain (with answers based on LanguageWire's list of industries[48]), usual language combination and most frequently used tool(s). All questions were close-ended, except the language combination, where respondents were advised to enter their answer in the format "Source>Target".

The second section focused on the UX of post-editors. It was divided into four subsections: a) overall satisfaction, b) PE actions, c) PE environment characteristics and d) UX elements extracted from various usability or UX models (such as Nielsen, 1994; Brooke, 1996 and Hassenzahl, Burmester and Koller, 2003). This section comprised 37 questions in total (33 of which were ratings, three were multiple-choice, and one was open-ended). The outcomes of this UX section were analysed in two phases. First,

[48] https://www.languagewire.com/en/industries (consulted: 30/04/2025)

a) overall satisfaction and b) PE actions are analysed in the remainder of this section. Next, an in-depth study of the responses in c) PE environment characteristics and d) UX elements is presented in 9.3, as it served as a foundation for designing the UX pillar of the BF.

The overall satisfaction subsection included two questions about the general satisfaction with tools for PE and with MT quality. The PE actions subsection was composed of three multiple-choice questions:

- *Which type of MT error(s) do you consider the most challenging to correct?* For this question, a list of 25 MT errors inspired by Costa et al. (2015) was provided, and respondents were offered an *Other* option to indicate errors which did not appear in this list.
- *Which editing action(s) do you find the most challenging?* The possible answers included the four editing actions (deleting, inserting, moving and replacing; do Carmo, 2017), as well as *It depends on the context* and *None*.
- *Which decision(s) do you find the most challenging?* Four decisions were provided (deciding whether to edit a segment, which correction to apply to an MT error, order in which corrections should be addressed, when and how often to refer to external resources) together with an *Other* option.

Finally, the last section consisted of a feature wish list, where respondents were asked to indicate the relevance of potential PE-centred features on a 5-point Likert scale. The list contained 11 features derived from 1.3 on Latest Advances in translation technology. For each feature, a short definition was also provided as follows:

- ***Interactive post-editing****: the translation is generated and updated on the fly, as you are typing.*
- ***Segment-based alternative suggestions****: offers alternative translations for each segment.*
- ***Word-based alternative suggestions****: offers alternative translations for each target word.*
- ***Confidence scores (quality estimation)****: for each MT segment, a score indicates the confidence estimation of MT quality.*
- ***Knowledge feature****: provides definitions, examples or images upon request to help you understand difficult concepts in the source text.*
- ***Rewriting assistance****: offers rewriting suggestions for predefined situations, such as changing the style or shortening a translation.*
- ***Multimodal interactions****: using different interaction modalities, including touch and speech input, as well as text to speech.*
- ***Document view****: switch from segment to document view, where the text can be revised and modified in context.*
- ***Effort prediction****: an expected time to completion for the current project is provided based on past performance.*

- ***Attention monitoring****: an analysis of behavioural patterns detects drops in attention and displays a warning or offers assistance.*
- ***Post-task feedback****: after project completion, provides immediate feedback based on the proportion of accepted, amended and rejected segments.*

In addition, two open-ended questions were included to collect feedback on the items listed above and to gather new feature ideas.

9.1.4 Analysis

The survey remained open for approximately one month (from 12/02/2024 to 10/03/2024). It was distributed to professional translators via direct emails or messages and was shared as a public post via LinkedIn. In addition, a second LinkedIn post was published in a private group for freelance language experts working for LanguageWire. In total, 126 answers were collected.

9.1.4.1 Demographics

Most survey respondents were between 25 and 44 years old (Figure 37), 50% (63) of them considered they were proficient with technology (Figure 38), and the majority were freelancers (Figure 39). The seven respondents who answered "other" included one student, one LSP owner, two freelancers with another full-time job, and three scholars. The distribution of the PE workload was rather unbalanced, with 35% (43) completing less than five PE jobs per month and 25% (32) completing more than 20 (Figure 40). Nevertheless, these figures only provided an overall idea, as the word count per PE job was not specified here. 42% (53) respondents had over 10 years of experience in translation (Figure 41). For PE, however, most respondents had between 1 and 6 years of experience (Figure 42), with more specifically 1-3 years for 33% (42), and 4-6 years for 38% (48).

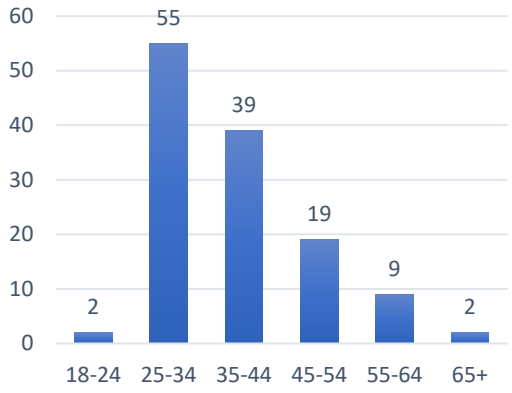


Figure 37: Age distribution.

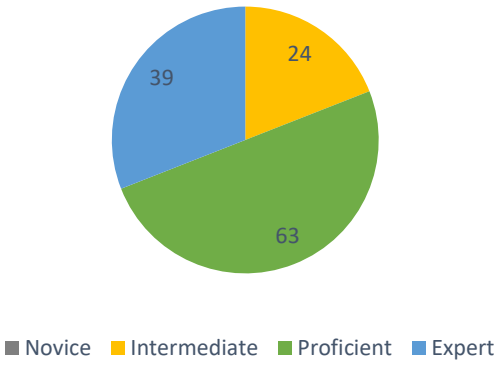


Figure 38: Technical literacy level.

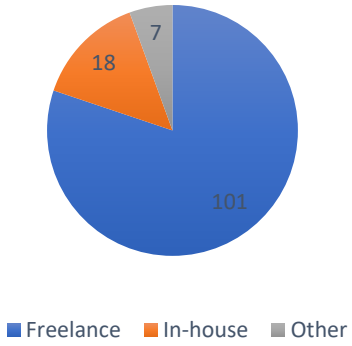


Figure 39: Employment type.

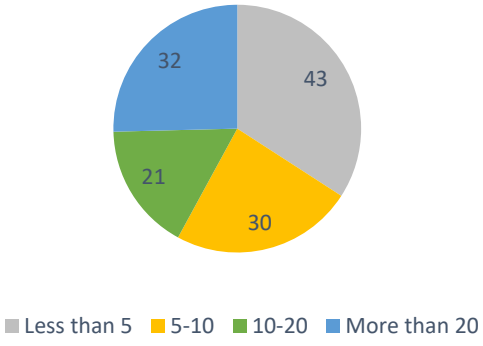


Figure 40: Monthly workload (PE jobs).

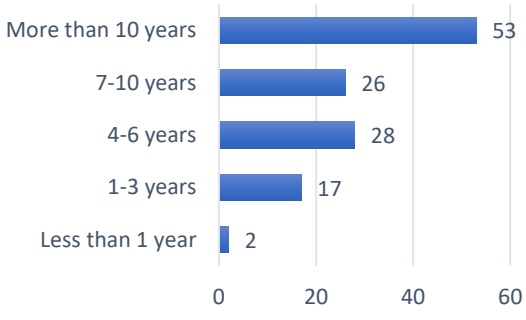


Figure 41: Experience in translation.

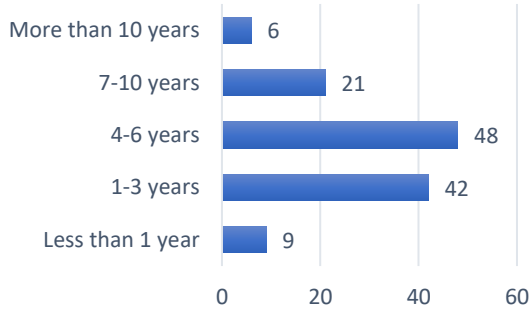


Figure 42: Experience in PE.

The most recurring target languages (Figure 43) were Italian (32), Spanish (27) and English (22). Respondents worked in a wide variety of domains, with the top three including marketing and

advertising, technology, and life sciences and healthcare (Figure 44). Apart from the domains listed in the close-ended question, some respondents mentioned video games (5), literature (2), education (2) and human resources (2).

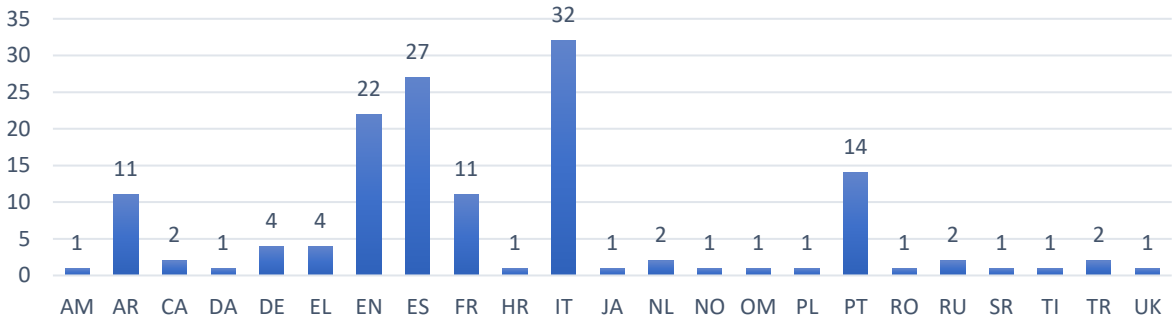


Figure 43: Target languages, following two-letter codes of ISO 639:2023 (ISO, 2023).

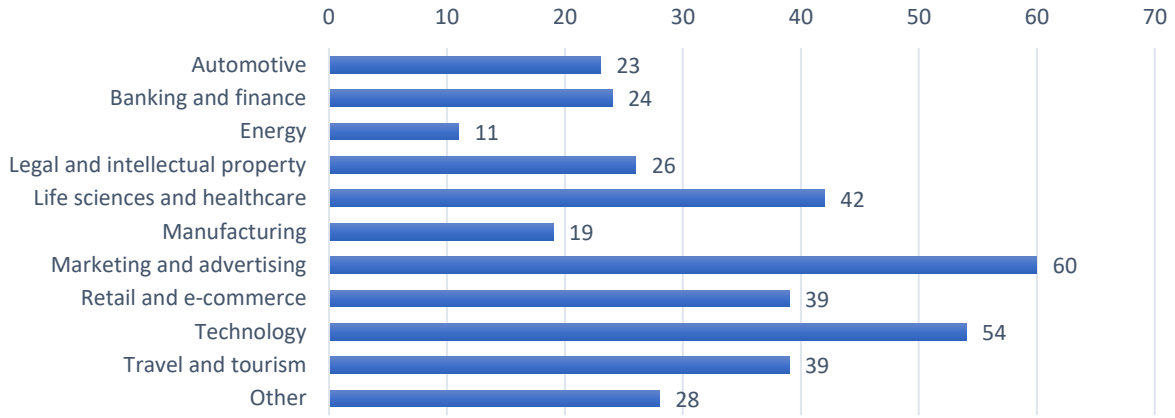


Figure 44: Domains.

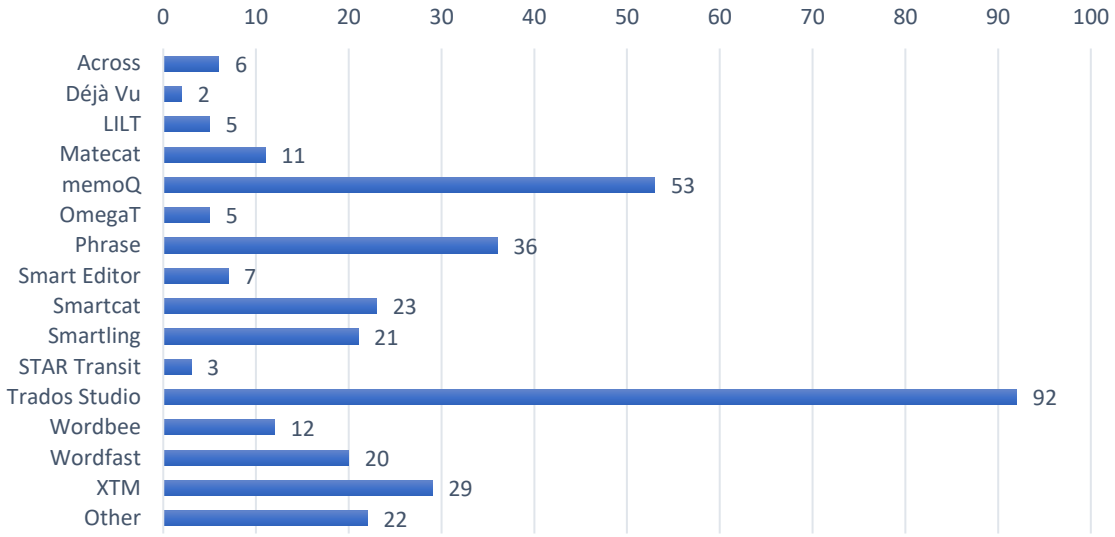


Figure 45: Most frequently used tools.

As far as tool usage is concerned, Trados Studio with 92 users and memoQ with 53 users led the list, followed by Phrase with 36 users (Figure 45). Several respondents also mentioned GlobalLink (4) and Polyglot (3).

9.1.4.2 Satisfaction and Post-Editing Needs

The overall satisfaction with translation tools for PE received an average score of 3.68 (on a 5-point Likert scale, where 1 = Not satisfied at all and 5 = Extremely satisfied.), with 46% (59) of respondents giving a score of 4, while 30% (38) were more neutral and gave a score of 3 (Figure 46). The outcome was more nuanced in terms of the overall satisfaction with MT quality. While the average satisfaction was 3.29, 40% (51) respondents were neutral, and 38% (48) were satisfied (Figure 47).

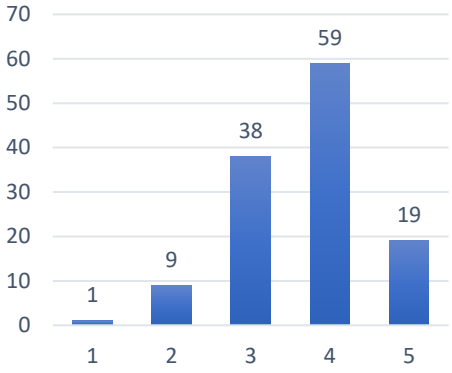


Figure 46: Satisfaction with tools for PE.

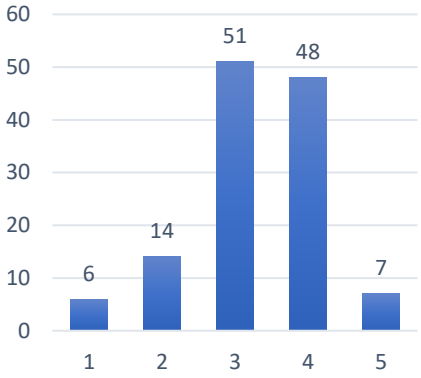


Figure 47: Satisfaction with MT quality.

Regarding MT error handling, cultural adaptation appeared to be the most challenging (76), followed by mistranslation (60) and source language interference (60). Frequently mentioned errors also comprised adherence to guidelines and reference material, document-level errors and idiomaticity (Figure 48). It should be mentioned here that out of the seven free-text answers collected for this question, four respondents mentioned issues related to tag handling. As illustrated in Figure 49, the difficulty of editing actions appeared to be correlated with the context (81), with the most challenging one being replacing parts of the MT output (40).

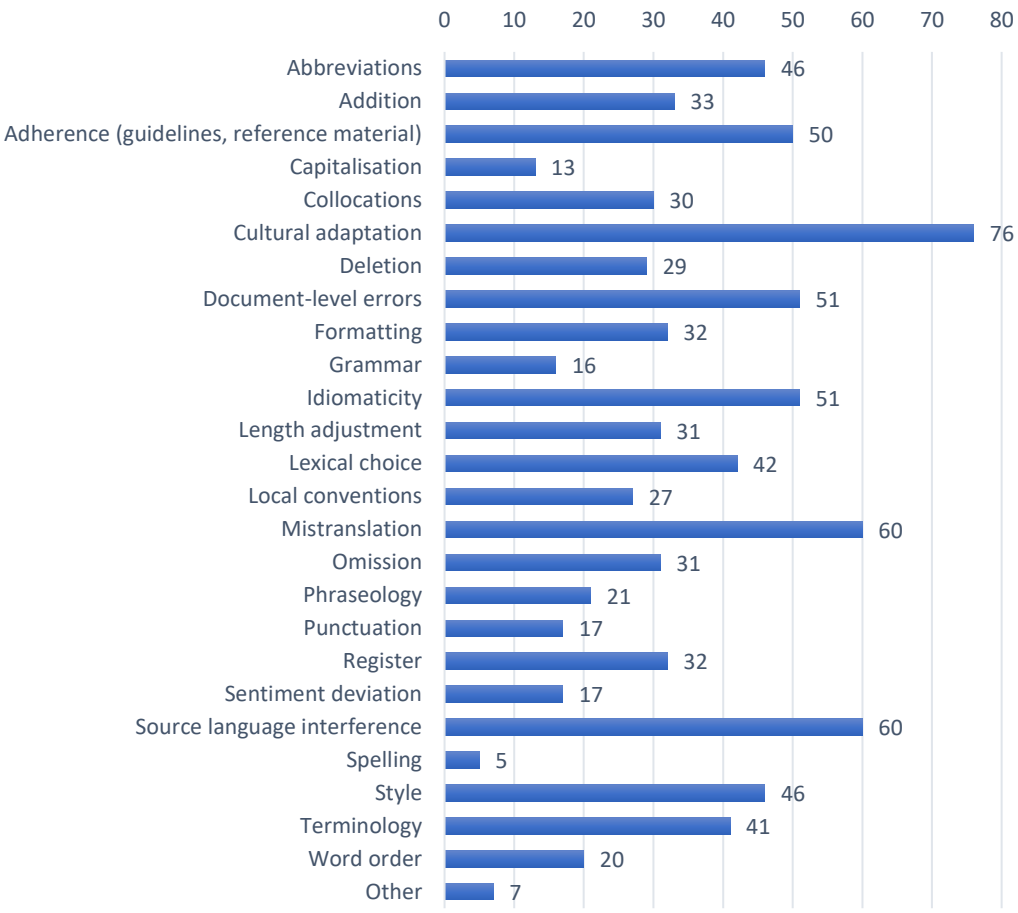


Figure 48: MT error handling difficulty.

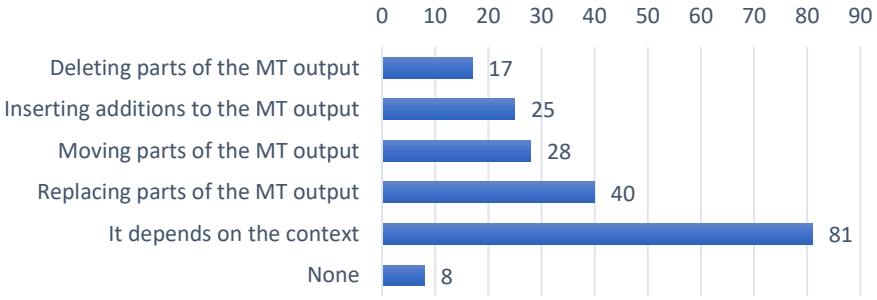


Figure 49: Editing action difficulty.

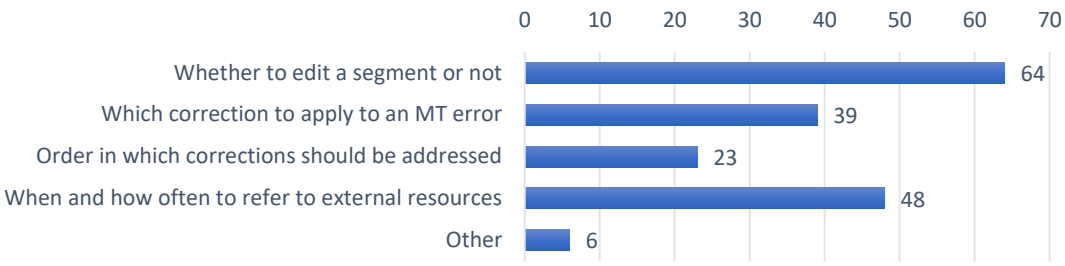


Figure 50: Decision difficulty.

When it comes to decision making (Figure 50), the most challenging decisions seemed to be deciding whether to edit a segment or not (64), followed by deciding when and how often to refer to external resources (48). In addition, six comments were provided in the *Other* answer option. Two focused on consistency, including consistency between the MT output and TM content, as well as consistency in style and terminology, which can be challenging for long projects. Two other comments emphasised the difficulty of estimating the severity of an error and the extent to which a segment should be modified.

9.1.4.3 Feature Wish List

Most features in the wish list appeared to be relevant since none received an average score below 3 (where 1 = Not relevant at all and 5 = Very relevant; Figure 51). The feature with the highest score was "document view" with an average of 4.23, followed by the "knowledge feature" (4.06) and "rewriting assistance" (3.96). These features also received the highest number of votes in the most positive category, with 63 for "document view", 61 for the "knowledge feature", and 49 for "rewriting assistance" (Figure 51). Conversely, the least popular items, "attention monitoring" (3.03) and "multimodal interactions" (3.10) received the highest number of negative votes (21 in both cases).

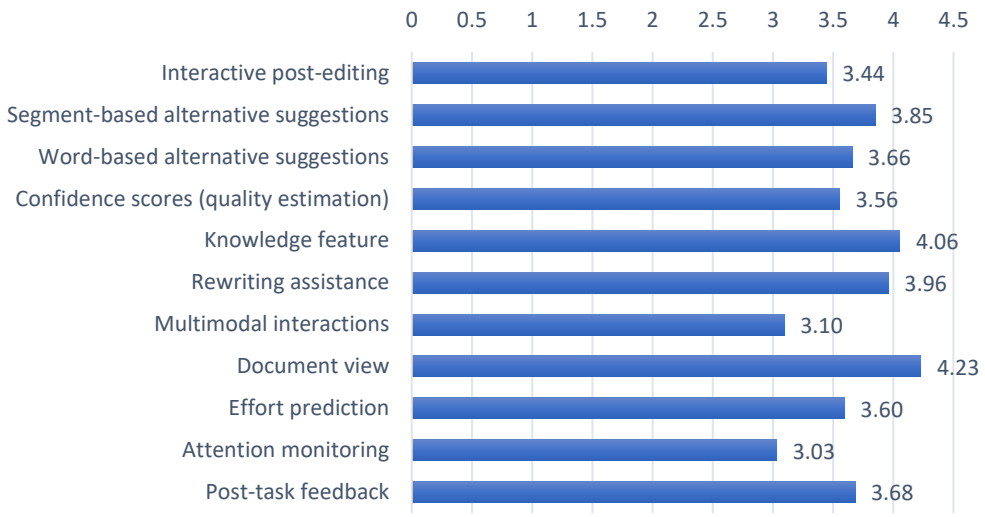


Figure 51: Average feature scoring.

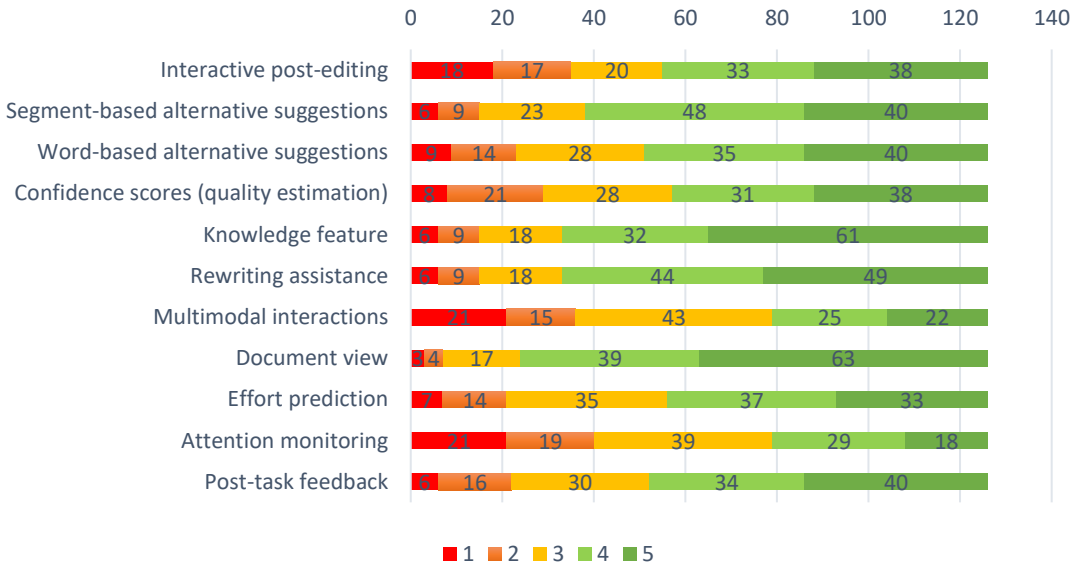


Figure 52: Score distribution.

Respondents had the possibility to leave additional feedback on the features of the wish list. QE attracted the attention of four linguists, who seemed to be in favour of this functionality, but also advised that it could also "*be a double-edged sword*" because it can be "*very useful if the estimation is extremely accurate*" but "*distracting otherwise*". One respondent mentioned that confidence scores could help in the decision-making process since "*the tricky part of MT is to spot when it's actually not that good even though it seems so*". Nonetheless, pricing appears to be a concern in this case, since LSPs can use QE to filter out segments above a predefined confidence score, and one person reported a negative

experience in that regard. Similarly, another respondent was worried that an attention monitoring feature could result in "*micro-managing and hyper-analysis of productivity*". Alternative translations, in contrast, generated enthusiasm. One respondent suggested to provide alternatives based on past corrections implemented during PE, and another one proposed to combine this feature with the knowledge feature and thus providing alternative translations together with "*definitions or more context, to better understand which option to choose*". Another respondent highlighted the importance of designing features that let translators stay in control and focus on their task without interfering heavily in their thinking process: "*I do not enjoy features that interfere too much with my input, but I value features that support the translation work (instead of feeling like I support the machine, not the other way around) and enhance my productivity by cutting corners (e.g., features such as Confidence scores and Knowledge feature […]) and allowing me to focus fully in the translated text*".

Finally, respondents were given the chance to describe other feature ideas. In total, 12 people shared their suggestions. Some answers mentioned features such as tag insertion, search and replace, access to TBs, history tracking and quality assessment. However, these functionalities are already available in various CAT tools today, and they are not specific to PE. Therefore, we focus here on more original suggestions. Below is the list of features extracted from these answers:

- **Unpopulated segments**: leaving target segments empty when the MT quality is not satisfactory in order to avoid the "*anchoring effect*"
- **Highlighting unedited MT output**: clearly differentiating segments which have been post-edited from those which still require to be checked ("*with MT target language already present in the target segment, it's often hard to discern at a glance what you have and haven't edited and makes it harder to 'trust' that you're 100% ready to sign off on a segment, which slows down decision-making*")
- **AI assistance** (i.e. integrating GPT technology), including
    - Idiomaticity support: assistance to provide support for idiomatic expressions
    - Inclusive suggestions: rewriting target segments in a more inclusive language
- **Correction propagation**: automatically applying a lexical or terminological change to the rest of the text
- **Time tracker**: integrated time tracking functionality which could provide productivity statistics for each project
- **MT engine comparison**: analysing the differences between outputs from different MT systems in order to identify which one is the most similar to the final text after PE ("*so that, if one engine seems to 'think' more in the way I do, I can realise that and start using that engine for that type of text*")

Arguably, differentiating unedited MT output is to some extent already covered in several tools. For example, in Trados Studio and Smart Editor, segments which were checked are indicated with a specific icon (a green checkmark in both cases). However, this suggestion may reveal a need to distinguish more clearly unedited machine-generated content. AI assistance does not come as a surprise in the era of GenAI. Some integrations, like the OpenAI Translator plug-in for Trados Studio, already offer the possibility to reformulate translations based on predefined prompts. The proposal for correction propagation is reminiscent of online adaptation of automatic PE systems (e.g. Simard and Foster, 2013), the incremental adaptation introduced by Escribe and Mitkov (2023), and the algorithm to automate repetitive PE actions proposed by Terribile (2024a). While having an online model run continuously in production can be expensive, this suggestion shows that adaptive systems capable of leveraging human input on the fly would be highly beneficial.

The average satisfaction with translation tools for PE (3.68) indicates that there is substantial room for improvement. The difficulties encountered in MT error handling seemed to be rather well aligned with the feature wish list. Cultural adaptation, for example, could be supported by AI assistance for rewriting thanks to tailored prompts. In the case of other errors considered challenging (mistranslation, source language interference, and idiomaticity), providing support can be more difficult since these errors require a thorough understanding of the source and target languages. However, QE may come in handy to spot such errors. Adherence to guidelines and reference material also appeared to be challenging, which is in line with the decisions considered difficult, in particular deciding when and how often to refer to external resources. In that regard, the knowledge feature, which was among the top voted items, could be helpful, in particular if it can be customised to extract contextual information from reference documentation. As suggested by one respondent, such a feature could also use this contextual knowledge to provide alternative translations. Document-level errors, which also appear among the most challenging MT errors, could be easier to handle with a propagation mechanism and potentially with enhanced visualisation of the document, which was the highest scoring feature.

The other feature ideas, in particular highlighting unedited MT and leaving unpopulated segments, emphasise the need for clarity and for making enough space for the post-editors' thinking process. As pointed out by one respondent, the PE environment should not interfere with the main task at hand but rather provide support for decision making and other activities related to the main task, such as researching unfamiliar concepts or checking reference material. This approach is reminiscent of Kay's proposal (Kay, 1980), whereby technology would gradually take charge of certain tasks, but without supplanting the human translator, who would remain in the driver's seat. In this sense, the knowledge feature appears particularly pertinent. Furthermore, some concerns were raised regarding the use of QE and attention monitoring. The consequences of using such features should be investigated further, notably when it comes to pricing models.

9.1.4.4 Perceived Importance of Each BF Dimension

In addition to gathering feedback on the PE experience of respondents, this survey also allowed for finding out what importance post-editors attribute to each of the pillars identified for the BF. The results (1-5 ratings for each pillar) are presented in Figures 53-56.

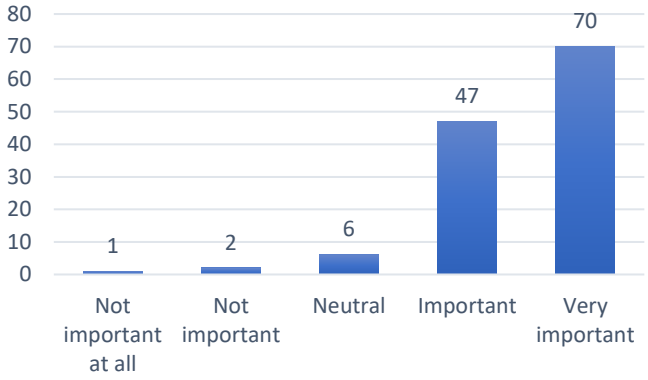


Figure 53: Importance of productivity.

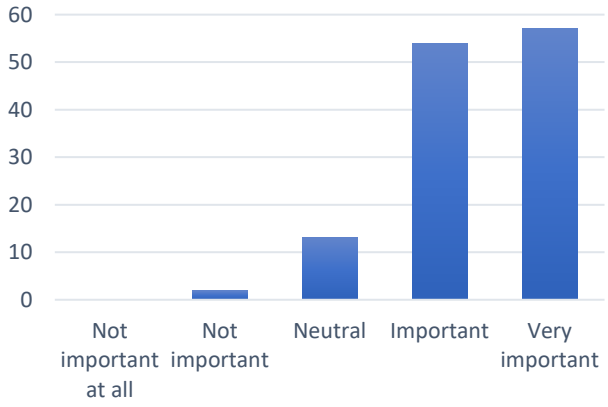


Figure 54: Importance of UX.

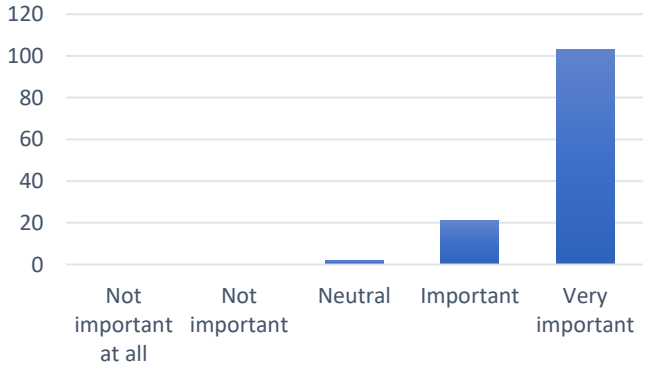


Figure 55: Importance of translation quality.

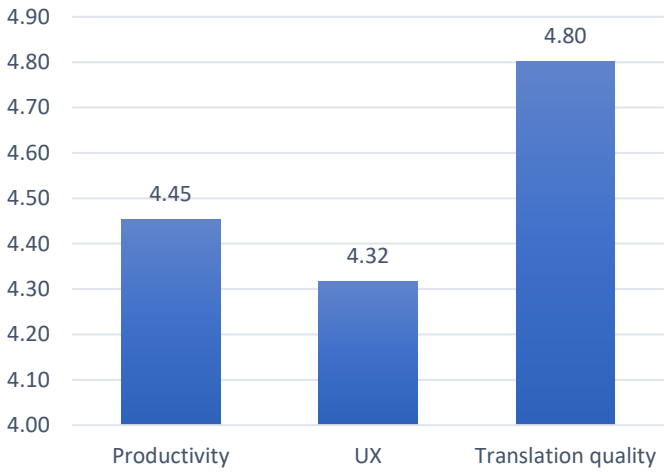


Figure 56: Average scores for the three dimensions.

Variable	Average rank
Productivity	4.45
UX	4.32
Translation quality	4.80
$\chi 2$ (Friedman Test Statistic)	44.450
p-value	2.23e-10

Table 26: Average ranks and Friedman test of rank significance (productivity, translation quality and UX).

Translation quality emerged as the most important aspect, with the highest average rank of 4.80, followed closely by productivity (4.45) and UX (4.32). While this shows that respondents tend to consider translation quality as the most relevant aspect, the gap between the other dimensions is rather narrow, which suggests that all three dimensions have a comparable importance. A Friedman test was performed to find out whether the ranked preferences of respondents are statistically significant (Table 26). The Friedman test yielded a significant result ($\chi 2 = 44.450$, $p < 0.001$), thus indicating significant differences in respondents' preferences for the three dimensions.

This implies that respondents do not uniformly prioritise these aspects, and there are notable variations in their preferences. While the perceived importance of the three BF pillars appears comparable, knowing which aspects are the most important to post-editors is relevant to design evaluation processes that accurately reflect their expectations.

To pinpoint where the significant differences identified by the Friedman test lie, post-hoc pairwise Wilcoxon signed-rank tests with a Bonferroni correction were conducted (adjusted $\alpha = 0.0167$ for 3

comparisons). This method compares every possible pair of BF pillars while adjusting for the increased risk of false positives that arises when making multiple comparisons. Significant differences emerged between *translation quality* vs *UX* and *translation quality* vs *productivity*, but no significant difference was found between *productivity* and *UX*.

Furthermore, the relationship between the satisfaction with tools and the importance attributed to each of the three dimensions is also worth exploring.

Variable 1	Variable 2	Spearman's ρ	*p*-value
Productivity	Satisfaction	0.07	0.41
UX	Satisfaction	-0.02	0.86
Translation quality	Satisfaction	0.04	0.69

Table 27: Relationship between productivity, translation quality and UX vs satisfaction with tools (Spearman test).

To that end, a Spearman rank test was performed to determine whether a correlation exists between these variables (Table 27). The results indicate no clear relationship between satisfaction and the BF pillars.

9.1.5 Summary, Outcomes & Future Research

This study provided insights into post-editors' needs by elucidating which MT errors, editing actions and decisions are the most challenging during PE, and by identifying possible features to cater to these needs. The results show that cultural adaptation, mistranslation and SL interference are the most laborious when it comes to error handling, followed by adherence to guidelines and reference material, document-level errors and idiomaticity; the difficulty of editing actions depends on the context; and the most challenging decisions were whether to edit a segment or not and when and how often to refer to external resources. Overall, the difficulties identified align with the preferred features, namely, the document view, the knowledge feature and rewriting assistance. In addition, further suggestions were proposed by respondents, including, for example, highlighting unedited MT and automatic correction propagation. These findings should help in guiding software developers to design useful improvements to translation tools that cater to PE needs. In addition, the importance attributed to each proposed BF pillar should guide future evaluation frameworks. Here, post-editors attributed a slightly higher importance to translation quality, followed by productivity and UX.

It should be acknowledged here that while the sample size is considered sufficient to formulate practical recommendations, it can be limited in comparison to other studies (e.g. Moorkens and O'Brien, 2017). The clear prevalence of Romance languages may also have affected the results. It would thus be recommended to extend this survey by disseminating it to more post-editors working with different

target languages. Furthermore, it should be noted that the number of open-ended questions was kept to a minimum in order to reduce the time required to complete the survey given the total number of questions. However, open-ended questions are essential to capture certain details and nuances; therefore, it would appear relevant to combine the outcomes obtained here with interviews or focus groups to gain a deeper understanding of user needs.

The results obtained in this survey, apart from providing valuable insights into common challenges faced by post-editors, also served as the backbone for the remainder of the present research. Notably, the MT errors judged the most difficult to handle supported the creation of the error typology created when investigating the relationship between MT error type and editing effort (in 9.4.1) and the identification of the most popular CAT tool features for PE informed the selection of the PE assistance feature selected for the BF application (in 10.2).

9.2 Productivity Pillar Definition

In the context of the BF framework, the productivity pillar is quantified as the number of source Words Per Hour (WPH), a metric widely adopted by LSPs and researchers to assess the impact of MT adoption and the efficiency of MT systems.

Several industry experts indicate drastic rises in productivity with the use of MT. According to Slator, while human translators can deliver 2,000-3,000 words daily, using MT can boost that number up to 7,000 (Stasimioti, 2022).

Blagodarna (2018) found 500-700 WPH to be a common throughput, with experienced post-editors reporting higher productivity at 900-1,000 WPH. Similarly, a study by Läubli et al. (2019) PE productivity of 494-934 WPH depending on language pair and task complexity. However, productivity is highly dependent on individual experience, language pairs, and content type. Such factors highly influence throughput, to the point where it becomes difficult to identify a real industry benchmark. Notably, Terribile (2024b) conducted a large-scale study, involving 90 million words translated across 2.5 years by 879 linguists in 11 language pairs and reported mixed results. Although PE was found to be 66% faster than HT on average, the efficiency of PE varied significantly from one language pair to the other, with average speed variations ranging between +130% and -7%. This suggests that further research is needed in this area to further determine the conditions under which PE could be improved and which variables are the most likely to affect speed. For instance, high WPH rates may indicate high MT quality or tool efficacy – or even both – but could mask usability flaws if achieved through excessive cognitive strain. Therefore, it is crucial to contextualise the productivity pillar within the benchmarking context, and interpret the values based on the other BF pillars.

9.3 UX Pillar Definition: Design of a PE UX Questionnaire

This section presents the work conducted to define the UX dimension of the BF. Specifically, it studies the UX dimensions that are deemed the most important by post-editors based on the results from the exploratory user survey. The outcomes of this study then serve as the basis for elaborating a custom PE UX questionnaire which is used to assess the UX dimension of the BF. It should be highlighted here that in this section, "questionnaire" (part of the BF, used to assess the UX dimension) should not be confused with "survey" (conducted with a view to designing the "questionnaire").

9.3.1 Rationale

The UX principles to assess translation tools described in 2. highlight a noticeable lack of a common framework to assess the UX of translation tools, particularly in the case of PE. Since the main advantage in using MT lies in productivity gains, the general satisfaction of language professionals has tended to be overlooked. Briva-Iglesias and O'Brien (2023) contended that MT tended to be examined from a productivity and quality perspective, without considering the overall interaction of linguists with it. Beyond the MT output, the same claim can be extended to translation tools in which PE takes place. However, user-centred studies have tended to focus on evaluating the usability of raw MT output from the perspective of post-editors (Castilho et al., 2014; Doherty and O'Brien, 2014; Koponen et al., 2020; Karakanta et al., 2022). Studying the experience of post-editors beyond the MT output, and also considering the environment in which PE occurs, therefore appears essential. While some user studies of CAT tools have been conducted (Coppers et al., 2018; Rodríguez Vázquez, Fitzpatrick and O'Brien, 2018; Alotaibi, 2020; Lee et al., 2021; Briva-Iglesias, O'Brien and Cowan, 2023), they systematically reused predefined usability or UX models, such as UEQ or SUS. Therefore, although these studies provided valuable insights into the experience of CAT tool users, they employed generic methods widely used in HCI, without exploring which dimensions of UX are the most important to post-editors, which calls for specific research on that topic. This gap is even more significant since users themselves tend to feel excluded from the design of the tools they work with (Fırat, 2021), when they should in fact be included in the development process (Ehrensberger-Dow and O'Brien, 2015).

9.3.2 Objectives & Research Questions

Therefore, this study aims to elucidate Research Questions (RQs) in two main categories:

1. Importance of PE needs

RQ 1.a: Which PE needs are the most important to translation technology users?

RQ 1.b: What is the relationship between PE needs and the current satisfaction with tools?

2. Importance of UX dimensions

RQ 2.a.: Which UX aspects are crucial for language professionals performing PE?

RQ 2.b.: What is the relationship between UX aspects and the current satisfaction with tools?

RQ 2.c.: Are there any other UX aspect(s) which are essential to post-editors?

Based on the answers to these RQs, this research also aims to design a proposal for a novel UX framework which accurately reflects the needs and views of post-editors.

9.3.3 Methodology

This study takes inspiration from the work of Zaretskaya (2016), which was based on a feature checklist where the weights of each component were determined by their popularity in a user survey. Hence, it is proposed to find out the importance of UX aspects for PE via a user survey, and to formulate an evaluation framework based on the results of this survey.

More specifically, the study presented in this section analyses the remaining part of the user survey already presented in 9.1. The PE UX section of the survey was further divided into two subsections. The first one consisted of questions related to respondents' actual experience around translation tools for PE via a generic question (*How satisfied are you with the tools you are currently using for post-editing projects?*). More detailed aspects, including productivity and the identification of MT errors, were also explored. The aim was to identify potential issues which may impact users' satisfaction. More specifically, users were asked about their current satisfaction with MT quality and with the tools they are using for PE; they were also asked about the types of MT errors, editing actions and decisions which they consider the most challenging. Further details on the survey as well as the answers to those questions were discussed in 9.1. Then, a list of seven PE needs was identified and included the below items:

- ***Completeness****: The post-editing environment provides the features that I need to complete post-editing projects*
- ***Productivity****: The post-editing environment allows me to be more productive.*
- ***Decision making****: I am able to make decisions efficiently when working in the post-editing environment.*

- ***Guidelines****: I am able to follow the guidelines easily when working in the post-editing environment.*
- ***Control****: I feel a sense of control when working in the post-editing environment.*
- ***Cognitive load****: The post-editing environment does not interfere with my thinking process and allows me to focus on my task.*
- ***MT error handling****: The post-editing environment provides relevant support for the identification and handling of MT errors.*

The selection of these items was motivated by several reasons. *Completeness* and *control* were included to assess whether the PE environment provides the necessary support and places the user at the centre; *productivity* was added because it is the main argument for the shift to PE workflows; *cognitive load* was considered important to evaluate whether the CAT tool adds to the extraneous load of the linguist; finally, *decision making*, *guidelines* application and *MT error handling* were incorporated as they represent core skills of PE (Ginovart Cid, 2021). For each of these statements, respondents were asked to rate the importance on a 5-point Likert scale. The choice of this scale was motivated by the fact that it allows for more nuanced insights by providing a neutral option, and the higher data quality and reliability it proved to yield (Adelson and McCoach, 2010).

The objective of the second subsection was to find out what respondents deemed important in a PE environment, independently of their current experience. For example, a post-editor may be dissatisfied with the documentation of the tool they are using but still consider this aspect important. For this section, major UX elements were extracted from the literature. To that end, the usability and UX models referenced in 2. were compared. This analysis led to the extraction of 49 dimensions in total, with 10 for general attractiveness, 23 for pragmatic quality and 16 for hedonic quality, as shown in Table 28.

UX Dimension	**Description**	**Present in**
Attractiveness		
Desirability	Unpleasant – Pleasant	5 Es SUS SUMI PSSUQ UMUX AttrakDiff UEQ
Enjoyment	Annoying – Enjoyable	UEQ
Engagement	Disagreeable – Likeable	AttrakDiff
User Friendliness	Unfriendly – Friendly	UEQ
Affinity	Unlikeable – Pleasing	UEQ

Hospitality	Rejecting – Inviting	AttrakDiff
Visual Appeal	Ugly – Attractive (AttrakDiff) Unattractive – Attractive (UEQ)	AttrakDiff UEQ
User Motivation	Discouraging – Motivating (AttrakDiff) Demotivating – Motivating (UEQ)	AttrakDiff UEQ
Overall Satisfaction	Bad – Good	ISO 9241 Nielsen (1994) PSSUQ CATTUM AttrakDiff UEQ
Overall Attractiveness	Repelling – Appealing	AttrakDiff
Pragmatic quality		
Effectiveness	Ineffective – Effective	ISO 9241 5 Es TAM CATTUM
Efficiency	Inefficient – Efficient	ISO 9241 Nielsen (1994) 5 Es TAM SUMI PSSUQ ASQ CATTUM UEQ UEQ-S
Practicality	Impractical – Practical	AttrakDiff UEQ
Usefulness	Useful – Useless	Nielsen (1994) TAM
Ease of Use	Cumbersome – Straightforward	TAM SUS SUMI PSSUQ UMUX

		UMUX-LITE ASQ SEQ AttrakDiff
Simplicity	Complicated – Simple (AttrakDiff) Complicated – Easy (UEQ, UEQ-S)	SUS PSSUQ AttrakDiff UEQ UEQ-S
Learnability	Difficult to learn – Easy to learn	Nielsen (1994) 5 Es TAM SUS QUIS SUMI PSSUQ CATTUM UEQ
Memorability	Difficult to remember – Easy to remember	Nielsen (1994) CATTUM
Understandability	Not understandable – Understandable	UEQ
Support	Obstructive – Supportive	UEQ UEQ-S
Documentation	Well documented – Not well documented	QUIS PSSUQ ASQ CATTUM
Predictability	Unpredictable – Predictable	SUS AttrakDiff UEQ
Clarity	Confusing - Clearly structured (AttrakDiff) Confusing – Clear (UEQ, UEQ-S)	AttrakDiff UEQ UEQ-S
Orderliness	Cluttered – Organised	PSSUQ UEQ
Error resistance	Error prone – Error resistant	Nielsen (1994)

		5 Es
Error recovery	Easy error recovery – Difficult error recovery	Nielsen (1994) 5 Es QUIS CATTUM UMUX
Control	Unruly – Manageable	SUMI AttrakDiff
Human Centricity	Technical – Human	AttrakDiff
Completeness	Does not meet expectations – Meets expectations	PSSUQ UMUX UMUX-LITE UEQ
Adaptability	Rigid – Adaptable	CATTUM
Interoperability	Compatible – Incompatible	CATTUM
Speed	Slow – Fast	UEQ
Security	Not secure – Secure	UEQ
Hedonic quality		
Professionalism	Unprofessional – Professional	AttrakDiff
Appearance	Unpresentable – Presentable	AttrakDiff
Social Integration	Isolating – Connective	AttrakDiff
Collaboration	Separates me - Brings me closer	AttrakDiff
Inclusivity	Alienating – Integrating	AttrakDiff
Aesthetic Elegance	Tacky – Stylish	AttrakDiff
Perceived Value	Cheap – Premium (AttrakDiff) Inferior – Valuable (UEQ)	AttrakDiff UEQ
Engagement Appeal	Dull – Captivating (AttrakDiff) Boring – Exciting (UEQ, UEQ-S)	AttrakDiff UEQ UEQ-S
Interest	Not interesting – Interesting	UEQ UEQ-S
Stimulation	Undemanding – Challenging	AttrakDiff
Sophistication	Ordinary – Novel	AttrakDiff

Originality	Conventional – Inventive	AttrakDiff UEQ UEQ-S
Innovativeness	Conservative – Innovative	AttrakDiff UEQ
Novelty	Usual – Leading edge	UEQ UEQ-S
Creativity	Unimaginative – Creative (AttrakDiff) Dull – Creative (UEQ)	AttrakDiff UEQ
Adventurousness	Cautious – Bold	AttrakDiff

Table 28: List of extracted UX dimensions.

However, several overlaps were identified. For example, the items *Impractical – Practical* (AttrakDiff and UEQ), *Complicated – Easy* (UEQ-S), *Complicated – Simple* (AttrakDiff), *Not understandable – Understandable* (UEQ), *Cumbersome – Straightforward* (AttrakDiff), *Easy to use* (SUS), *Simple to use* (PSSUQ) all revolve around the ease of use. Since UX dimensions were extracted with a view to asking users to rate their importance via the survey, a list of 49 items, comprising overlaps, was not deemed acceptable. Therefore, several dimensions were merged for the sake of simplicity. This was done following a careful review of the dimensions presented in Table 28 and their corresponding definitions and merging dimensions when the definitions were sufficiently close and when the item was comprised in different models. For example, the AttrakDiff item of *Engagement (Disagreeable – Likeable)* was considered reasonably close to the UEQ item of *Affinity (Unlikeable – Pleasing)*. Admittedly, minor differences can be identified between these items (*Engagement* in AttrakDiff emphasises a broader emotional connection and engagement with the tool, potentially reflecting how well the product resonates with users, and *Affinity* in UEQ focuses specifically on the pleasing nature of the product, which might lean more toward immediate emotional satisfaction). While these subtle differences are acknowledged, the primary aim of this exercise was to identify core UX dimensions from the literature.

This resulted in a list of 21 dimensions containing most items in the models analysed. Table 29 provides an overview of the resulting list, including the elements from other models encompassed in each item. It should be noted that some models, such as SUMI and QUIS, require a paid licence, and were therefore excluded from the final comparison.

	ISO 9241	Nielsen	5 Es	TAM	SUS	PSSUQ	UMUX	UMUX-LITE	ASQ	SEQ	AttrakDiff	UEQ	CATTUM
Enjoyment	✓		✓									✓	
Ease of use				✓	✓	✓	✓	✓	✓	✓	✓	✓	
Motivation											✓	✓	
Effectiveness	✓		✓	✓									✓
Efficiency	✓	✓	✓	✓		✓			✓			✓	✓
Usefulness				✓		✓							
Learnability		✓		✓	✓	✓						✓	✓
Memorability		✓											✓
Documentation						✓			✓				✓
Clarity				✓		✓					✓	✓	
Error resistance		✓	✓										
Error recovery		✓	✓			✓	✓						
Human centricity												✓	
Adaptability													✓
Interoperability													✓
Speed												✓	
Security												✓	
Visual appeal						✓					✓	✓	
Collaboration											✓		
Intellectual engagement											✓	✓	
Innovation											✓	✓	

Table 29: Comparison of UX dimensions.

As for the previous section, for each item, respondents were asked to rate the importance on a 5-point Likert scale. In addition, an open-ended question was included at the end of this list to let respondents

indicate if any other aspect should be considered when assessing the PE environment. The survey questions of the second PE UX subsection are presented hereafter:

Please indicate how important the following usability aspects are to you in a post-editing environment.

- ***Enjoyment**: I find the post-editing environment pleasant to use.*
- ***Ease of use**: I find the post-editing environment easy to use.*
- ***Motivation**: the post-editing environment motivates me to complete my tasks.*
- ***Effectiveness**: the post-editing environment allows me to achieve my goals with accuracy and completeness.*
- ***Efficiency**: the post-editing environment allows me to achieve my goals efficiently (using minimum resources and effort).*
- ***Usefulness**: the post-editing environment makes my work easier.*
- ***Learnability**: the post-editing environment is easy to learn.*
- ***Memorability**: the functionalities of the post-editing environment are easy to remember.*
- ***Documentation**: the post-editing environment's documentation (user guides, tutorials) is helpful to solve my doubts.*
- ***Clarity**: the post-editing environment's structure is clear and well-organised.*
- ***Error resistance**: the post-editing environment prevents me from making mistakes, such as misusing a feature.*
- ***Error recovery**: the post-editing environment allows to easily undo mistakes.*
- ***Human centricity**: the post-editing environment is designed with a strong focus on user needs, ensuring a human-centric experience.*
- ***Adaptability**: the post-editing environment is adaptable, allowing for customisation when needed.*
- ***Interoperability**: the post-editing environment is well integrated with other software.*
- ***Speed**: the post-editing environment is fast.*
- ***Security**: the post-editing environment is secure (it keeps data secure and prevents unauthorised access).*
- ***Visual appeal**: the post-editing environment has an attractive design.*
- ***Collaboration**: the post-editing environment facilitates collaboration with other users.*
- ***Intellectual engagement**: the post-editing environment is exciting, it captivates my interest and keeps me engaged.*
- ***Innovation**: the post-editing environment showcases innovative features which set it apart from conventional solutions.*

Is there any other aspect which you think should be considered when assessing a post-editing environment? (Optional, open-ended)

9.3.4 Analysis and Discussion

The details of the survey distribution and the characteristics of respondents were already presented in 9.1. This section therefore focuses on the analysis of data not already included in 9.1 and is organised based on the main categories identified in the objectives, namely the analysis of PE needs, then UX dimensions, and finally the design of a new PE UX evaluation model.

A. Importance of PE needs

RQ 1.a: Which PE needs are the most important to translation technology users?

The responses to the survey allow for determining which PE needs are deemed the most important, and whether the differences in ranking are statistically significant. To that end, the average ranks and results of a Friedman test are reported in Table 30. The Friedman test was selected for its suitability to assess three or more related groups or conditions. This test helps determine whether there are significant differences between these groups when the same subjects are measured multiple times under different conditions, and it is used when the data is not normally distributed. It works by ranking the data and then comparing the obtained ranks to detect if any differences are statistically significant. It should be noted here that the Friedman test assumes ordinal rankings and does not quantify the magnitude of differences between ranks. It also does not specify which groups differ from each other; it only identifies whether differences exist. Thus, while it can be stated that certain needs are prioritised over others, the exact distance between ranks should be interpreted cautiously.

PE need	**Average rank**
Completeness	4.20
Productivity	4.09
Decision making	4.06
Guidelines	3.99
Control	3.89
Cognitive load	3.86
MT error handling	3.84
$\chi 2$ (Friedman Test Statistic)	26.96
p-value	1.48e-4

Table 30: Average ranks and Friedman test of rank significance for PE needs.

The ranked preferences of respondents show that *completeness*, *productivity* and *decision making* received the highest average ranks. The Friedman test yielded a significant result ($\chi 2 = 26.96, p < 0.001$), indicating that respondents' rankings of PE needs were not uniform. In other words, respondents do not uniformly prioritise all PE needs, and some needs are considered more important than others to a statistically significant degree.

While the Friedman test confirmed that respondents prioritised certain PE needs over others, it did not reveal which specific needs were ranked differently. To address this, post-hoc pairwise comparisons were performed using the Wilcoxon signed-rank test with a Bonferroni correction. This method compares every possible pair of PE needs while adjusting for the increased risk of false positives that arises when making multiple comparisons. Three statistically significant differences emerged (all $p < 0.0024$, Bonferroni-adjusted), all for pairs involving *completeness* – more specifically:

- *Completeness* ↔ *MT error handling*
- *Completeness* ↔ *Cognitive load*
- *Completeness* ↔ *Control*

No other pairs showed significant differences. This finding suggests that *completeness* is uniquely prioritised as a non-negotiable need in the PE environment, surpassing the importance of other PE needs. In addition, although *productivity*, *decision making*, and *guidelines* were highly ranked, these elements were not statistically distinguishable from one another or from lower-ranked needs like *control*. This implies that users perceive them as similarly important in practice. This lack of significant differences among other needs suggests that user preferences are more nuanced or context-dependent. However, from the software development perspective, this result calls developers to prioritise the completeness of translation tools.

RQ 1.b: What is the relationship between PE needs and the current satisfaction with tools?

It also appears relevant to assess whether a correlation exists between PE needs and satisfaction with tools. To that end, a Spearman rank test was performed to explore the relationship between PE needs and the overall satisfaction. This test is a statistical method used to measure the strength and direction of the relationship between two variables. Unlike traditional correlation tests, it works with ranked data, making it suitable for situations where the data does not follow a normal distribution or when the relationship between variables is monotonic (one consistently increases or decreases as the other changes).

Variable 1 (PE need)	Variable 2	Spearman's ρ	*p*-value
Control	Satisfaction	0.14	0.11
Decision making	Satisfaction	0.13	0.14
Completeness	Satisfaction	0.12	0.17
Productivity	Satisfaction	0.08	0.40
Cognitive load	Satisfaction	0.08	0.38
Guidelines	Satisfaction	0.07	0.42
MT error handling	Satisfaction	0.01	0.87

Table 31: Relationship between PE needs and satisfaction with tools (Spearman test).

As shown in Table 31, all correlations were weak in magnitude ($\rho < 0.15$), following Cohen's (1988) guidelines where $\rho < 0.10$ is considered trivial and 0.10-0.30 as weak. Furthermore, none was found to be statistically significant (all with $p > 0.05$). The strongest associations were observed for *control* ($\rho = 0.14$), *decision making* ($\rho = 0.13$), and *completeness* ($\rho = 0.12$), which were positively correlated with *satisfaction*, though even these relationships remain weak in practice. These results indicate that despite the detectable statistical relationship, users' prioritisation of PE needs has a minimal association with their actual satisfaction levels with current tools. This may suggest that other unmeasured factors (e.g., workflow integration, compensation models) play larger roles in determining satisfaction.

B. Importance of UX aspects

RQ 2.a: Which UX aspects are crucial for language professionals performing PE?

The ranked preferences revealed that *usefulness* (rank = 1, mean rank = 4.294), *speed* (rank = 2, mean rank = 4.262), and *effectiveness* (rank = 3, mean rank = 4.246) were considered the most important by respondents (Table 32). Overall, the top 10 UX aspects (all with a mean rank ≥ 4) reveal that respondents tend to attribute a higher importance to pragmatic dimensions. Another notable finding is that the UX aspects which were considered the most important seem to be related to productivity: *usefulness* has to do with the capacity of the tool to make the user's work easier, *speed* is about the responsiveness of the tool, and *efficiency* revolves around minimising the resources and effort (which can include time) required to complete a task. It should also be argued that some dimensions with lower ranks can still be indirectly related to some of the top-ranking ones. For example, *documentation* can be associated with *learnability*. Similarly, *motivation* and *intellectual engagement* can be connected to *enjoyment*.

UX Aspect	Rank	Average rank
Usefulness	1	4.29
Speed	2	4.26
Effectiveness	3	4.25
Efficiency	4	4.23
Ease of use	5	4.22
Memorability	6	4.09
Error recovery	7	4.07
Learnability	8	4.06
Security	9	4.03
Clarity	10	4
Human centricity	11	3.89
Error resistance	12	3.82
Enjoyment	13	3.79
Adaptability	14	3.79
Interoperability	15	3.77
Documentation	16	3.67
Motivation	17	3.66
Innovation	18	3.37
Collaboration	19	3.33
Intellectual engagement	20	3.21
Visual appeal	21	3.18
$\chi 2$ (Friedman Test Statistic)		421.07
p-value		8.62e-77

Table 32: Average ranks and Friedman test of rank significance for UX aspects.

To find out whether these differences were statistically significant, a Friedman test was performed and confirmed significant variation in rankings across UX aspects (χ^2 = 421.07, p <0.0001). Despite statistical significance, the practical implications of these rankings should be interpreted cautiously. While the Friedman test indicated non-uniform preferences, the mean rank differences between adjacent items (e.g., *usefulness* [4.29] vs *speed* [4.26]) are marginal, suggesting that the priorities attributed by respondents to the different dimensions were nuanced.

To identify which specific aspects were prioritised differently, post-hoc Wilcoxon signed-rank tests with a Bonferroni correction were conducted (210 pairwise comparisons). The results showed that the top-ranked aspects (namely, *usefulness*, *speed* and *effectiveness*), which are of a pragmatic nature, were consistently rated higher than hedonic dimensions like *visual appeal* and *intellectual engagement* (e.g. *usefulness* ↔ *visual appeal*: p = 1.01e-14). No significant differences were found between the top

10 aspects (e.g., *usefulness* ↔ *speed*, $p = 0.60$), suggesting they are perceived as being of a similar importance. Therefore, it appears that post-editors prioritise task-oriented UX aspects (e.g., *speed*, *error recovery*) over aesthetic or engagement-focused features. (e.g. *visual appeal*). *Enjoyment* (ranked 13th) was the only hedonic aspect to outperform *visual appeal* ($p = 3.39e-09$), which can be explained by its indirect link to productivity (e.g., reducing fatigue).

RQ 2.b: What is the relationship between UX aspects and the current satisfaction with tools?

Furthermore, it appears relevant to explore the relationship between the importance given to usability aspects and satisfaction with tools for PE. Here, a Spearman rank test was performed to ascertain whether users' satisfaction with the tools and the significance of usability characteristics are significantly correlated (Table 33).

Variable 1 (UX aspect)	**Variable 2**	**Spearman's ρ**	***p*-value**
Visual appeal	Satisfaction	0.22	0.01
Enjoyment	Satisfaction	0.13	0.15
Intellectual engagement	Satisfaction	0.11	0.21
Collaboration	Satisfaction	0.09	0.29
Error resistance	Satisfaction	0.07	0.44
Motivation	Satisfaction	0.05	0.55
Innovation	Satisfaction	0.05	0.61
Security	Satisfaction	0.03	0.77
Error recovery	Satisfaction	0.02	0.81
Efficiency	Satisfaction	0.01	0.91
Learnability	Satisfaction	0.01	0.95
Clarity	Satisfaction	0	0.95
Ease of use	Satisfaction	-0.01	0.91
Usefulness	Satisfaction	-0.01	0.99
Memorability	Satisfaction	-0.01	0.92
Documentation	Satisfaction	-0.01	0.93
Human centricity	Satisfaction	-0.01	0.93
Effectiveness	Satisfaction	-0.02	0.80
Speed	Satisfaction	-0.03	0.73
Interoperability	Satisfaction	-0.04	0.69
Adaptability	Satisfaction	-0.07	0.45

Table 33: Relationship between importance given to UX aspects and satisfaction with tools (Spearman test).

The correlations between UX aspects' importance and satisfaction ranged from ρ = -0.07 to 0.22, all falling below the conventional threshold for weak effects (ρ < 0.30; Cohen [1988]). Only one statistically significant relationship (p = 0.01) can be observed in this data: the one between visual appeal and satisfaction, yet the strength of this relationship remains weak (ρ = 0.22). On the one hand, *ease of use*, *usefulness*, *memorability*, *documentation*, *human centricity*, *effectiveness*, *speed*, *interoperability* and *adaptability* show a weak negative correlation with the overall satisfaction. This suggests that when these UX characteristics increase, the satisfaction tends to decrease slightly, which in turn may indicate that these aspects could be improved in current tools. On the other hand, *visual appeal*, *enjoyment*, *intellectual engagement*, *collaboration*, *error resistance*, *motivation*, *innovation*, *security*, *error recovery*, *efficiency* and *learnability* exhibit a weak positive correlation with the satisfaction. This finding indicates that when these UX characteristics increase, the satisfaction tends to increase slightly.

C. Additional feedback

RQ 2.c: Are there any other UX aspect(s) which are essential to post-editors?

In addition, 17/126 (13.5%) respondents provided additional insights to the optional open-ended question "*Is there any other aspect which you think should be considered when assessing a post-editing environment?*". Below is a list of themes mentioned extracted from the answers:

- **Centralisation**: gathering all the necessary information in the tool itself ("*ensuring that all the information is available at a glance to make quick and effective translation decisions*" and "*convenient access to client terminology, website, resources, examples and specifications*").
- **Customisation**: users should be able to set up the interface to match their preferences ("*visual appearance should be customisable [fonts, font size, size and position of different sections on the screen]*") and to activate or deactivate functionalities as needed ("*the ability to toggle features on and off*").
- **Visualisation**: making all relevant information available without interfering too much with the main task. In that regard, a respondent pointed out that "*making any error flags clearly visible but not intrusive at the same time is something that's rarely achieved*".
- **Interoperability**: easy integration with external tools or resources (such as MT systems, QA tools, plug-ins).
- **Supported languages**: number of languages that can be handled in the tool.
- **Source quality**: the quality of the original text may affect the PE experience. A respondent indeed indicated that "*high-quality source texts can greatly facilitate the post-editing process,*

while poor-quality source material may require more time and effort to edit effectively", and another one advised for pre-editing "*to create files that respond well to MT, improve PE output and ease the editor's task*".

- **MT quality**: the quality of the MT output is also important when it comes to the PE effort.
- **Continuous retraining**: implementing models which are capable of learning from human input on the fly ("*the machine should be able to learn from corrections*").
- **Productivity**: features that allow for working faster, including "*propagation that works, ability to change 1 or more term in the whole project, QA features that don't produce false negatives*".
- **Project specifications**: certain project-specific factors (e.g. "*deadlines, client expectations, and the tools and resources available*") also heavily impact the PE experience.
- **Cost**: price of the tool.

From those themes, some are directly related to the UX dimensions extracted for this study. For example, *interoperability* was already listed in this section, and *customisation* appeared in the definition of *adaptability*. Others are also implicitly connected. For example, *centralisation* may be linked to *effectiveness* and *usefulness*; *visualisation* has to do with *clarity* and *visual appeal*; and *productivity* is related to *efficiency*. Similarly, *continuous retraining* could be encompassed in *human centricity*, as it involves that the technology learns from and adapts to human actions. Some topics are dependent on the texts themselves (*source quality* and *MT quality*) and their impact on the task at hand. Finally, the remaining topics have a more practical dimension (e.g. *supported languages, cost*) and are related to the working conditions of linguists. For example, the *project specifications* may add extra pressure if the deadline is tight.

9.3.5 Resulting PE UX Questionnaire

The results reported above allowed for designing a new questionnaire to evaluate the UX of post-editors working with translation tools. The questionnaire adopts the format and scoring system of the SUS, but with content adapted to evaluate the UX of post-editors in CAT tools. This format was selected over other alternatives (e.g., UEQ, AttrakDiff) because of its brevity (10 items) and simplicity for respondents (as opposed to semantic differential formats, which likely require a higher cognitive effort).

The items to be included in the questionnaire were selected primarily based on the top 10 UX aspects (i.e., all aspects with a rank ≥ 4 in Table 32). While the top 10 aspects were of a pragmatic nature, it was judged relevant to include one item of a hedonic nature. Therefore, *security* was removed and replaced with *engagement*. This choice was motivated by several factors. While *security* ranked moderately (ranked 9th, mean rank = 4.032, Table 32), it showed a weak practical impact, with a near-zero correlation with satisfaction ($\rho = 0.03$, Table 33). Conversely, *enjoyment* (ranked 13th, mean

rank = 3.794, Table 32) demonstrated a stronger relationship with satisfaction (with ρ = 0.13 being the second highest value, Table 33). Furthermore, *security* is often a binary requirement (either present or absent), thus leaving little room for UX gradation.

Each item comes with the name of the UX aspect as well as a brief description based on the corresponding definition, adapted to match how SUS items are formulated. All statements start by "I" and positive and negative statements are alternated (odd: positive, even: negative), but negative statements do not use negation ("not"). For each item, the answer is a number on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). The resulting questionnaire is as follows:

1. ***Usefulness****: I think the post-editing environment makes my work easier.*
2. ***Speed****: I find the post-editing environment slow.*
3. ***Effectiveness****: I feel that the post-editing environment allows me to achieve my goals with accuracy and completeness.*
4. ***Efficiency****: I need to use a high amount of resources and effort to achieve my goals when working in the post-editing environment.*
5. ***Enjoyment****: I find the post-editing environment pleasant to use.*
6. ***Ease of use****: I find the post-editing environment complex to use.*
7. ***Memorability****: I think that the functionalities of the post-editing environment are easy to remember.*
8. ***Error recovery****: I find that it is difficult to undo mistakes in the post-editing environment.*
9. ***Learnability****: I think that the post-editing environment is easy to learn.*
10. ***Clarity****: I find the post-editing environment's structure unclear and disorganised.*

The scoring model is also based on the SUS logic, and the final score is obtained by following these steps:

- For odd-numbered questions: [User Rating] – 1 = ___ points
- For even-numbered questions: 5 – [User Rating] = ___ points
- Add all points
- Multiply total by 2.5 (to convert scale to 100)
- Average all user's scores together

A mock-up is provided in Table 34 for illustration purposes.

Question	Rating
1. Usefulness: I think the post-editing environment makes my work easier	4
2. *Speed*: I find the post-editing environment slow	5
3. *Effectiveness*: I feel that the post-editing environment allows me to achieve my goals with accuracy and completeness	3
4. *Efficiency*: I need to use a high amount of resources and effort to achieve my goals when working in the post-editing environment	2
5. *Enjoyment*: I find the post-editing environment pleasant to use	4
6. *Ease of use*: I find the post-editing environment complex to use	1
7. *Memorability*: I think that the functionalities of the post-editing environment are easy to remember	5
8. *Error recovery*: I find that it is difficult to undo mistakes in the post-editing environment	2
9. *Learnability*: I think that the post-editing environment is easy to learn	5
10. *Clarity*: I find the post-editing environment's structure unclear and disorganised	2
SCORE	**72.5**

Table 34: Simulation of the PE UX questionnaire.

As far as interpretation is concerned, the same threshold as for SUS is suggested (i.e., benchmark of 68). For a more fine-grained analysis, it is however recommended to analyse individual dimensions more closely. More specifically, average scores of 4-5 for even and 1-2 for odd items could reveal weaknesses in the system under evaluation.

The proposed questionnaire was then validated in a pilot study involving 12 in-house linguists working at LanguageWire, with diverse professional and linguistic profiles (Table 35), who were asked to post-edit a marketing text before taking the questionnaire.

Category	Details
Language pairs	EN>FR (3), EN>ES (2), EN>DE (2), EN>NL (2), EN>IT (1), EN>PT (1), EN>SV (1)
Experience in translation	5-35 years in the industry; 5 with Master's degrees; 6 with Bachelor's degree or equivalent diploma in translation or related disciplines
Experience in PE	6 limited; 3 occasional; 3 regular; 8 receive training (university or industry)
Experience in marketing	9 occasional exposure; 3 regular exposure
Trados Studio proficiency	6 advanced; 4 intermediate; 2 minimal

Table 35: Overview of participants' background.

Participants covered seven language combinations, with the most frequent being English to French (three participants), followed by English to Spanish and English to German (two participants each). The remaining language pairs (English to Italian, Dutch, Portuguese, and Swedish) were each represented by a single participant. In terms of translation experience, participants ranged from early-career linguists (5 years) to seasoned professionals with over 35 years in the field. The majority held university degrees, including five with Master's degrees in translation, interpreting, or linguistics, while six had Bachelor's degrees or equivalent diplomas in translation-related disciplines. One participant specialised in linguistics without receiving formal translation training. Regarding PE, three participants had substantial experience, often supported by formal training (e.g., European Commission workshops or company-led programmes); three others had occasional exposure, while three reported limited involvement in PE tasks. Exposure to marketing translation was occasional for most participants: nine reported periodic exposure, and three had regular experience in this field. Familiarity with Trados Studio also varied: six linguists were long-term users (10+ years), employing the tool regularly for translation or TM management; four used it occasionally, and two had minimal exposure. Formal training for Trados Studio was rare; most proficiency was self-taught or acquired on the job. The limited exposure for some of the participants can be explained by the use of LanguageWire's proprietary CAT tool, Smart Editor.

To evaluate the reliability of the proposed PE UX questionnaire, a controlled pilot study was conducted with the 12 linguists. Each one post-edited a 517-word marketing text using Trados Studio. For each participant, the task completion time was recorded, which allowed for calculating productivity as the number of source words processed per hour. Immediately following the task, participants completed the proposed questionnaire. The questionnaire's internal consistency was assessed using Cronbach's alpha, a widely used metric for assessing the interrelatedness of Likert-scale items within a survey (Cronbach, 1951). A high alpha value (closer to 1) suggests that the items consistently measure the same underlying construct, reinforcing the questionnaire's validity. It yielded a score of $\alpha = 0.706$, which falls slightly above the threshold of 0.70 for acceptable reliability in exploratory research (Nunnally, 1978; Tavakol and Dennick, 2011), indicating that the items cohesively measure the underlying UX construct. It could be hypothesised that replicating the study with a larger sample comprising participants with more similar backgrounds could yield a higher alpha value.

To explore the relationship between UX scores and productivity (Table 36), Pearson correlation was used, as it is a suitable method to measure the relationship between two continuous variables, assuming both are normally distributed and linearly related. The analysis revealed a moderate negative correlation ($r = -0.30$), which was not statistically significant ($p = 0.338$). This suggests that UX may not linearly predict efficiency in PE. This aligns with qualitative feedback from participants who noted that working with creative texts often required slower, more deliberate editing despite appreciating the CAT tool's features. However, the small sample size limits definitive conclusions. Therefore, future work should validate these trends with larger samples.

Participant ID	UX score	Productivity (WPH)
P1	90.0	326.53
P2	82.5	553.93
P3	92.5	470.00
P4	82.5	608.24
P5	75.0	430.83
P6	75.0	633.06
P7	77.5	316.53
P8	57.5	775.50
P9	65.0	373.73
P10	70.0	517.00
P11	60.0	408.16
P12	62.5	544.21

Table 36: UX scores and productivity for each participant.

Furthermore, respondents unanimously confirmed the questionnaire's comprehensiveness, with multiple participants explicitly stating they found it complete enough for evaluating the PE environment. However, qualitative comments revealed two important contextual factors influencing responses: several participants noted their limited recent experience with Trados Studio or identified as unaccustomed users, while one emphasised how their lack of formal training affected all ratings. These observations suggest that while the instrument adequately covers essential UX elements, future implementations may benefit from including user expertise and training history as demographic variables to better contextualise evaluation outcomes. However, since no substantive omissions in UX coverage were identified, the proposed questionnaire can be considered valid for assessing the PE environment across different user proficiency levels.

9.3.6 Summary and Outcomes

The findings of this study shed light on the PE needs considered important among post-editors, within a sample of 126 respondents. *Completeness*, *productivity*, and *decision-making* were prioritised in respondents' ratings, though with limited correlations with satisfaction. Furthermore, the alignment between user requirements and tool functionality is critical in enhancing the overall experience and effectiveness of translation technologies. In that regard, the results also revealed a weak positive correlation between PE needs and user satisfaction with current translation tools. In addition, *usefulness*, *speed*, and *effectiveness* emerged as the most important UX aspects influencing the experience of post-editors. Additionally, several other UX aspects not initially included in the analysis were identified as essential for post-editors, such as centralisation, customisation, visualisation, interoperability, supported

languages, source and MT output quality, continuous retraining, productivity, project specifications, and cost. While some of these aspects are directly related to the PE environment itself (e.g. customisation and visualisation), these responses also showed that the experience of post-editors is also affected by broader factors beyond the tool itself (e.g. source quality and project specifications).

These findings motivated the design of a novel framework which accurately reflects the needs and views of post-editors. A new PE-specific questionnaire was created and tested with 12 professional linguists in a pilot study, which allowed for validating it. Although further testing is needed, this method can be used in user-centred studies of the PE environment. Furthermore, integrating this method within an extended evaluation framework that includes multiple pillars is essential for a comprehensive benchmarking of PE environments.

9.4 Quality Pillar Definition

This section describes the process behind the definition of the Quality dimension of the BF. It explores two key aspects: the relationship between MT quality and PE effort, and the quality of the final post-edited text. By addressing these dimensions, this section provides a comprehensive study of translation quality in a PE context.

9.4.1 Quality Element 1: MT Quality as Predictor of PE Effort

In light of the findings discussed in 4.4 (Towards New MT Quality Evaluation Methods), establishing a new evaluation model for assessing MT quality based on the PE effort emerges as a relevant approach. This section discusses the rationale behind this work, the proposal for the new evaluation method and outlines the experimental setup designed to test it. It also presents the findings and their implications for the BF.

9.4.1.1 Rationale

The main objective was to examine the relationship between specific types of MT errors and their corresponding PE effort. The aim was to determine whether certain types of MT errors tend to require a higher PE effort and to assess whether these errors align with those identified in 4.4.

More specifically, based on this section of the theoretical framework, it was hypothesised that certain MT errors (reordering, lexical edits, source language interference, mistranslation, grammar and document level) would be perceived as requiring a greater PE effort compared to other errors (word form, removing, function words, locale conventions and typography).

If this assumption holds true, it would have two major implications:

1) it would allow for the creation of a new MT evaluation model based on the PE effort (i.e., after identifying MT errors that systematically entail a low or high effort, a typology could be designed with pre-assigned weights based on the predicted effort, in a similar way that traditional metrics like DQF assign weights based on severity)
2) it would allow for accounting for MT quality as a satellite parameter in the BF to further explain the Productivity dimension (i.e., the number of MT errors encountered, and their associated PE effort would provide further insights into the overall task completion time).

9.4.1.2 Proposed Method to Evaluate MT for PE

In order to conceptualise an error typology in the context of the present work, the three main categories suggested by Ovchinnkova (2020) were adopted (i.e. fluency, accuracy and function).

The error types belonging to each category were identified and pre-assigned an effort level (high or low) based on the above discussion. Therefore, the initial effort distribution was as follows:

- High
    - Reordering
    - Lexical edits
    - Source language interference
    - Mistranslation
    - Grammar
    - Document level
    - Sentence length
- Low
    - Word form
    - Removing additions
    - Function words
    - Locale conventions
    - Typography

For other error types, the effort perception was not clear; therefore, the pre-assigned effort was attributed based on the findings of the exploratory user survey (9.1). If the error was considered difficult for more than one third of the respondents, the effort is considered high. More specifically, the error difficulty for the remaining error types was as follows in the survey results:

- Omission: low effort (considered difficult for 24%)
- Domain mismatch: high effort (since terminology was considered difficult for 33%)
- Punctuation: low (considered difficult for 13%)
- Style: high (was considered difficult for 37%)
- Adherence to guidelines: high (was considered difficult for 40%)
- Sentiment deviation: low (was considered difficult for 13%)

After refining the typology further, the categorisation presented in Table 37 was designed.

Error category	**Error type**	**Definition**	***Examples (EN-US > ES-ES)***
Adequacy: The target text does not accurately reflect the content of the source text.	***Addition***	The MT output includes content that is not present in the source text, resulting in the addition of fictitious or erroneous information in the target text.	*Source: The study examined the effects of climate change on local ecosystems.* *MT: El estudio examinó los efectos del cambio climático en los ecosistemas locales, concluyendo que no hay impacto significativo.* *Post-edited: El estudio examinó los efectos del cambio climático en los ecosistemas locales.*
	Omission	Content which is present in the source is missing in the MT output.	*Source: The manager approved the budget and the project timeline.* *MT: El gerente aprobó el presupuesto.* *Post-edited: El gerente aprobó el presupuesto y el cronograma del proyecto.*
	Lexical error	The MT output contains incorrect or inappropriate word choices, resulting in a target text that is inaccurate or lacks clarity.	*Source: The tools are in the pen.* *MT: Las herramientas están en el bolígrafo.* *Post-edited: Las herramientas están en el corral.*
	Document level	The MT output shows internal inconsistency or coherence throughout the document. This can involve issues with logical flow and structure across sentences and paragraphs, as well as inconsistent terminology, style, or tone.	*Source:* *Section 1: The company policy states that employees must submit their timesheets by Friday.* *Section 2: According to the company policy, employees should turn in their weekly work hours by the end of the week.* *MT:* *Section 1: La política de la empresa establece que los empleados deben presentar sus hojas de tiempo el viernes.*

			Section 2: Según las directrices corporativas, los empleados deben entregar sus horas de trabajo semanales antes de que acabe la semana. *Post-edited:* *Section 1: La política de la empresa establece que los empleados deben presentar sus hojas de tiempo el viernes.* *Section 2: Según la política de la empresa, los empleados deben entregar sus horas de trabajo semanales antes de que acabe la semana.*
	Terminology	In the MT output, a specialised term is translated with a term other than the one expected for the domain or otherwise specified (for example, contained in the Term Base or guidelines).	*Source: The patient's airway was blocked.* *MT: La vía aérea del paciente estaba bloqueada.* *Post-edited: Las vías respiratorias del paciente estaban bloqueadas.*
Fluency: The target text contains errors which affect its readability.	***Typography***	The MT output contains incorrect typography, such as misspellings, punctuation errors, or formatting issues, resulting in a target text that is incorrect or inconsistent with the source text.	*Source: The website URL is www.example.com* *MT: La URL del sitio web es www.examplecom.* *Post edited: La URL del sitio web es www.example.com.*
	Locale convention	The MT output does not adhere to locale-specific conventions and violates requirements for the presentation of content in the target locale. This includes issues with date formats, currency, units of measurement, address formats, idiomatic expressions, and other locale-specific elements.	*Source: The meeting is scheduled for 07/04.* *MT: La reunión está programada para el 07/04.* *Post-edited: La reunión está programada para el 4 de julio.*

	Source language interference	The MT output reflects structures or idiomatic expressions from the source language that do not exist or are not natural in the target language, leading to a target text that is unnatural or difficult to understand.	*Source: The director came up with a brilliant idea.* *MT: El director vino con una idea brillante.* *Post-edited: Al director se le ocurrió una idea brillante.*
	Grammar	The MT output contains issues related to the grammar or syntax of the text (other than spelling).	*Source: I hope that she finishes the project on time.MT: Espero que ella termina el proyecto a tiempo.Post-edited: Espero que ella termine el proyecto a tiempo.*
Function: The target text contains errors that distort its objective or affect its perception in the target culture.	***Sentiment deviation***	The MT output conveys a different sentiment (polarity or intensity) compared with the source text, leading to a target text that misrepresents the intended emotions or attitudes.	*Source: I really cannot praise the staff highly enough.* *MT: Realmente no puedo alabar el personal.* *Post-edited: Realmente no puedo alabar lo suficiente al personal.*
	Cultural element	The MT output contains cultural elements that are inappropriate or meaningless in the target culture.	*Source: Her son scored high on the SAT.* *MT: Su hijo obtuvo una alta puntuación en el SAT.* *Post-edited: Su hijo obtuvo una alta puntuación en el examen de ingreso a la universidad.*
	Non-compliance with guidelines	The MT output does not adhere to specific style, terminology, formatting, length, or other type of guidelines provided for the translation task, leading to a target text that deviates from the desired standards or requirements.	*Guideline: Do not include abbreviations in the translation.* *Source: The patient underwent a coronary artery bypass graft (CABG) surgery.* *MT: El paciente se sometió a una cirugía de injerto de derivación de arteria coronaria (CABG).* *Post-Edited: El paciente se sometió a una cirugía de injerto de derivación de arteria coronaria.*

	Domain mismatch	The MT output is inappropriate for the specified subject matter or context, leading to a target text that does not align with the intended domain or topic.	*Source: Do not leave the machine within the reach of children. MT: No deje la máquina al alcance de los niños. Post-edited: Mantenga el aparato alejado de los niños.*
	Style	The MT output does not conform to the specified or expected writing style or tone, resulting in an inappropriate target text.	*Source: Could you send the project details to the director? MT: ¿Podrías enviar los detalles del proyecto al director? Post-edited: ¿Podría enviar los detalles del proyecto al director?*
Other		Any other issue.	

Table 37: MT error typology with examples.

"Hallucination" was not included as a separate error type, as it is considered that it would be encompassed in either "addition" or "lexical error". However, it should be mentioned that "addition" typically occurs in HT, when the translator purposely inserts additional content in the target text, using certain techniques such as explicitation or compensation. For example, an English translator may render "l'Albufera" as "the Albufera lake", intentionally adding "lake" to provide the target reader with the necessary background to understand the text. Such an approach is not typical of MT; hence it would be more accurate to consider the example provided for the "addition" error type as a case of "hallucination". Nevertheless, "addition" was preferred here, as annotators are more likely to be familiar with this wording.

9.4.1.3 Experimental Setup

This experiment involved five freelance translators who were asked to post-edit and annotate a text. The source text selected for this evaluation is a leaflet for Daktacort, a cream used to treat inflamed sweat rash and Athlete's Foot[49]. The translation was performed from English (UK) to Spanish (Spain).

[49] https://www.boots.com/wcsstore/cmsassets/Boots/Content/Products/Athletes%20foot%20-%20CAT:%20A00000330/10033188.P/Daktacort%20Hydrocortisone%20Cream%20PIL%5B1%5D.21796.2.pdf (consulted: 30/04/2025)

The selected text falls within the medical domain, as it is a medicine leaflet. It was originally produced by a UK-based manufacturer, which indicates that the text was originally written in English. With a length of 1,368 words, the text was chosen in its entirety to address document-level phenomena, such as coherence, which are critical when evaluating translation quality. Further characteristics of this text are presented in Table 38. Two measures of lexical diversity are included to provide an indication of the complexity of the selected text. The first one is the Type-Token Ratio (TTR; Johnson, 1944), calculated by dividing the number of unique words (types) by the total number of words (tokens) in a text (this ratio gives a number between 0 and 1, with higher values indicating greater lexical diversity). While TTR is sensitive to text length, the second measure used here, the Measure of Textual Lexical Diversity (MTLD; McCarthy and Jarvis, 2010), addresses this limitation. MTLD calculates the mean length of sequential word strings in a text that maintain a given TTR value (usually 0.72). Higher MTLD values indicate greater lexical diversity. Both TTR and MTLD were calculated using Reuneker (2017).

Domain	Medical
Number of segments	126
Number of words	1,368
Number of characters	6,724
TTR	0.3
MTLD	58.94

Table 38: Source text characteristics (Daktacort).

The Vademecum website[50] was checked to ensure that the selected medicine was not commercialised in Spain. This precaution was necessary to prevent any potential bias in the results that might arise if the MT system had been trained on the translation of this leaflet. Furthermore, it should be noted that both the language combination and the domain represent common translation requests processed at LanguageWire. The MT output was produced by LanguageWire's MT system, an in-house model based on neural technology and regularly updated with customer data for customisation purposes. To ensure the authenticity and validity of the results, no artificial errors were introduced into the output. This approach aims to reflect the engine's performance under realistic conditions.

[50] https://www.vademecum.es/buscar?q=daktactort&cc=es (consulted: 30/04/2025)

Before the experiment, two LanguageWire in-house linguists participated in a pilot study to ensure the clarity of instructions for the annotation task. The guidelines were improved based on their feedback prior to the main experiment.

For the main experiment, five freelance linguists working for LanguageWire, experienced in PE and with native-level proficiency in Spanish, were recruited. All participants were compensated based on their agreed hourly rate, and the estimated time set to complete the task was three hours. They were not informed that the PE task was requested as part of an experiment; they were asked to post-edit the text as usual, with the extra task of annotating.

The annotators were provided with a briefing that included overall instructions for the annotation task, including definitions of the error types and examples (see Table 37). In addition to the definitions, the following additional guidelines were provided for errors which could be ambiguous to classify:

- Document level: *if you encounter several occurrences of the same error, please annotate each one, specify which is the first and which are the related ones in the Comments column. For terminological and stylistic errors that have to do with the document level, please use this category instead of "Terminology" or "Style".*
- Terminology: *if you encounter inconsistent terminology at the document level, please classify such errors as "Document-level error".*
- Style: *if you encounter inconsistent style at the document level, please classify such errors as "Document-level error".*
- Other: *please specify which error you encountered in the Comments column.*

Definitions of error severities and editing actions were also provided, as in Tables 39 and 40 respectively.

Error severity	Definition
Low	Errors that do not lead to loss of meaning and would not confuse or mislead the user but would be noticed, would decrease stylistic quality, fluency or clarity, or would make the content less appealing.
Medium	Errors that may confuse or mislead the user or hinder proper use of the product/service due to significant change in meaning or because errors appear in a visible or important part of the content.
High	Errors that may carry health, safety, legal or financial implications, or negatively modify/misrepresent the functionality of a product or service, or which could be seen as offensive.

Table 39: Definitions of error severities.

Editing actions	Definition
Deleting	Removing parts of the MT output.
Inserting	Adding new content to the MT output.
Moving	Changing the position of words or phrases within the MT output.
Replacing	Substituting parts of the MT output.

Table 40: Definitions of editing actions.

Annotators were provided with a dedicated file for the annotation. This file contained the following columns: segment ID, source text, target text, post-edited text, MT-PE match (to indicate if the MT output was edited), effort, error type, error severity, editing action, and comments.

They were instructed to post-edit the MT output when necessary, and to specify the corresponding effort, error type, error severity and editing action. If there were multiple errors in a segment, they were asked to annotate the one they deemed the most important, and to specify the other one(s) in the comments column.

9.4.1.4 Research Questions

The primary RQs this experiment sought to answer were:

- RQ 1: How similar or different is the PE experience across various linguists?
    - Did the annotators edit the same segments?
    - What is the inter-annotator agreement?
- RQ 2: Did annotators identify the same errors?
- RQ 3: Did annotators judge severity in a similar way?
- RQ 4: Did annotators experience similar levels of effort?
    - Is the overall effort comparable across the annotators?
    - Is the overall perceived effort consistent with the TER scores?
    - Is there a correlation between the error categories and the effort? In other words, are the error types belonging to specific categories systematically perceived as more difficult to correct than others?
    - Is there a correlation between the error types and the effort? In other words, are certain types of errors systematically perceived as more difficult to correct than others?
    - Is there a correlation between the editing action and the effort? In other words, are certain editing actions systematically perceived as requiring more effort than others?

9.4.1.5 Results and Analysis

This section presents the experimental results and offers a detailed analysis of the findings. It explores key aspects of the PE experience across several dimensions, providing both quantitative and qualitative insights.

A. Inter-Annotator Agreement

The agreement on edited segments (i.e. whether the same segments were edited or not by all annotators) was calculated as the percentage of agreement among annotators. To that end, each segment was analysed to determine how many annotators chose to edit it, as illustrated in Table 41.

The overall agreement on edited segments was found to be 0.14, meaning that only 14% of segments were edited by all annotators (18 segments out of 126), indicating a low consensus among the annotators about which segments required editing. However, the same analysis for segments left unedited by all annotators yielded a higher agreement of 0.27, meaning that 27% of the segments (35 segments out of 126) were not edited by any annotator. Therefore, combining these results, all annotators agreed on the editing decision (editing the MT output or not) for 41% of the segments. This indicates that no agreement was reached on the editing decision for more than half of the text.

Segment ID	Edited by annotators	Agreement score
1	A4	0.20
2	A2, A5	0.40
3	A1, A2, A3, A4, A5	1.00
4	A4	0.20
5	A3, A4	0.40
…	…	…
	Average	0.14

Table 41: Calculation of the agreement on edited segments.

To gain further insight, the pairwise agreement was also calculated to find out the agreement rate between individual annotators. The pairwise agreement matrices presented in Figure 57 below are particularly relevant to identify potential deviations in the work of individual annotators. In the case of the agreement on the editing decision, A3 and A4 seemed to strongly agree (0.8), whereas A4 and A5 have the lowest agreement rate (0.64), and the other values fall within the [0.64-0.8] range, which indicates that there was no specific outlier.

Figure 57: Pairwise agreement (Daktacort).

Moreover, to further explore the differences between annotators, the inter-annotator agreement was also calculated for the error types, the error severity and the editing actions (see Table 42).

Agreement	Score
Error types (Fleiss' kappa)	0.07
Error severity (Krippendorff's alpha)	0.23
Editing actions (Fleiss' kappa)	-0.02
Effort (Krippendorff's alpha)	0.06

Table 42: Agreement on error types, error severity, editing actions and effort.

More specifically, the inter-annotator agreement for error types was assessed using Fleiss' kappa, a measure suitable for categorical data involving more than two annotators. It yielded a value of 0.07, indicating poor agreement among annotators regarding error classification. This suggests that annotators had different perceptions or interpretations of what constitutes a specific error type. To evaluate agreement on error severity, Krippendorff's alpha was used, as it is appropriate for ordinal data. The alpha value of 0.23 reflected a fair – albeit still low – agreement on the perceived severity of errors. This indicates substantial variability in how annotators judged severity. For editing actions, Fleiss' kappa was again applied, yielding a value of -0.02, indicating no agreement among annotators, suggesting that annotators were inconsistent in their choice of editing actions for similar errors. Similarly, the agreement on the effort, calculated using Krippendorff's alpha, was also low, with a value of 0.06. This indicates that annotators had varied perceptions of the effort required to correct segments, regardless of the error types identified. These findings highlight the diverse approaches taken by annotators when correcting errors.

B. Editing Analysis

The distribution of edited segments, editing actions and the TER for each annotator appears in Table 43.

	A1	A2	A3	A4	A5
Total edited segments	46 (36%)	38 (30%)	61 (48%)	60 (47%)	64 (50%)
Deleting	2 (4%)	2 (5%)	0 (0%)	1 (2%)	1 (2%)
Inserting	2 (4%)	3 (8%)	2 (3%)	5 (8%)	1 (2%)
Moving	0 (0%)	0 (0%)	1 (2%)	1 (2%)	1 (2%)
Replacing	43 (91%)	33 (87%)	58 (95%)	53 (88%)	61 (95%)
TER	9.4	8.93	20.2	15.46	14.22

Table 43: Edited segments, editing actions and TER for each annotator.

A5 edited the highest number of segments (50%), yet the associated TER score for this annotator was 14.22, which indicates that they did not edit the highest proportion of the text. A3, in contrast, edited slightly fewer segments (48%), but made more extensive changes to the text, resulting in a higher TER of 20.2. A2, on the other hand, edited the lowest number of segments (30%) and also had the lowest TER score of 8.93. In terms of edit operations, the most frequent action across all annotators was replacing, ranging from 87% to 95% of the total number of edits. Deleting, inserting and moving were observed in significantly lower proportions. These results suggest that at least half of the MT output was acceptable without modification and that the majority of the identified errors required replacements.

C. Error Classification Analysis

The analysis of error classification reveals significant disparity in the errors identified by each annotator (Table 44), which is in line with the low inter-annotator agreement observed.

	A1	A2	A3	A4	A5
ADEQUACY - Addition	1 (2%)	1 (3%)	0 (0%)	0 (0%)	0 (0%)
ADEQUACY - Omission	2 (4%)	0 (0%)	1 (2%)	1 (2%)	0 (0%)
ADEQUACY - Lexical error	5 (11%)	7 (18%)	7 (11%)	48(80%)	36 (56%)
ADEQUACY - Document level	0 (0%)	10(26%)	1 (2%)	1 (2%)	0 (0%)
ADEQUACY - Terminology	8 (17%)	3 (8%)	10 (16%)	0 (0%)	19 (30%)
ADEQUACY (total)	16 (34%)	21 (55%)	19 (31%)	50 (83%)	55 (86%)
FLUENCY - Grammar	0 (0%)	9 (24%)	4 (7%)	2 (3%)	2 (3%)
FLUENCY - Typography	3 (6%)	1 (3%)	0 (0%)	3 (5%)	0 (0%)
FLUENCY - Source language interference	1 (2%)	0 (0%)	35(57%)	0 (0%)	7 (11%)
FLUENCY - Locale convention	1 (2%)	0 (0%)	3 (5%)	3 (5%)	0 (0%)
FLUENCY (total)	5 (11%)	10 (26%)	42 (69%)	8 (13%)	9 (14%)
FUNCTION - Sentiment deviation	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
FUNCTION - Cultural element	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
FUNCTION - Non-compliance with guidelines	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
FUNCTION - Domain mismatch	10 (21%)	4 (11%)	0 (0%)	0 (0%)	0 (0%)
FUNCTION - Style	16 (34%)	3 (8%)	0 (0%)	2 (3%)	0 (0%)
FUNCTION (total)	26 (55%)	7 (18%)	0 (0%)	2 (3%)	0 (0%)

Table 44: Error type classification for each annotator.

On the one hand, the agreement appears to be high for certain error types. In the fluency category, no annotator identified instances of sentiment deviation, cultural elements or non-compliance with guidelines. This may indicate that the MT system handled these elements well, that the source text did not present challenges related to those errors, or that the annotators may not have fully grasped the definitions of these errors. Similarly, three annotators did not report any occurrence of addition or domain mismatch. Other error types, such as omission and typography, also showed relatively consistent distributions across annotators.

On the other hand, notable differences can be spotted in the number of errors identified by each annotator for each error type. For example, while three annotators (A1, A2 and A3) detected only a few lexical errors, the remaining two annotators reported significantly higher numbers (with 36 lexical errors identified by A5 and 48 by A4). Similarly, the annotations for the terminology error type varied greatly: A4 did not annotate any terminological error, whereas this error type accounted for 30% of the errors found by A5. Grammatical errors represented less than 10% for four annotators, but 24% of the errors spotted by A2. Domain mismatch and stylistic errors also exhibited low agreement: while four annotators found little to no occurrence of these errors, they constituted 21% and 34% of the errors identified by A1, respectively.

These findings highlight a key challenge in translation quality assessment – subjectivity. Beyond editing different segments, annotators also disagreed on the nature of the errors in the raw MT output. This task thus faces a dual problem: not only is translation quality inherently subjective, but an additional layer of subjectivity exists in how errors are categorised. As a result, the same error might be classified in numerous ways by different annotators. For example, while A4 and A5 both identified the majority of errors (over 80%) as being related to adequacy, A1's perspective was markedly different, attributing only 34% of errors to adequacy, with 11% classified as fluency errors and 55% as functional errors. A4 also considered most of the errors to be lexical (80% of all errors identified), which may explain why this annotator did not flag any terminological error. This suggests potential overlaps between error categories, further complicating the task of consistent error classification.

D. Error Severity Analysis

Most annotators rated the majority of errors as low in severity (Table 45), with the exception of A1, who more frequently assigned a medium severity rating.

	A1	**A2**	**A3**	**A4**	**A5**
Low	6 (13%)	19 (50%)	26 (43%)	37 (62%)	62 (97%)
Medium	23 (49%)	9 (24%)	23 (38%)	16 (27%)	1 (2%)
High	18 (38%)	10 (26%)	12 (20%)	7 (12%)	1 (2%)

Table 45: Error severity distribution for each annotator.

Furthermore, while A5 edited the highest number of segments, 97% of the errors identified were rated as low severity. In contrast, A1, despite modifying fewer segments overall, identified a significantly higher proportion of high-severity errors (38%). These discrepancies suggest that subjectivity extends not only to error identification but also to the attribution of error severity.

The distribution of severity across the main error categories (Table 46) reveals further discrepancies.

	A1	A2	A3	A4	A5
ADEQUACY					
Low	1	6	6	29	54
Medium	4	7	6	14	0
High	11	8	7	7	1
FLUENCY					
Low	1	9	20	6	8
Medium	3	1	17	2	1
High	1	0	5	0	0
FUNCTION					
Low	4	4	0	2	0
Medium	16	1	0	0	0
High	6	2	0	0	0

Table 46: Error severity distribution by error category for each annotator.

For both A4 and A5, the majority of low-severity errors were classified under the adequacy category, whereas A3 found that low-severity errors were primarily related to fluency.

E. Effort Analysis

Examining the PE effort required by each annotator constitutes one of the central focus points in the present study. The overall effort distribution is presented in Table 47. This table also comprises an effort score, which was calculated by multiplying the number of errors identified by a coefficient corresponding to the perceived effort required to correct each error. The effort categories and their respective coefficients are as follows: low effort (1), medium effort (2) and high effort (5). These coefficients were assigned to reflect the increasing complexity of edits, although it should be noted that the values are to some extent arbitrary and primarily intended to provide a relative measure of the PE effort.

	A1	A2	A3	A4	A5
Low	13 (28%)	26 (68%)	28 (46%)	30 (50%)	61 (95%)
Medium	20 (43%)	8 (21%)	19 (31%)	15 (25%)	2 (3%)
High	14 (30%)	4 (11%)	14 (23%)	15 (25%)	1 (2%)
Effort score	123	62	136	60	71

Table 47: Effort distribution and score for each annotator.

When it comes to the effort annotations, A5 stands out as a notable case, having edited the highest number of segments, yet not having the highest effort score. This is explained by the fact that 95% of the errors A5 identified required low editing effort. In contrast, A3's editing effort was more consistent. Despite editing a high number of segments (48%), A3 had the highest TER score (20.2) and the highest effort score (136), suggesting that PE was most challenging for this individual. The case of A4 illustrates the opposite situation. Although A4 also edited a high number of segments (47%) and had the second highest TER score (15.46), their effort score was the lowest (60). These findings reveal significant variation in how annotators perceived editing effort, underscoring the subjective nature of this dimension.

	A1	A2	A3	A4	A5
ADEQUACY					
Low	4	10	9	24	52
Medium	5	7	5	12	2
High	7	4	5	14	1
FLUENCY					
Low	3	10	19	5	9
Medium	1	0	14	2	0
High	1	0	9	1	0
FUNCTION					
Low	6	6	0	1	0
Medium	14	1	0	1	0
High	6	0	0	0	0

Table 48: Effort distribution by error category and score for each annotator.

Similarly, the distribution of editing effort across error categories shows different patterns across annotators, as illustrated in Table 48. For A4 and A5, most of the low-effort edits were concentrated in the adequacy category. In contrast, A3 and A2 primarily associated low-effort edits with fluency errors, while A1 attributed them mostly to functional errors.

Figures 58-62 provide a more fine-grained view of the effort distribution for the different error types in the case of each annotator.

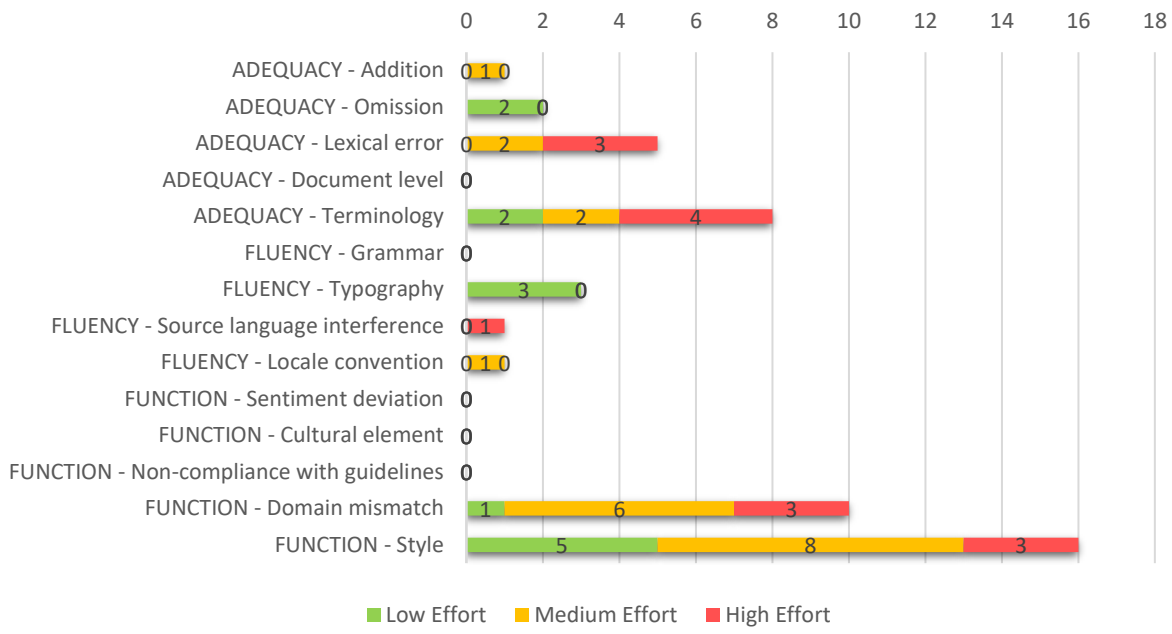


Figure 58: Effort distribution by error category for A1.

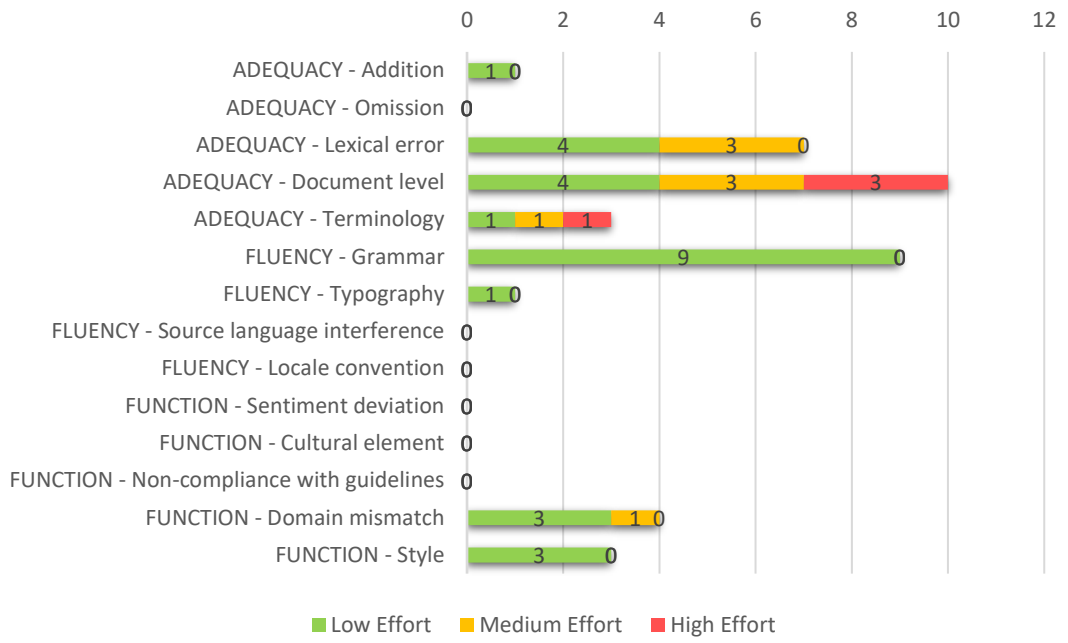


Figure 59: Effort distribution by error category for A2.

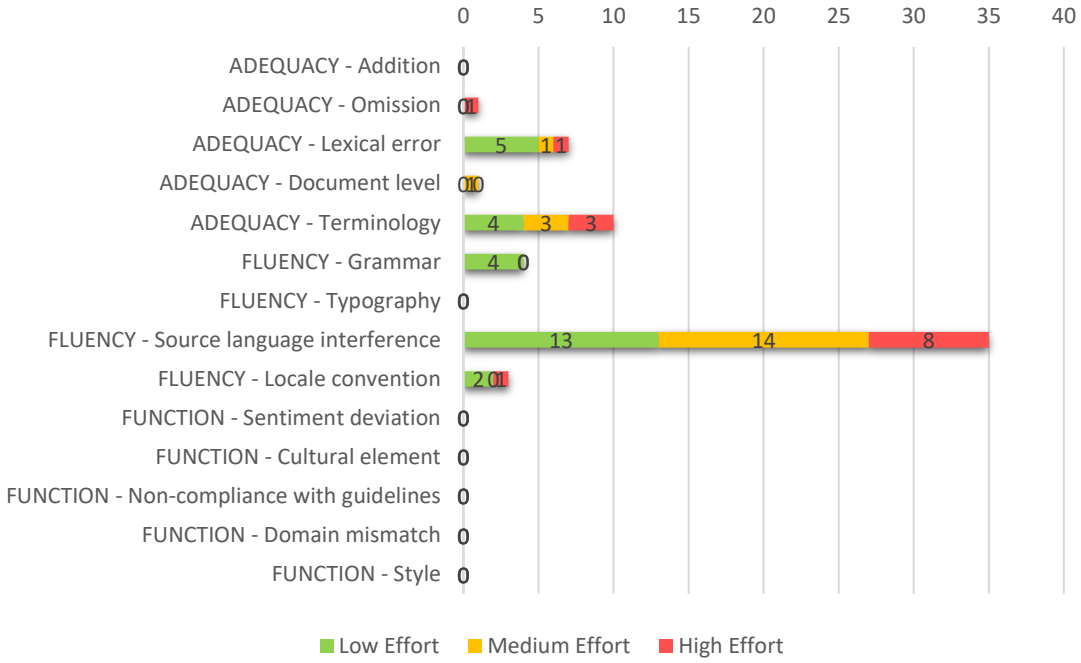


Figure 60: Effort distribution by error category for A3.

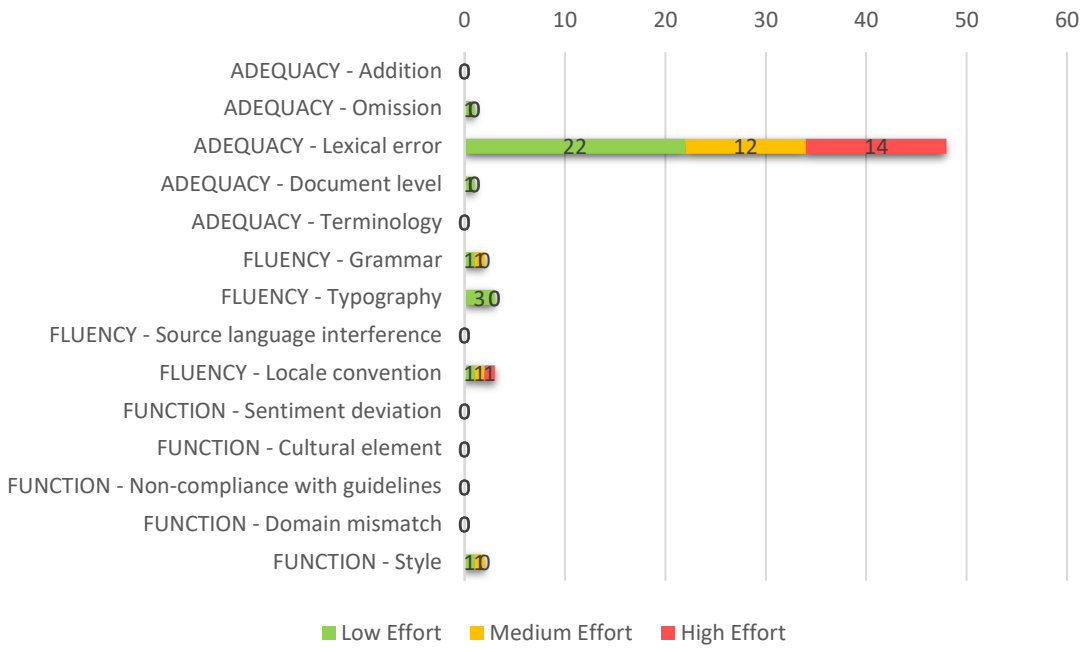


Figure 61: Effort distribution by error category for A4.

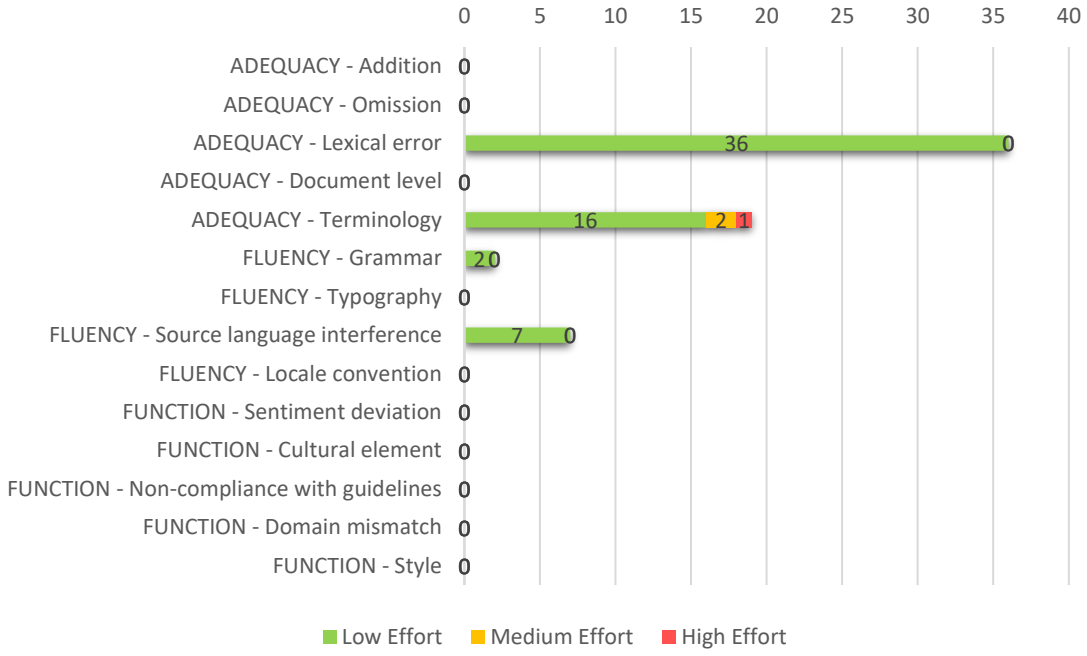


Figure 62: Effort distribution by error category for A5.

Overall, the graphs illustrate that each annotator had a distinct PE experience, largely influenced by the different errors they identified. For example, A3 primarily annotated SL interference, and A4 mostly found lexical errors; therefore, they naturally encountered more high-effort errors in these categories. In

the case of A1, the most difficult errors to correct seemed to be mainly terminology, lexical errors, domain mismatch and style. For A2, the difficult errors were mostly concentrated in the document-level error type. Although A5 was among those who made the highest amount of changes to the text, only one error (related to terminology) was considered difficult to correct. These findings emphasise once again the inherent subjectivity in both error categorisation and effort perception.

Furthermore, the distribution of effort levels across annotators does not completely align with the initial hypothesis. The divergence in error categorisation hinders drawing solid conclusions in that regard. Errors considered difficult to correct in previous studies (4.4.2) were observed as requiring a high effort only for some of the annotators. For example, document-level errors were challenging for A2, but only this annotator found such errors. Similarly, SL interference was perceived as difficult on several occasions by A3, but this annotator was the one who found the highest amount of errors belonging to this type. Conversely, other errors which were assumed to be challenging were viewed as easy to correct, as is the case with grammatical errors (systematically easy for A3 and A5, the only annotators who identified such errors).

In addition, the relationship between different error types and perceived effort was analysed using Spearman's correlation coefficient for each annotator. The results (Table 49) indicated generally weak or no significant correlations.

Annotator	**Spearman's ρ (error type ↔ effort)**	***p*-value**
A1	0.20	0.19
A2	-0.02	0.91
A3	0.11	0.38
A4	0.12	0.37
A5	0.06	0.63

Table 49: Relationship between error type and effort (Spearman test).

This finding suggests that specific error types could not be identified as systematically more difficult to correct than others. This outcome seems to contradict the findings of the user survey, according to which specific MT errors were considered more challenging to handle by respondents (in particular, cultural adaptation, mistranslation, and SL interference, as discussed in 9.1). Further research, with longer texts and more participants, would thus appear beneficial to investigate this phenomenon. The survey respondents did not have access to real examples and worked with more diverse languages, which may

have also skewed the results. In addition, it would also appear relevant to study the difficulty attributed to a specific correction before and during PE, as pre-task perception may differ from the experience when working on a PE project.

Crossing the effort with the editing actions performed (Table 50) does not allow for drawing conclusive findings given the prevalence of replacing for all annotators.

	A1	**A2**	**A3**	**A4**	**A5**
LOW	13	26	28	30	61
Deleting	1	2	0	1	1
Inserting	2	3	1	5	1
Moving	0	0	1	1	1
Replacing	10	21	26	23	58
MEDIUM	20	8	19	15	2
Deleting	1	0	0	0	0
Inserting	0	0	0	0	0
Moving	0	0	0	0	0
Replacing	19	8	19	15	2
HIGH	14	4	14	15	1
Deleting	0	0	0	0	0
Inserting	0	0	1	0	0
Moving	0	0	0	0	0
Replacing	14	4	13	15	1

Table 50: Effort distribution by editing action for each annotator.

The relationship between editing actions performed and the associated effort was examined using Spearman's correlation. The values obtained (Table 51) showed no significant correlations for most annotators, with the highest weak positive correlation observed for A3 (0.13).

Annotator	Spearman's ρ (editing action ↔ effort)	*p*-value
A1	-0.03	0.83
A2	0.09	0.61
A3	0.13	0.30
A4	0.00	1.00
A5	0.02	0.90

Table 51: Relationship between editing action and effort (Spearman test).

This result entails that specific error actions could not be identified as systematically more difficult to perform than others. As for the correlation between error types and perceived effort, this finding is not aligned with the outcomes of the user survey, where specific editing actions were found to involve a higher difficulty (as discussed in 9.1). However, in this case, the survey respondents linked the difficulty of editing actions to the context, which may explain the lack of a clear correlation here.

F. Qualitative Analysis

Since the quantitative analysis revealed significant divergence in the PE experience across annotators, a qualitative analysis based on selected examples is presented here to better illustrate these differences.

- ***Example 1***

Source	This medicine is used to treat inflamed sweat rash and Athlete's Foot skin infections.
Raw MT output	Este medicamento se utiliza para tratar la sudoración inflamada y las infecciones de la piel del pie de atleta.
Back translation	This medicine is used to treat inflamed sweating and athlete's foot skin infections.

Table 52: Source, raw MT output and back translation for Example 1.

A	Post-edited version	Back translation	Effort	Error type	Severity	Editing action
A1	Este medicamento se utiliza para tratar el sarpullido cutáneo por sudoración y la	This medicine is used to treat skin rash caused by sweating and athlete's foot skin infection.	Medium	Adequacy (Terminology)	Medium	Replacing

	infección cutánea del pie de atleta.					
A2	Este medicamento se utiliza para tratar la irritación por sudor y las infecciones cutáneas causadas por hongos, como el pie de atleta.	This medicine is used to treat sweat irritation and skin infections caused by fungi, such as athlete's foot.	High	Adequacy (Document level)	High	Replacing
A3	Este medicamento se utiliza para tratar la erupción por sudor con inflamación e infecciones de la piel provocadas por el pie de atleta.	This medicine is used to treat inflamed sweat-induced eruption and skin infections caused by athlete's foot.	High	Adequacy (Terminology)	High	Replacing
A4	Este medicamento se utiliza para tratar erupciones por sudoración con inflamación e infecciones cutáneas como el pie de atleta.	This medicine is used to treat inflamed sweat-induced eruptions and skin infections such as athlete's foot.	High	Adequacy (Lexical error)	Medium	Replacing
A5	Este medicamento se utiliza para tratar la dermatitis por sudor inflamada y las infecciones del pie de atleta.	This medicine is used to treat inflamed sweat dermatitis and athlete's foot infections.	Medium	Adequacy (Terminology)	Low	Replacing

Table 53: Post-edited versions and annotations for Example 1.

Example 1 illustrates a case of a highly specialised source sentence, containing two terminological units ("inflamed sweat rash" and "Athlete's Foot skin infections"), which were post-edited differently by each annotator. "Inflamed sweat rash" was rendered as "sudoración inflamada" ("inflamed sweating") in the MT output, which was judged incorrect by all annotators. However, each provided a distinct correction. For example, "rash" was translated as "sarpullido" (A1), "irritación" (A2), "erupción" (A3, A4) and "dermatitis" (A5). A similar variation occurred for "Athlete's Foot skin infections". While A1 stayed close to the source structure, A2 opted for an addition with explicitation ("skin infections caused by

fungi"); A3 and A4 also implemented a similar strategy as A2 (adding "caused by" and "such as"); on the other hand, A5 chose to omit "skin" (mentioning only "foot infections").

It should also be noted that all annotators judged the errors to belong to the adequacy category, yet their specific classifications varied: three selected terminology (A1, A3, A5), one chose document level (A2), and another identified it as a lexical error (A4). Overall, the effort was perceived as relatively high, with three annotators assigning a high effort score and two rating it as medium. More pronounced differences were observed in the perception of error severity, with two annotators judging it as high, two as medium, and one as low.

- ***Example 2***

Source	How do you catch Athlete's Foot?
Raw MT output	¿Cómo conseguir el pie de atleta?
Back translation	How to get athlete's foot?

Table 54: Source, raw MT output and back translation for Example 2.

A	Post-edited version	Back translation	Effort	Error type	Severity	Editing action
A1	¿Cómo se contrae el pie de atleta?	How does one contract Athlete's Foot?	High	Adequacy (Lexical error)	High	Replacing
A2	¿Cómo se contagia el pie de atleta?	How is Athlete's Foot transmitted?	Low	Function (Domain mismatch)	High	Replacing
A3	¿Cómo se contrae el pie de atleta?	How does one contract Athlete's Foot?	Medium	Fluency (Source language interference)	Medium	Replacing
A4	¿Cómo se contrae el pie de atleta?	How does one contract Athlete's Foot?	Medium	Adequacy (Lexical error)	Medium	Replacing

A5	¿Cómo se contagia el pie de atleta?	How is Athlete's Foot transmitted?	Low	Adequacy (Lexical error)	Low	Replacing

Table 55: Post-edited versions and annotations for Example 2.

In the case of ***Example 2***, the raw MT output presents an error in meaning, as "catch" was translated as "conseguir", which carries a positive connotation and is therefore not appropriate for diseases. All annotators changed this verb, opting for either "contraer" (A1, A3, A4) or "contagiar" (A2, A5). While both terms are equivalent in meaning, they differ in nuance: "contraer" refers to acquiring or developing a disease, whereas "contagiar" involves transmitting a disease from one person to another.

Although the solution is comparable for all annotators, their perception of the error is rather different: it was a lexical error for three of them (A1, A4, A5), but a domain mismatch error for A2 and an error related to source language interference for A3. All categorisations are nevertheless equally valid: the MT output presented a lexical error since one lexical item was not translated correctly; one could argue that this issue could be attributed to the interference of the source ("catch" can indeed have a more positive meaning); and "contraer" or "contagiar" are specific to the medical domain, which explains the selection of a domain mismatch error. The perception of the effort and severity also shows significant discrepancies across annotators.

9.4.1.6 Summary & Outcomes

This experiment explored the relationship between the type of MT error and its associated effort, severity and editing action. While the initial objective revolved around establishing a new evaluation model for MT based on the PE effort, the findings showed that the PE experience was significantly different for each of the five linguists who participated in this experiment. A more detailed summary of the results and the answers they brought to the research questions is presented below.

- RQ1: How similar or different is the PE experience across various linguists?

The linguists who took part in this experiment disagreed on the editing decision for more than half of the segments (specifically, 59%). The TER scores also varied greatly (the lowest being 9.4, and the highest being 20.2) and did not align with the proportion of edited segments. Hence, the annotator who changed the highest number of segments is not the one who modified the highest proportion of the text.

The inter-annotator agreement scores also showed that the perception of error types, severity and effort differed greatly from one annotator to another. Only the editing actions seemed to be consistent, with all annotators performing primarily replacements. These findings show that the PE experience differed

significantly among annotators, suggesting that the perception of quality, error types and effort is highly subjective.

- RQ2: Did annotators identify the same errors?

The analysis of error type classification revealed significant divergence in the errors identified by each annotator, since each post-editor found a significantly different number of occurrences for each error type. This finding highlights the high degree of subjectivity involved in error categorisation.

- RQ3: Did annotators judge severity in a similar way?

The analysis of severity annotations demonstrated that although most errors were classified as having a low severity by the majority of annotators, significant discrepancies occurred regarding the overall error severities found by each annotator. Therefore, the variability observed in error severity annotations also suggests that this task is equally influenced by subjectivity.

- RQ4: Did annotators experience similar levels of effort?

The overall effort varied significantly from one annotator to the other, highlighting the subjective nature of effort perception. Moreover, the overall effort was not necessarily consistent with the TER scores, indicating that the difficulty did not seem to increase with the amount of editing performed. Additionally, no correlation could be established between the effort and the other dimensions (namely, error categories, error types and editing actions). This suggests that the effort perception is not influenced by any of these variables but is rather a subjective phenomenon that varies from one individual to another in the PE process.

Consequently, the initial hypothesis, according to which specific MT errors would be perceived as requiring a greater PE effort, cannot be confirmed. This finding impedes the design of an MT evaluation model based on the PE effort, as originally planned. Nevertheless, it should be acknowledged that this experiment was conducted with a limited number of annotators; therefore, replicating it with a larger sample may allow for identifying more recurring patterns. It could also be argued that the typology could be simplified to reduce the disagreement between annotators, though finding an optimal balance between granularity of the typology and informativeness of the annotations is likely to be an arduous task. Additional effort measures could also be considered in follow-up studies, including the temporal effort (time to complete the PE task) and the cognitive effort (e.g., eye-tracking data). This would allow for further exploring the relationship between each dimension.

Finally, since the results are highly affected by subjectivity, further exploring the subjectivity component would also be a relevant research angle. More specifically, if subjectivity can be quantified and if it appears to be constant from one PE task to the other, it could be incorporated as a dedicated component (e.g. a penalty) in the BF.

9.4.1.7 Additional Exploration: Subjectivity in Post-Editing

The above findings called for further investigation into the role of subjectivity in PE. Therefore, a follow-up analysis was performed to explore this phenomenon. The primary research question addressed here was: To what extent is the subjectivity of the editing decision stable across various PE tasks?
In the initial experiment presented above, an agreement rate of 41% was observed among five post-editors regarding the editing decision (i.e., whether a segment required modification). This raised questions about the generalisability and representativeness of this finding.

To examine this further, an additional analysis was conducted using data from a PE task carried out by institutional translators employed by the Valencian Government. These professionals routinely engage in translation workflows that incorporate MT followed by PE, which constitutes a central aspect of their day-to-day professional practise. The PE task was designed to reflect realistic working conditions and was intended to capture performance data that is representative of their habitual interaction with machine-generated texts. As such, the findings provide insights into PE behaviour and outcomes within a professional institutional context, where MT technologies are systematically integrated into the translation process. The ST used for this task was a leaflet for Pantoprazol, a medicine used for digestive issues such as heartburn and acid reflux.

Domain	Medical
Number of segments	186
Number of words	2,872
Number of characters	15,925
TTR	0.28
MTLD	62.91

Table 56: Source text characteristics (Pantoprazol).

The characteristics of this text are summarised in Table 56. Apart from belonging to the same domain (medical), the TTR value is also close to the text used in the previous experiment (0.28 here vs 0.3 in the first text), suggesting similar lexical diversity between the two texts. The MTLD scores show a slightly different pattern. The second text has a higher MTLD (62.91 here vs. 58.94), indicating greater lexical diversity while accounting for text length. Overall, these values indicate that both texts are of comparable lexical complexity. The experiment was conducted taking a Spanish source text from the

medical domain, translated into Valencian using Salt[51], a free tool developed by the Generalitat Valenciana that provides bidirectional translation between Valencian and Spanish.

A total of 13 participants provided post-edited versions, together with annotations of the effort, error type and severity. Although each annotator performed the same PE task, the portions of the text processed varied across individuals. Table 57 below provides a summary of each annotator's progress.

Annotator	ID of last segment edited
A1	41
A2	44
A3	31
A4	34
A5	22
A6	38
A7	30
A8	39
A9	29
A10	28
A11	44
A12	19
A13	39

Table 57: Progress for each participant.

Consequently, various analyses were conducted on different samples, excluding some annotators based on their task completion level:

- ***Analysis 1***: original analysis with Daktacort text
- ***Analysis 2***: sample of original analysis of Daktacort text, taking the same number of segments as for Analysis 5
- ***Analysis 3***: sample of Pantoprazol text, excluding annotators who processed less than 28 segments (A5, A12)

[51] https://salt.gva.es/es/traductor (consulted: 30/04/2025)

- ***Analysis 4***: sample of Pantoprazol text, excluding annotators who processed less than 38 segments (A3, A4, A5, A7, A9, A10, A12)
- ***Analysis 5***: sample of Pantoprazol text, excluding annotators who processed less than 39 segments (A3, A4, A5, A6, A7, A9, A10, A12)

An overview of the agreement levels obtained across these analyses is provided in Table 58.

Analysis ID	Source text	Number of annotators	Segments	Words	Segments edited by all	Segments left unedited by all	Total segments with same editing decision from all
1	Daktacort	5	126	1,368	14% (18/126)	27% (35/126)	41% (53/126)
2	Daktacort	5	39	413	13% (5/39)	18% (7/39)	31% (12/39)
3	Pantoprazol	11	28	328	25% (7/28)	25% (7/28)	50% (14/28)
4	Pantoprazol	6	38	549	34% (13/38)	18% (7/38)	52% (20/38)
5	Pantoprazol	5	39	557	51% (20/39)	26% (10/39)	77% (30/39)

Table 58: Overview of editing trends for both PE tasks, based on different samples.

The results revealed significant variability in the agreement rates, ranging from 31% to 77% across the analyses. However, the agreement on editing decisions appeared to be consistently lower for the Daktacort text than for Pantoprazol. This discrepancy was expected for *Analysis 1*, given that Daktacort is a longer text (agreement may decrease as the text length increases due to greater potential for variability in perception). However, this hypothesis was invalidated, as shown by the unexpectedly low agreement rate (31%) obtained in *Analysis 2*, despite the reduced segment count.

A particularly marked difference emerged between *Analyses 2* and *5*, which present the same number of segments and annotators, yet a stark divergence in agreement (31% for Daktacort vs 77% for Pantoprazol).

Furthermore, agreement patterns were found to differ between the texts. For Daktacort, annotators tended to reach consensus mainly on segments that did not require editing. In contrast, for Pantoprazol, annotators appeared to agree more frequently on segments that did require edits (except in *Analysis 3*, where the agreement rates were similar for both edited and unedited segments). This suggests that text characteristics may influence editing consistency across annotators.

The remainder of this analysis delves into a more detailed comparison between *Analyses 2* and *5*, as these two scenarios present the closest similarities in terms of segment count and annotators involved. An overview of the editing behaviour and TER scores observed in these analyses is presented in Table 59.

A. Analysis of Edits

	Daktacort **Analysis 2**						**Pantoprazol** **Analysis 5**					
	A1	A2	A3	A4	A5	x̄	A1	A2	A8	A11	A13	x̄
Total edited segments	21 (54%)	11 (28%)	16 (41%)	21 (54%)	20 (51%)	17.8 (46%)	25 (64%)	28 (72%)	25 (64%)	24 (62%)	26 (67%)	25.6 (66%)
TER	16.4	12	18.1	16.1	10.8	14.7	9	8.3	12.4	7.9	9.1	9.34

Table 59: Overview of edited segments and TER.

The analysis revealed a notable variation in the number of edited segments for the Daktacort text, ranging from 11 to 21 segments. In contrast, a more stable editing behaviour was observed in the case of the Pantoprazol text, with the number of edited segments falling between 24 and 28. This stability is further reflected in the TER scores, which also exhibit less fluctuation for the Pantoprazol text (7.9 to 12.4) compared to Daktacort (12 to 18.1).

Interestingly, when examining the averages, it appeared that annotators working with the Daktacort text edited fewer segments on average, with 46% of segments being modified, compared to 66% for the Pantoprazol text. However, despite the lower number of edited segments, annotators made more changes on average in the Daktacort analysis, as indicated by a higher average TER of 14.7, compared to 7.34 for the Pantoprazol text. This suggests that while fewer segments were edited in the Daktacort text, the edits made were more extensive.

B. Inter-Annotator Agreement

Furthermore, as done with the Daktacort analysis, the pairwise agreement was also calculated for the Pantoprazol text. Here, the agreement matrices were computed for Analyses 3, 4 and 5 (Figures 63-65).

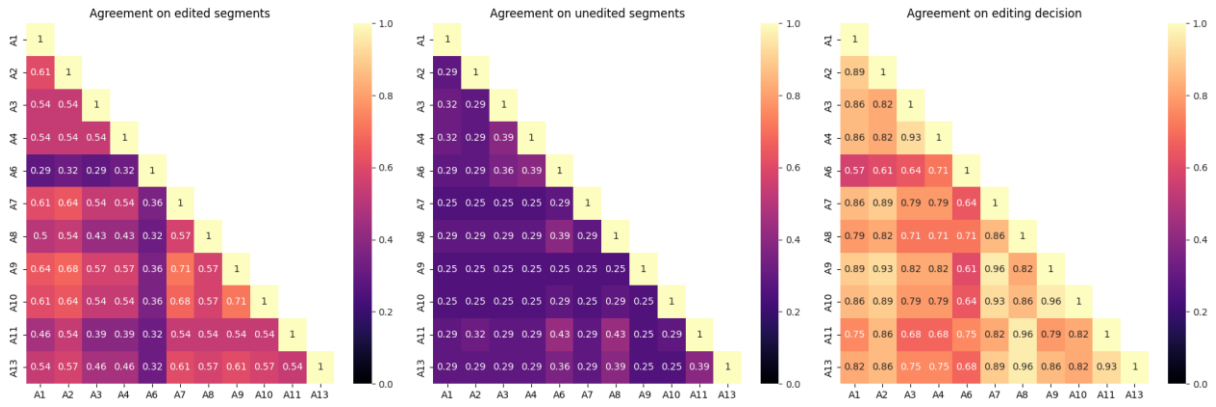


Figure 63: Pairwise agreement for Analysis 3.

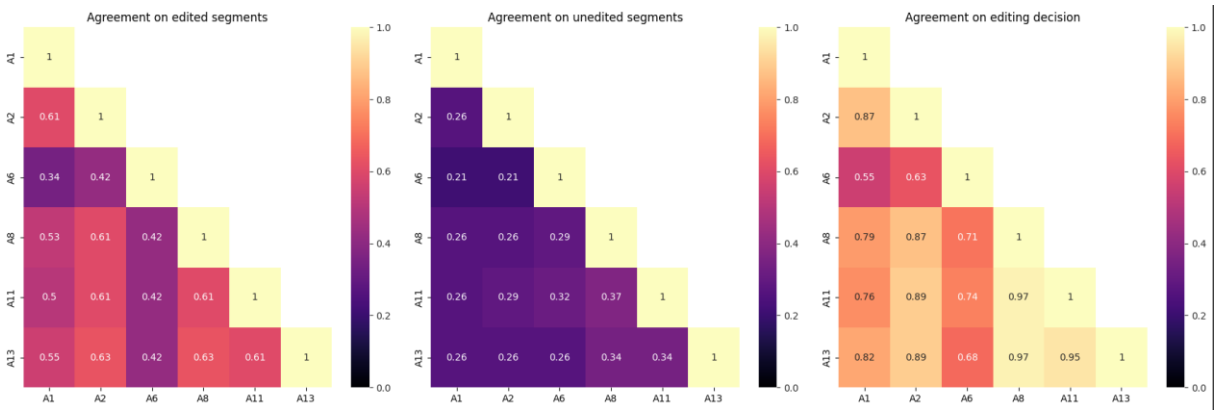


Figure 64: Pairwise agreement for Analysis 4.

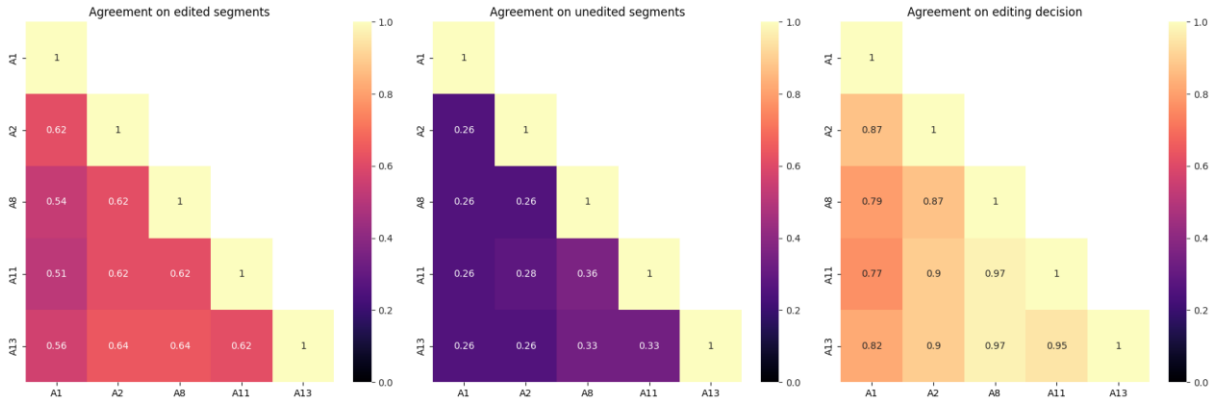


Figure 65: Pairwise agreement for Analysis 5.

A clear pattern can be identified in Analysis 3, where A6 seemed to systematically have the lowest agreement with all other annotators (Figure 63), in particular for the agreement on edited segments and editing decision, which may have impacted the overall agreement rate. The same pattern can be observed for Analysis 4 (Figure 64), where the agreement between A6 and the other annotators also remained low. However, such a pattern cannot be identified in the case of Analysis 5 (Figure 65), where the agreement rates are more homogeneous, due to having a smaller sample and excluding A6.

C. Error Analysis

In addition, it also appears relevant to study the annotated errors for both PE tasks (data presented in Table 60). The total number of errors identified by the annotators demonstrated greater stability in the Pantoprazol text, with counts ranging consistently between 25 and 27 errors. In contrast, the Daktacort text exhibited a wider variability, with the number of errors found fluctuating between 11 and 21. Furthermore, when comparing the average number of errors across both tasks, fewer errors were detected in the EN>ES task involving the Daktacort text (with an average of 18 errors). Conversely, the Pantoprazol text yielded an average of 26 errors.

The analysis of the distribution patterns across the different error types revealed both similarities and discrepancies between the Daktacort and Pantoprazol texts. Notably, a similar trend was observed for lexical errors, where annotators either identified a high number of errors or none at all. For instance, in the Daktacort text, A4 identified 18 lexical errors and A5 found four, while the other annotators reported only 0 to 1 lexical error. In the case of Pantoprazol, A11 found 15 lexical errors, but all other annotators did not find any. This pattern suggests a polarised perception of lexical errors among annotators, highlighting inconsistencies in their approach to error identification.

	Daktacort Analysis 2						Pantoprazol Analysis 5					
	A1	A2	A3	A4	A5	x̄	A1	A2	A8	A11	A13	x̄
Addition	0	0	0	0	0	0	1	2	1	0	2	1.2
Omission	0	0	0	1	0	0.2	0	0	0	0	0	0
Lexical error	0	0	1	18	14	6.6	0	0	0	15	0	3
Document level	0	7	1	0	0	1.6	0	0	0	0	0	0
Terminology	5	0	5	0	6	3.2	8	11	0	0	12	6.2
Mistranslation*	X	X	X	X	X	X	1	1	1	0	0	0.6
ADEQUACY (total)	*5*	*7*	*7*	*19*	*20*	11.6	*9*	*13*	*1*	*15*	*14*	10.4
Grammar	0	3	1	1	0	1	8	5	23	6	12	10.8
Typography	2	0	0	0	0	0.4	1	0	0	0	0	0.2
Source language interference	0	0	5	0	0	1	2	0	0	0	0	0.4
Locale convention	0	0	3	1	0	0.8	6	8	0	0	0	2.8
FLUENCY (total)	*2*	*3*	*9*	*2*	*0*	3.2	*17*	*13*	*23*	*6*	*12*	14.2
Sentiment deviation	0	0	0	0	0	0	0	0	0	0	0	0
Cultural element	0	0	0	0	0	0	0	0	0	0	0	0
Non-compliance with guidelines	0	0	0	0	0	0	0	0	0	0	0	0
Domain mismatch	5	0	0	0	0	1	0	0	0	0	0	0
Style	10	1	0	0	0	2.2	0	0	0	4	0	0.8
FUNCTION (total)	*15*	*1*	*0*	*0*	*0*	3.2	*0*	*0*	*0*	*4*	*0*	0.8
Total errors found	***22***	***11***	***16***	***21***	***20***	***18***	***27***	***27***	***25***	***25***	***26***	***26***

* Note that “mistranslation” was not included in the annotation of Daktacort for simplification purposes.

Table 60: Overview of errors found by each annotator.

Additionally, the annotators’ perception of lexical and terminological errors showed an unexpected pattern. In the Daktacort text, A4 identified 18 lexical errors but did not find any terminological error. Similarly, A11 found 15 lexical errors in Pantoprazol and did not identify any terminological error. On the other hand, annotators who identified terminological errors, such as A1 for Daktacort, and A1, A2, and A13 for Pantoprazol, did not flag any lexical error. This observation highlights the ambiguous

distinction between these two error types, suggesting that annotators may have distinct interpretations of what distinguishes a lexical error from a terminological one. Furthermore, the function category exhibited the lowest number of errors across both texts, with no annotation at all for certain types such as sentiment deviation, cultural element, and non-compliance with guidelines. Additionally, only a few annotators identified errors related to style and domain mismatch, which suggests that either the MT outputs did not contain such errors, or that the annotators tended to agree on the definition of these error types. In addition, a significant difference occurred in the number of grammatical errors identified between the two tasks, with an average of one grammatical error for the Daktacort text compared to 10.8 for Pantoprazol. This difference may be attributed to several factors, including the specific language combination. More specifically, the amount of data each MT system was trained on and the influence of the SL interference between Spanish and Valencian are likely to have affected the results, as further studied with examples hereafter (E. Qualitative Analysis).

D. Effort Analysis

An overview of the perceived effort is presented in Table 61, which summarises the varying levels of effort reported by annotators across the different texts, together with the resulting effort scores.

	Daktacort						**Pantoprazol**					
	Analysis 2						**Analysis 5**					
	A1	A2	A3	A4	A5	x̄	A1	A2	A8	A11	A13	x̄
Low	5	5	5	14	18	9.4	18	16	18	22	20	18.8
Medium	11	3	2	2	2	4	8	8	8	4	5	6.6
High	6	3	9	5	0	4.6	0	0	0	0	0	0
Effort score*	***57***	***26***	***54***	***43***	***22***	***40.4***	***34***	***32***	***34***	***30***	***30***	***32***

* Low effort (1), medium effort (2) and high effort (5)

Table 61: Overview of effort distribution for each annotator.

For both tasks, the majority of errors were categorised as involving a low effort. However, in the case of Daktacort, annotators identified some high-effort errors, which was not the case for Pantoprazol. In addition, despite the average effort score being higher for Daktacort, the annotators working on this text reported finding fewer errors overall compared to the annotators working with Pantoprazol. This suggests that a high number of errors does not necessarily imply a higher editing effort.

E. Qualitative Analysis

As for the first analysis, a qualitative analysis was performed, and selected examples are included hereafter for illustration purposes.

- ***Example 3***

Source	Lea todo el prospecto detenidamente antes de empezar a tomar este medicamento, porque contiene información importante para usted.
Raw MT output	Llija tot el prospecte detingudament abans de començar a prendre aquest medicament, perquè conté informació important per a vosté.
English translation	Read all of this leaflet carefully before you start taking this medicine because it contains important information for you.

Table 62: Source and raw MT output for Example 3.

A	**Post-edited version**	**Effort**	**Error type**
A1, A2, A3, A4, A7, A9, A10	Llija tot el prospecte detingudament abans de començar a prendre este medicament, perquè conté informació important per a vosté.	Low (A1, A2, A4, A7, A9, A10) Medium (A3)	Locale convention (A1, A2, A3, A7, A9, A10) Style (A4)
A6, A8, A11, A13	Llija tot el prospecte detingudament abans de començar a prendre aquest medicament, perquè conté informació important per a vosté.	No changes	

Table 63: Post-edited versions and annotations for Example 3.

Example 3 revealed a notable divergence in the treatment of the demonstrative adjective “aquest” (MT output) versus “este” (the correction chosen by most linguists). While 7/11 annotators modified “aquest” to “este”, 4 retained the original MT output. The effort required for this correction was generally rated as low, indicating that the change was straightforward for most linguists. However, one annotator classified it as medium effort. In addition, most categorised the error as a *locale convention* issue, while one labelled it as *style*, reflecting differences in interpretation. Therefore, while there was moderate agreement among annotators, discrepancies were still observed.

- ***Example 4***

Source	Si tiene alguna duda, consulte a su médico o farmacéutico.
Raw MT output	Si té algun dubte, consulte al seu metge o farmacèutic.
English translation	If in doubt, consult your doctor or pharmacist.

Table 64: Source and raw MT output for Example 4.

A	Post-edited version	Effort	Error type
A1, A2, A3, A4, A6, A7, A8, A9, A11, A13	Si té algun dubte, consulte el seu metge o farmacèutic.	Low (A1, A4, A6, A7, A8, A11, A13) Medium (A2, A3)	Grammar (A1, A2, A3, A4, A6, A7, A8, A11, A13)
A9	Si té algun dubte, consulte el metge o farmacèutic.	Medium	Lexical error
A10	Si té algun dubte, consulte al metge o farmacèutic.	Low	Grammar

Table 65: Post-edited versions and annotations for Example 4.

The MT output in ***Example 4***, "consulte al seu metge", shows a clear case of SL interference, as the preposition "a" (mirroring Spanish "consulte a su medico") is incorrect in Valencian, where the verb "consultar" is used transitively ("consulte el seu metge"). This grammatical discrepancy was identified and corrected by the majority of annotators: 10/11 selected the *grammar* error category and post-edited the MT output accordingly. The effort required for this correction was predominantly rated as low, suggesting that this error type is easily recognisable. However, some variability emerged in the annotations. While most classified the error as grammatical, one annotator (A9) labelled it as a lexical issue. Additionally, A10 retained the preposition ("al metge"). This divergence, though minor, highlights the challenges in achieving full consistency even with seemingly straightforward grammatical rules.

When comparing both examples, a pattern emerges: while medical leaflets employ standardised phrases that should theoretically limit variation, PE outcomes still reveal subtle discrepancies in error classification and corrections. The moderate but incomplete agreement among annotators suggests that factors such as individual linguistic background or interpretation of guidelines can influence PE decisions and confirms the findings of the first part of this experiment regarding subjectivity.

F. Conclusions

This additional exploration suggests that subjectivity in PE cannot be quantified as a constant dimension. While the complexity of the source text is unlikely to influence the results, as the texts are comparable in nature, several other factors may have impacted the findings. First, since two different MT models and two different language pairs were involved, the participants in each task were likely working with raw MT outputs of different quality levels. This phenomenon may have been exacerbated here, given the quantity of training data available. Additionally, individual perception and experience may have significantly shaped the annotators' assessments, leading to varied interpretations of errors and the effort required to correct them. Lastly, the professional profile of the participants involved may also have influenced the results. The freelance translators post-editing the Daktacort text had broader translation experience, whereas the institutional translators working with the Pantoprazol text specialised solely in PE and did not have as much experience in translating from scratch. This observation highlights the crucial role of the individual background when it comes to translation quality assessment. In addition, these findings suggest that a "one size fits all" approach is inadequate for evaluating translation quality, which is why organisations such as TAUS have emphasised the need for a "dynamic" framework.

As a pilot study, this research lays the groundwork for further exploration into the nuances of translation quality assessment in a PE context. Several avenues for future research could thus be considered. One possibility is to conduct another PE experiment involving multiple participants and apply the same analytical framework used here. This would allow further examination of the agreement on the editing decision. While the agreement rate on the editing decision was 41% in the first experiment, the results obtained in this follow-up study are less encouraging, with two similar samples (in terms of length and number of annotators) yielding considerably different results (31% for the Daktacort sample vs 77% for the Pantoprazol sample). The MT quality and the different language pairs involved, as well as the background of the participants likely affected these results. Consequently, identifying which factors influence the agreement rate on the editing decision would be highly relevant. It should, however, be noted that a significant challenge associated with this idea is the potential cost involved in recruiting and compensating additional participants. Another approach could rely on leveraging existing PE datasets, but this would also present difficulties, particularly in identifying datasets that include both unedited segments and multiple post-edited versions of the same source text. For instance, the multilingual PE dataset (Fomicheva et al., 2022) appears to feature post-edited versions of all sentences. Similarly, the EU APE-QUEST dataset (Ive et al., 2020) contains only one post-edited version, and the original text from a parallel corpus, which restricts the scope for comparative analysis.

9.4.2 Quality Element 2: Quality of Post-Edited Texts

The previous section explored MT quality and how it may affect the PE effort. The focus was thus placed on the quality of raw MT output, before PE. While MT quality represents a critical element in the BF, it is not the unique focus in terms of quality assessment. MT quality is relevant to further explain productivity and UX; however, when adopting the point of view of a customer ordering PE services, quality assessment happens once the text has been post-edited. Final texts (after PE) are typically expected to meet the same quality requirements as human translations, especially when full PE is performed. However, the required quality level also depends on the project's specifications. The ISO 18587:2017 standard (ISO, 2017) on PE requirements indeed provides the following definition of full PE: "process of post-editing to obtain a product comparable to a product obtained by human translation".

Consequently, considering that quality expectations for PE are the same as for HT, the quality pillar of the BF, which revolves around the final quality of post-edited texts, does not propose a novel evaluation framework but rather adopts established metrics and standards for translation quality assessment.

It is acknowledged here that differences in quality can arise between texts coming from HT and PE, yet the study of these differences – although relevant to update the evaluation models to current practises of the industry – falls outside the scope of the present research.

10. Application of the Benchmarking Framework

With the BF fully designed and structured, this chapter shifts focus to its practical application in real-world scenarios. To that end, the framework is applied to two distinct contexts: the evaluation of Smart Editor for PE in an industrial setting, and the assessment of prompt-based alternative translations using the AI Professional Companion plug-in in Trados Studio. These applications demonstrate the framework's versatility and effectiveness in addressing both established tools and emerging technologies. In addition to validating the utility of the BF, this chapter also provides actionable insights for optimising PE environments.

10.1 Benchmarking Smart Editor – A Case Study at LanguageWire [Confidential]

The discussion begins with the first application, which takes place in an industrial context, as presented in 7. More specifically, the first application of the BF consists of an assessment of Smart Editor for PE.

This section contains proprietary information, technical details, and methodologies developed in collaboration with LanguageWire. In accordance with the confidentiality agreement and to protect the commercial interests and intellectual property of LanguageWire, all the content belonging to 10.1 is withheld from the publicly available version of this thesis. The complete work was made available only for review by authorised academic staff and examiners.

10.2 Benchmarking Augmented PE with Prompt-Based Alternative Translations

The second application of the BF explores the potential of augmented PE through the use of prompt-based alternative translations, using the AI Professional Companion plug-in in Trados Studio. This scenario represents an innovative development in translation technology, where AI-driven suggestions aim to enhance the efficiency and creativity of post-editors. Beyond testing the framework's adaptability to emerging technologies, this section also provides insights into the future of AI-assisted translation and its implications for post-editors.

10.2.1 Rationale

Since their conceptualisation in the 1980s, CAT tools have not significantly evolved to accommodate the specific needs of PE. However, with the advent of LLMs, new integrations of AI-driven functionalities have emerged. As a result, AI assistants can now be integrated into CAT tools, offering features such as alternative translations or in-context definitions. Among the major CAT tool providers, Trados Studio has introduced AI assistance through the AI Professional Companion plug-in, leveraging OpenAI's capabilities to provide alternative translations via prompting.

In this context, assessing the impact of AI-driven features on PE is paramount, and such evaluation is precisely the aim of the present work. This study aims to evaluate the impact of AI-based alternative translations on the PE workflow by comparing AI-enhanced PE with a traditional PE setting. The choice of this specific PE assistance technology over other CAT features that would be relevant for PE is motivated by its high ranking in the results of the exploratory user survey (9.1), as *rewriting assistance* reached the third position (after the *document view* and the *knowledge feature*).

The evaluation is structured around the established benchmarking dimensions: productivity, UX, and translation quality. Furthermore, this study provides insights into how post-editors interact with alternative translations and explores ways to further tailor AI assistance to PE needs.

10.2.2 Objectives & Research Questions

This study aims to explore the impact of using AI assistance to generate alternative translations in a PE workflow, using the AI Professional Companion plug-in for Trados Studio. The objective is therefore twofold: 1) to evaluate the influences of this technology on key aspects of the PE process and 2) to understand post-editors' perceptions of its usefulness.

The study thus addresses the following research questions:

- **Impact on PE**: how does working with alternative translations generated via prompts affect PE in terms of
    - Productivity?
    - UX?
    - Quality?
- **Post-editors' perception**: how do post-editors perceive this new PE technology?
    - Do they find it useful?
    - What are its strengths and weaknesses?
    - How could it be further tailored to PE needs?

10.2.3 Experimental Setup

Three source texts (T1, T2, T3) were selected for this study. They all consist of product descriptions for high-end perfumes from a UK-based company, which suggests that they were originally written in English. Furthermore, an online search did not reveal any existing translations of these descriptions, confirming that no translation was available as training material for MT. The key characteristics of the selected texts, in terms of length and lexical complexity, are summarised in Table 71. Notably, all three texts are comparable in length, each consisting of approximately 500 words.

	T1	**T2**	**T3**
Domain	Cosmetic	Cosmetic	Cosmetic
Number of segments	39	35	34
Number of words	517	504	507
TTR	0.46	0.5	0.5
MTLD	55.67	73.32	86.27

Table 71: Source texts characteristics.

Translations were produced for four target languages – namely, French, Spanish, German, and Dutch. It is worth noting that both the language combinations and the domain of the selected texts reflect typical translation requests handled by LanguageWire, and the MT output for this study was generated using LanguageWire's proprietary MT system.

A total of eight in-house linguists working at LanguageWire participated in the study, with two linguists assigned per target language. Before starting the task, participants completed a short pre-task

questionnaire to provide information about their professional background. The answers revealed that all participants had received higher education in Translation or related fields, such as Interpreting or Literature and Linguistics, and they all have several years of experience in the industry (at least five and up to 24). As far as PE is concerned, one participant reported to be experienced in this activity, six mentioned that they received training (at university or in a company context) but do not perform PE regularly, and only one had not received training and never performed PE. Five participants reported to be familiar with marketing or creative translation, while two engaged in such a task only for revision, and only one participant had no experience in this field. Regarding Trados Studio, five participants had extensive experience with this tool, but tended to use it less over the last few years, and the remaining three had little to no experience. Overall, the background of the participants, notably in terms of familiarity with PE and Trados Studio, can be attributed to their roles at LanguageWire. All participants were part of either the Linguistic Quality Management team or the Linguistic Services & Terminology team, which are both composed of linguists focused on either assessing translation quality and investigating customer feedback or maintaining linguistic databases. Furthermore, the participants' reduced use of Trados Studio is explained by LanguageWire's proprietary CAT tool, Smart Editor, which is more commonly used within the company.

To ensure familiarity with the study's objectives and the different tasks involved, participants were provided with detailed guidelines one week before the experiment. Additionally, they were encouraged to watch short instructional videos explaining the technology used in the experiment.

The study involved post-editing several short texts with the specific instruction to approach the task creatively. Participants completed PE both with and without the AI Professional Companion plug-in for Trados Studio. This setup was designed to compare UX, productivity, and translation quality in both scenarios. The model used in the plug-in for this study was GPT-4. Participants were instructed to use the Qualitivity plug-in to capture productivity metrics and to record their screens using Microsoft Teams for data collection purposes. The study was structured into several tasks:

- **Preparation**: participants familiarised themselves with the task by reading the guidelines and watching instructional videos.[52]
- **Pre-task questionnaire**: participants completed a short questionnaire about their background.
- **Post-Editing 1 (PE1)**: post-editing of T1 without using the plug-in, followed by a brief UX questionnaire.

[52] AI Professional Companion: https://community.rws.com/product-groups/trados-portfolio/trados-studio/b/weblog/posts/unleash-the-power-of-translation-with-ai-professional-plugin-for-trados-studio and Qualitivity: https://www.youtube.com/watch?v=vKnbt25YyKA (both consulted: 30/04/2025)

- **Post-Editing 2 (PE2)**: post-editing of T2 with the plug-in, using only one prompt (specifically designed for this task, based on the PE guidelines below). This was followed by a detailed UX questionnaire, which included the UX survey designed for the BF (see 9.3) and additional questions regarding the perception of AI assistance.
- **Voluntary Task – Post-Editing 3 (PE3)**: for linguists interested in exploring further or with available time, this optional task involved post-editing T3 using multiple prompts (including custom ones), followed by a detailed UX questionnaire.
- **Linguistic Quality Assessment (LQA)**: after PE1 and PE2 were completed, each linguist evaluated the post-edited versions of these two texts provided by the colleague working in the same language pair, using LanguageWire's error typology.

The estimated total duration for completing the PE tasks and questionnaires was approximately two hours. However, several participants exceeded this estimation, as explained hereafter.

Participants were also provided with the following PE guidelines, adapted from generic LanguageWire guidelines for PE:

"*Accurately post-edit the machine translation of the source text so that it sounds fluent and natural in the target language.*
Ensure that no information has been added or omitted.
Edit any inappropriate content.
Restructure sentences in the case of incorrect or unclear meaning.
Apply spelling, punctuation and hyphenation rules.
The fully post-edited text must be at a publishable quality level and indistinguishable from that of a translation performed by a native speaker.
Ensure that the translation is appropriate for the target audience. Please note that using creative language is crucial for this customer. It is essential that the translation is not too literal. Since the texts are about high-end products, it is very important to use creative, elegant and refined language.
You can refer to the corresponding source webpages for contextual information on the source texts.
You may also refer to external resources or perform information searches as necessary."

Additional guidelines were provided to help participants effectively use the AI Professional Companion plug-in, including:

- How to use the generation button (a dedicated shortcut [Ctrl + Shift + G] was also configured for this action)

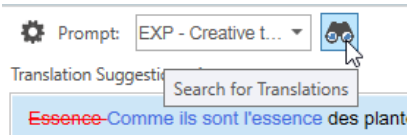


Figure 125: Generation button of the AI Professional Companion in Trados Studio.

- How to insert an alternative translation (another shortcut [Ctrl + Shift + I] was also created for this action)

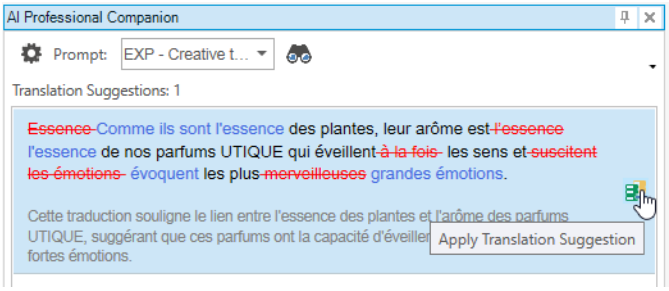


Figure 126: Inserting alternative translations generated by the AI Professional Companion in Trados Studio.

- Where to find the explanations (text in grey, displayed below the alternative translations)
- How to show or hide the differences between the initial MT output and the alternative translation

Figure 127: Show/hide differences in the alternative translations generated by the AI Professional Companion in Trados Studio.

- How to spot the origin of the applied pre-translation based on the segment metadata: after inserting an alternative translation suggestion, the origin goes from "AT" (Origin: Automated Translation; Provider: LanguageWire MT) to "NMT" (Origin: Neural Machine Translation; Provider: AI Professional) as illustrated below:

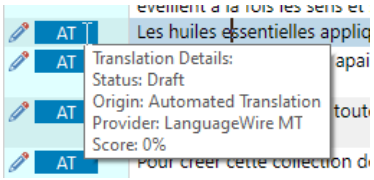


Figure 128: Original MT segment metadata.

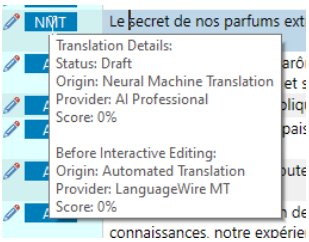


Figure 129: Segment metadata after inserting an alternative translation generated by the AI Professional Companion in Trados Studio.

Further instructions were included for T2 and T3.

- In the case of T2, these comprised:
  - **Use of one prompt**: linguists were instructed to use only one prompt to ensure consistency across the results obtained from the different participants. After testing various options, the following prompt was selected: *"Creative translation: Translate this text using creative, elegant, and refined language."*
  - **Testing alternative translations**: linguists were encouraged to explore alternative translations as much as possible, but were not required to use them for every segment. They were advised to select alternatives based on their specific needs.
  - **Regenerating alternatives**: linguists were allowed to regenerate alternative translations as many times as they liked for the same segment.

- For T3, the additional instructions emphasised how participants could experiment with different prompts, including:
  - **Selecting predefined prompts** from the drop-down menu

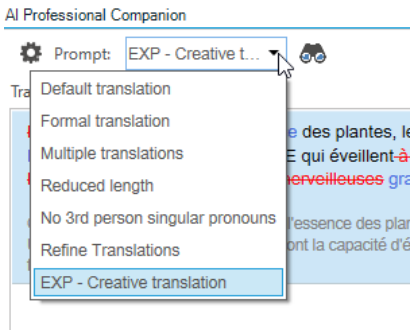


Figure 130: Predefined prompts available in the AI Professional Companion in Trados Studio.

- **Creating custom prompts** by
  - Clicking the wheel icon to open the settings

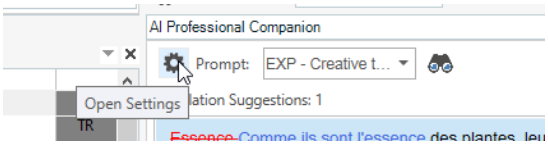


Figure 131: Accessing the settings of the AI Professional Companion in Trados Studio.

  - Under the Prompts tab, clicking "Add"

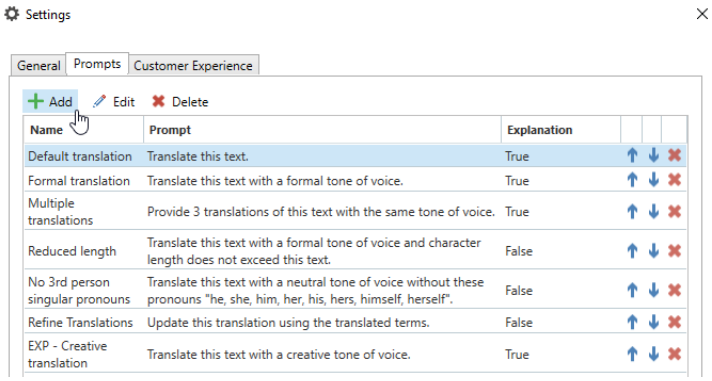


Figure 132: Prompt settings of the AI Professional Companion in Trados Studio.

- Adding a name and the prompt text. Linguists could tick the box named "Show an explanation with the translation" if they wanted the AI Professional Companion to display an explanation in grey, below the alternative translation.

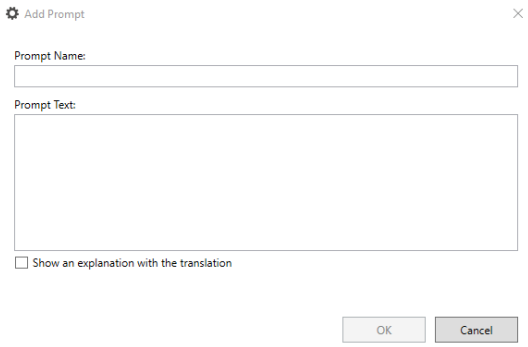


Figure 133: Prompt creation in the AI Professional Companion in Trados Studio.

10.2.4 Analysis of Benchmarking Pillars

This section presents the results of this study, organised around the three BF pillars: productivity, UX and quality. Each pillar is discussed in detail, with a focus on the impact of using the AI Professional Companion plug-in in the PE process. Additionally, extra parameters related to the use of the AI assistant are also analysed. In this section, the participants are named based on their target language, followed by a numerical identifier (e.g., ES1 for the first participant working with Spanish, DE2 for the second participant working with German). In addition, it should be noted here that although most participants (except one) expressed their interest in testing more prompts with T3, only five completed this task fully (NL1 did not perform PE3 at all, and ES1 and DE2 only completed a portion of it).

10.2.4.1 Productivity Pillar

The average completion times (Table 72) show a general trend where the completion time decreases when working with the AI assistant (PE2 and PE3), suggesting that the use of alternative translations contributes to increasing productivity.

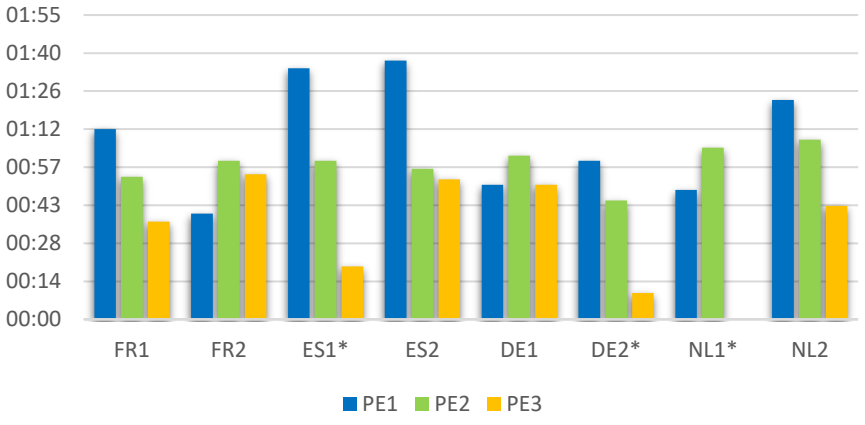
Task	Average completion time	Average productivity (WPH)
PE1	01:08	497.68
PE2	00:58	521.03
PE3	00:59*	650.63*

* For PE3, the average completion time is calculated only for the five participants who performed PE3 for the entire text, and the average productivity is calculated only for five participants (those who performed PE3 for the entire text)

Table 72: Completion times per PE task.

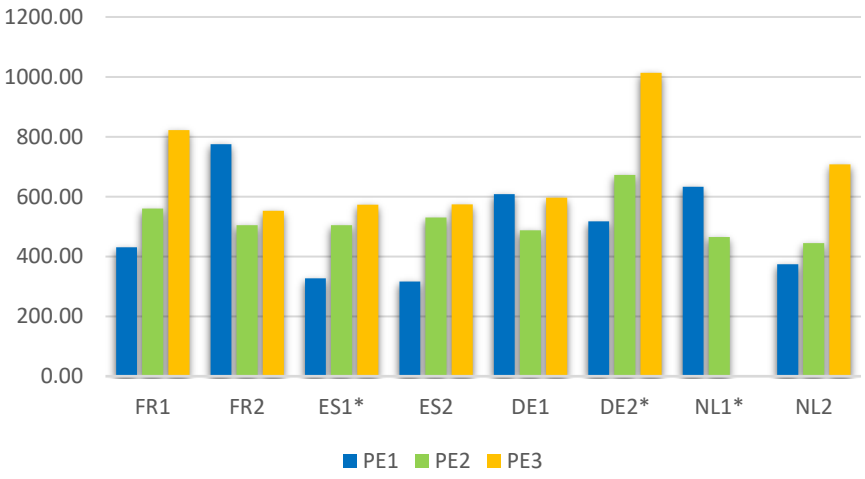
The similar values for PE2 and PE3 indicate that even with the time spent creating custom prompts, productivity remains higher when using the AI assistant. However, this result may be influenced by the fact that the PE3 average is based on a smaller sample of participants.

Furthermore, considering that in PE2 and PE3, post-editors had to wait for the alternative translations to be generated, the productivity gain could potentially be even higher if the time required for generation could be reduced.



* Incomplete (ES1, DE2) or no PE3 completion (NL1)

Figure 134: Task completion times per task and participant.



* Incomplete (ES1, DE2) or no PE3 completion (NL1)

Figure 135: Productivity (WPH) per task and participant.

The results across different language pairs exhibit some variability (Figures 134 and 135). FR1 and NL2 share the same pattern, whereby the time spent on the tasks decreases progressively from PE1 to PE2,

and then from PE2 to PE3. This could be attributed to the support provided by the AI assistant. Alternatively, the decrease in the task completion time could be due to the participants becoming more familiar with the task as they progressed through the study. In contrast, FR2 and NL1 displayed the opposite trend, with task completion times increasing when working with the AI assistant. This suggests that, for these participants, having to process alternative translations may have introduced additional complexity or inefficiencies into the PE process. In the case of DE1, however, apart from a slight increase for PE2, the completion time remained relatively stable. This indicates that the AI assistant's impact on productivity was less pronounced for this participant. For both ES participants, PE1 was the most time-consuming task. Apart from the observations above, there was no clear pattern that could be attributed to the language pairs themselves. This suggests that other factors, such as familiarity with the task or the effectiveness of the AI assistant, might have been more significant in influencing productivity.

The TER scores are relevant to further understand productivity, as they reflect the amount of editing required to transform the raw MT output into the final, post-edited version of the target text. A high TER score indicates that the MT output was heavily edited, while a low TER score suggests that the MT output was of better quality, requiring fewer edits.

Task	**Average TER**
PE1	66.66
PE2	58.95
PE3	62.36*

* PE3 is calculated for seven participants (NL1 is excluded; and for ES1 and DE2, the TER score is computed based on the number of words post-edited)

Table 73: Average TER per PE task.

The average TER scores per PE task (Table 73) show a general tendency to edit less of the MT output when working with the AI assistant. This is reflected in the decrease in TER from PE1 to PE2. Specifically, the average TER for PE1 is 66.66, indicating a significant amount of editing was required on the MT output. In PE2 and PE3, the TER scores are lower (58.95 and 62.36, respectively). This decrease could be explained by a change in perception: when seeing the suggestions provided by the AI assistant, linguists were able to compare different translations and may have found MT output to be comparatively better. Alternatively, the higher TER in PE1 may also indicate that the first text was more challenging, which resulted in a lower MT quality, and, in turn, a higher TER.

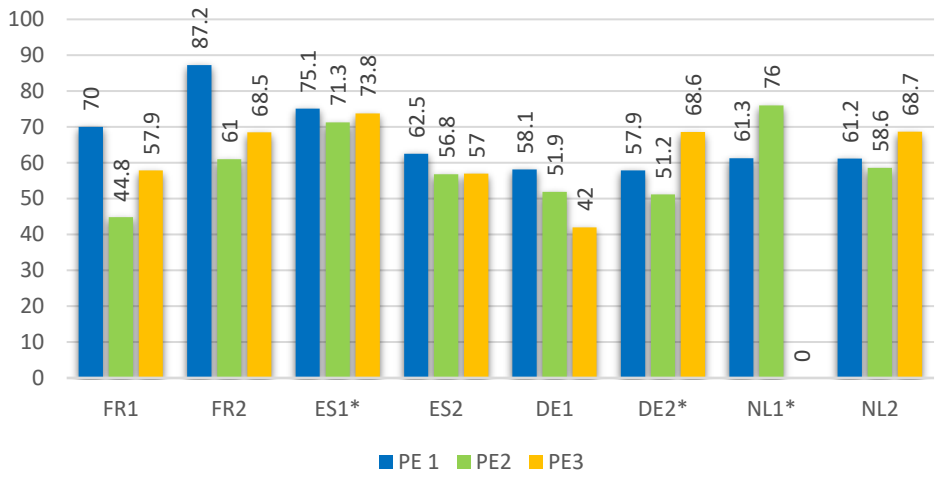


Figure 136: TER scores per task and participant.

When examining the TER scores for each participant (Figure 136), the lowest TER for PE2 was observed in FR1, FR2, ES1, DE2, and NL2. However, DE2 and NL2 displayed a higher TER in PE3, suggesting that these participants may have needed additional adjustments when using multiple prompts or generating alternative translations.

10.2.4.2 UX Pillar

The post-task questionnaires for PE1 and PE2 included a UX questionnaire that served to attribute a UX scoring for the PE environment in both cases (with and without the plug-in). The questionnaire used is the PE UX questionnaire (presented in 9.3), which is based on the SUS (Brooke, 1996) and consists of the 10 statements listed below:

- **Usefulness**: I think the post-editing environment makes my work easier.
- **Speed**: I find the post-editing environment slow.
- **Effectiveness**: I feel that the post-editing environment allows me to achieve my goals with accuracy and completeness.
- **Efficiency**: I need to use a high amount of resources and effort to achieve my goals when working in the post-editing environment.
- **Enjoyment**: I find the post-editing environment pleasant to use.
- **Ease of use**: I find the post-editing environment complex to use.
- **Memorability**: I think that the functionalities of the post-editing environment are easy to remember.
- **Error recovery**: I find that it is difficult to undo mistakes in the post-editing environment.

- **Learnability**: I think that the post-editing environment is easy to learn.
- **Clarity**: I find the post-editing environment's structure unclear and disorganised.

Respondents rate their agreement with each statement on a 5-point Likert scale (where 1 = strongly disagree, 5 = strongly agree). A UX score was then calculated based on the following logic:

- For odd-numbered questions: [User Rating] – 1 = ___ points
- For even-numbered questions: 5 – [User Rating] = ___ points
- Add all points
- Multiply total by 2.5 (to convert scale to 100)
- Average all user's scores together

The scores thus obtained are presented in Table 74. The average UX score is higher for PE1 (74.1) compared to PE2 (69.4). This suggests that, overall, participants had a slightly better UX when working without the AI assistant.

	PE1	**PE2**
FR1	75.0	77.5
FR2	57.5	65.0
ES1	90.0	72.5
ES2	77.5	80.0
DE1	82.5	67.5
DE2	70.0	65.0
NL1	75.0	65.0
NL2	65.0	62.5
Average	***74.1***	***69.4***

Table 74: UX scores per PE task.

The decrease in the UX score for PE2 could be attributed to several factors. Notably, the lack of familiarity with the UI and uncertainty about how the AI assistant works likely contributed to the lower ratings. Participants may not have fully understood what to expect from the AI suggestions, which was confirmed through open-ended responses in the post-task questionnaires.

In addition, the average UX scores per dimension are also presented in Table 75 to further examine the experience of participants for individual UX elements. The most impacted dimension is Speed, with an increase from 1.88 to 3.5, indicating a dissatisfaction with this element when working with the plug-in,

which is further discussed in the analysis of participants' feedback in 10.2.5, and likely is the factor that mostly contributed to the decrease in the UX score in PE2. Other dimensions for which the UX appeared to be slightly better in PE1 than in PE2 include Usefulness, Efficiency, Ease of use and Memorability, which could be explained by the lack of familiarity with the plug-in. On the other hand, Enjoyment improved (3.63 in PE1, 4.13 in PE2), which indicates that the post-editors had a more pleasant experience when working with alternative translations. It should, however, be noted that this phenomenon could also be explained by the enthusiasm of trying out a modern technology. Other dimensions which improved with the plug-in include Effectiveness, Error recovery, Learnability and Clarity, which are more surprising results, given that participants had never worked with this technology before.

	PE1	**PE2**
Usefulness	3.88	3.63
Speed*	1.88	3.50
Effectiveness	3.63	3.75
Efficiency*	2.75	3.00
Enjoyment	3.63	4.13
Ease of use*	2.00	1.75
Memorability	4.25	4.13
Error recovery*	1.63	2.25
Learnability	4.13	4.38
Clarity*	1.63	1.75

*Indicates negative statements in the questionnaire (therefore, the lower the score, the better the UX for these elements)

Table 75: Average UX scores per UX dimension.

10.2.4.3 Quality Pillar

The LQA results provide a detailed analysis of error types and severities across the three PE tasks. The LQA was performed using the LanguageWire typology of error categories and error severities (Tables 76 and 77).

Error Category	Error Subcategory	Definition
Accuracy	Mistranslation	The target content does not correspond to the source content.
	Addition	The target text includes information which is absent in the source.
	Omission	Information present in the source is missing from the target text.
	Under/ overediting	Not editing enough or editing too much the pre-translation (TM match or MT), including leaving TM fuzzy matches unedited.
	Untranslated	The content which should have been translated in the target language was left in the source language.
Non-adherence	Briefing instructions	The target text does not comply with the instructions specified in the brief, including non-adherence to reference material.
	Style guide	The target text does not comply with the style guide.
Language Quality	Grammar & Syntax	The target text contains issues related to grammar or syntax.
	Spelling & Capitalisation	The target text contains issues related to spelling or capitalisation of words.
	Punctuation	Punctuation is used incorrectly.
	Locale convention	Inappropriate conventions are used in the translation. This applies, among others, to address, date, currency, measurement, and telephone number formats.
Terminology	Inconsistent (job level)	Terminology is translated inconsistently in the target text.
	Inconsistent (company TB)	The terminology used in the target text does not comply with the terminology provided by the customer.
	Wrong terminology	The terminology used in the target text is not the one expected for the domain.
Formatting & Tags	/	Issues related to formatting (bold, underlined, italics) and tags (including missing tags or tag order errors).
Other	/	This category can be used to annotate issues which are not represented in the typology defined above.

Table 76: Error typology of the LanguageWire LQA model.

Severity Levels	Definition
Neutral	To be used when entering additional information or comments without affecting the final quality score (no penalty points for this level); this severity level can be used to indicate preferential changes.
Minor	To be used when the annotated error does not entail a change or loss in meaning but affects the overall quality of the translation (for example, decreased readability).
Major	To be used when the annotated error entails a change in meaning and may therefore confuse or mislead the final reader. If a minor error is particularly visible (because it is part of a title or header, for example), it can also be assigned a major severity so that it has a higher weight in the final score.
Critical	To be used when the annotated error implies an important change in meaning which could give rise to damaging consequences in terms of health, safety, legal, geopolitical, financial, ethical or moral considerations.

Table 77: Error severities of the LanguageWire LQA model.

A. Analysis of error categories

Overall, no clear pattern can be identified in terms of the total number of errors per participant (represented in Figure 137). Some linguists made a similar number of mistakes in all tasks (3 in PE1, 2 in PE2 and 4 in PE3 for FR1), some made more errors when using the AI assistant (the translations produced by NL2 contained 1 error in PE1, 2 in PE2 and 6 in PE3), while others made slightly fewer errors throughout the task (DE1 made 9 errors in PE1, 8 in PE2 and 6 in PE3). However, it is worth noting that certain participants consistently made more errors than others. For example, ES1, FR2 and ES1 systematically made a high number of errors, whereas FR1 and DE2 tended to systematically produce fewer errors – which could also reflect the subjective dimension inherent to LQA, since some evaluators tend to be stricter than others.

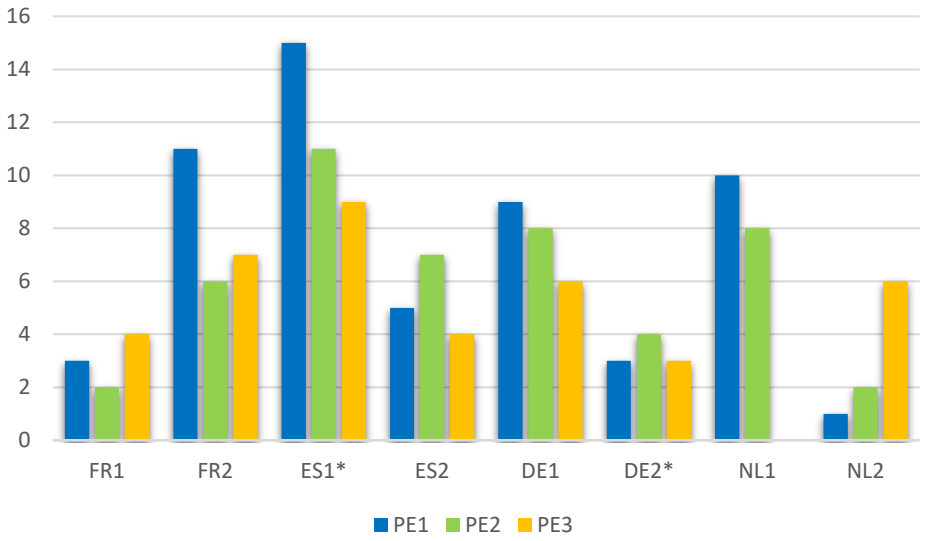


Figure 137: Total errors per task and participant.

The case of NL2 is particularly striking, since this participant produced a very low number of errors in PE1 (1), a somewhat higher number of errors in PE2 (2), but a significantly higher number of errors in PE3 (6). Although this could suggest that it was challenging for this participant to maintain a high quality while exploring the custom prompts, a closer analysis of the recordings revealed that this participant used three of the default prompts and did not create a new one. This result may therefore have been caused by fatigue or a perceived higher difficulty of the third text. Conversely, in the case of FR2 and DE1, the number of errors is lower in both tasks in which the AI assistant was used, which indicates a potential quality benefit.

To gain further insight into the different error categories found in each task and language pair, summaries are presented in Tables 78-80.

	FR1	FR2	ES1	ES2	DE1	DE2	NL1	NL2	Avg.
Accuracy	0	9	2	1	0	1	0	0	1.63
Non adherence	0	0	0	0	0	0	0	0	0.00
Style	0	1	0	0	1	0	4	0	0.75
Language quality	3	1	7	4	5	2	6	1	3.63
Terminology	0	0	3	0	3	0	0	0	0.75
Formatting & Tags	0	0	0	0	0	0	0	0	0.00
Other	0	0	3	0	0	0	0	0	0.38
Total	3	11	15	5	9	3	10	1	7.13
		Grand total	57						
	Average per linguist		7.13						

Table 78: Error categories found in PE1.

	FR1	FR2	ES1	ES2	DE1	DE2	NL1	NL2	Avg.
Accuracy	1	5	0	0	1	0	2	0	1.13
Non adherence	0	0	0	0	0	0	0	0	0.00
Style	1	1	1	0	2	1	1	0	0.88
Language quality	0	0	6	4	1	3	4	2	2.50
Terminology	0	0	3	0	4	0	0	0	0.88
Formatting & Tags	0	0	0	0	0	0	1	0	0.13
Other	0	0	1	3	0	0	0	0	0.50
Total	2	6	11	7	8	4	8	2	6.00
		Grand total	48						
	Average per linguist		6						

Table 79: Error categories found in PE2.

	FR1	FR2	ES1*	ES2	DE1	DE2*	NL1*	NL2	Avg. (excl. *)
Accuracy	0	4	0	0	0	1	N/A	2	1.20
Non adherence	0	0	0	0	0	0	N/A	0	0.00
Style	0	2	1	0	2	2	N/A	1	1.00
Language quality	3	1	6	4	1	0	N/A	3	2.40
Terminology	1	0	2	0	3	0	N/A	0	0.80
Formatting & Tags	0	0	0	0	0	0	N/A	0	0.00
Other	0	0	0	0	0	0	N/A	0	0.00
Total	4	7	9	4	6	3	0	6	5.40
Grand total			39						
Average per linguist (excl. *)			7.8						

Table 80: Error categories found in PE3.

Across all PE tasks, no error related to non-adherence was found, and only one formatting error was spotted. Similarly, the number of stylistic errors was consistently low for all tasks (between 0 and 2), with only one exception (NL1: 4 errors in PE1), and the average number of stylistic errors per PE task is between 0.75 and 1 (excluding the data for participants who did not complete PE3). Accuracy errors were slightly more frequent, with an average per task higher than one, and decreasing when using the AI assistant (1.63 in PE1, 1.13 in PE2 and 1.20 in PE3). However, accuracy errors were generally low among all participants (usually comprised between 0 and 2), except for FR2 (up to nine errors in PE1). This outlier likely affected the results. Terminology errors were also constant across the different PE tasks (around 0.7-0.8 on average), with several participants not making any terminological mistake (FR2, ES2, DE2 and NL2). No particular pattern differentiated the performance when using the AI assistant for those who did produce terminology errors. Language quality errors appeared to have occurred more often across all tasks, but a notable decrease was observed when using the AI assistant (3.63 in PE1, 2.50 in PE2 and 2.40 in PE3 on average). This suggests that the AI assistant was beneficial in improving language quality.

B. Analysis of error severities

The distribution of errors by severity highlights further nuances (Tables 81-83). No critical error was found in any of the tasks, and the errors spotted were mostly minor, and a few were major. The number of both minor and major errors tended to decrease from PE1 to PE2 and from PE2 to PE3. This could

indicate a positive impact of the AI assistant; however, these results could also be explained by the lower number of participants who fully completed PE3.

	FR1	FR2	ES1	ES2	DE1	DE2	NL1	NL2	Avg/PE task
Neutral	0	0	2	4	0	0	0	0	0.75
Minor	3	8	12	0	8	3	9	1	5.50
Major	0	2	1	1	1	0	1	0	0.75
Critical	0	0	0	0	0	0	0	0	0.00

Table 81: Error severities in PE1.

	FR1	FR2	ES1	ES2	DE1	DE2	NL1	NL2	Avg/PE task
Neutral	0	0	1	4	0	0	0	0	0.63
Minor	2	3	10	2	8	4	7	2	4.75
Major	0	3	0	1	0	0	1	0	0.63
Critical	0	0	0	0	0	0	0	0	0.00

Table 82: Error severities in PE2.

	FR1	FR2	ES1*	ES2	DE1	DE2*	NL1*	NL2	Avg./PE task (excl. *)
Neutral	0	2	1	3	0	1	N/A	1	1.20
Minor	4	3	8	1	6	2	N/A	5	3.80
Major	0	2	0	0	0	0	N/A	0	0.40
Critical	0	0	0	0	0	0	N/A	0	0.00

Table 83: Error severities in PE3.

In addition, the TER scores (see Table 73 and Figure 136) indicate that post-editors tended to edit the MT output less when working with alternatives (with PE2 showing the lowest average TER across tasks). This does not seem to have impacted the final quality; therefore, it could indicate that the MT output for PE2 was of a higher quality.

The average number of errors per linguist decreased slightly in PE2 (6 vs 7.13 in PE1 and 7.8 in PE3). Regarding the frequency of specific error categories, the main difference appeared for language quality, with a reduction in the average number of errors (PE1: 3.63, PE2: 2.50, PE3: 2.40). This could suggest

that PE with AI assistance leads to increased language quality. Accuracy errors seemed to decrease only slightly with AI assistance (PE1: 1.63, PE2: 1.13, PE3: 1.20). Conversely, style (PE1: 0.75, PE2: 0.88, PE3: 1.00) and terminology (PE1: 0.75, PE2: 0.88, PE3: 0.80) errors were found to increase when using the AI assistant. Nevertheless, these variations remain minimal and do not appear to signify any direct impact of AI assistance on quality.

10.2.4.4 Interplay Between the Pillars

As done for the previous application of the BF, the relationship between each pilar was also studied. To that end, a Spearman test was performed to find out whether correlations could be identified between them.

Pair	**PE1** **Spearman's ρ (*p*-value)**	**PE2** **Spearman's ρ (*p*-value)**
UX ↔ Productivity	-0.47 ($p = 0.24$)	0.49 ($p = 0.22$)
UX ↔ MT quality	0.00 ($p = 1.00$)	-0.32 ($p = 0.44$)
UX ↔ Post-edited text quality	0.38 ($p = 0.35$)	0.27 ($p = 0.52$)
Productivity ↔ MT quality	-0.02 ($p = 0.96$)	-0.64 ($p = 0.09$)
Productivity ↔ Post-edited text quality	0.26 ($p = 0.51$)	-0.28 ($p = 0.51$)
Post-edited text quality ↔ MT quality	0.59 ($p = 0.13$)	0.59 ($p = 0.12$)

Table 84: Relationship between the BF pillars (Spearman test) in the second BF application.

The correlation analysis (Table 84) revealed weak trends, but none reached statistical significance, likely due to the small sample size ($N = 8$). In PE1, UX showed a weak negative relationship with productivity ($\rho = -0.47$, $p = 0.24$), suggesting that higher UX scores might marginally correlate with slower task completion. However, this pattern was reversed when using the AI assistant (PE2), where UX and productivity were weakly positively correlated ($\rho = 0.49$, $p = 0.22$). This could imply that the use of AI altered the dynamics between UX and efficiency, though the effect remains uncertain.

The correlation between UX and MT quality was null in PE1, and PE2 exhibited a slight negative correlation ($\rho = -0.32$, $p = 0.44$), hinting that poorer raw MT output might have slightly diminished perceived UX in PE2. The relationship between productivity and MT quality strengthened from PE1 ($\rho = -0.02$, $p = 0.96$) to PE2 ($\rho = -0.64$, $p = 0.09$), suggesting that linguists may have worked more slowly when using the AI tool, though this trend was not definitive.

Both tasks showed a consistent, moderate positive correlation between post-edited text quality and MT quality ($\rho = 0.59$ for PE1 and PE2), indicating that better initial MT output tended to align with higher final quality. However, the lack of significance across all metrics underscores the need for caution in drawing conclusions. These observations may serve as a foundation for future studies with larger samples to explore whether the trends – particularly the potential shift in UX ↔ Productivity dynamics with AI assistance – reflect meaningful patterns.

10.2.5 Further Insights into User Perception

In addition to the BF pillars, it was deemed relevant to further explore the user's perception of the PE2 and PE3 tasks, based on the answers to the post-task questionnaires and user behaviour.

10.2.5.1 User Perception after Performing PE2

First, this analysis turns to the user's satisfaction after performing PE2.

A. Average Satisfaction

The participants' average satisfaction with the AI Professional Companion was 3.63 out of 5, indicating a slightly below-average experience. This suggests room for improvement in the tool's usability and performance.

B. Pre vs Post-Task Perception

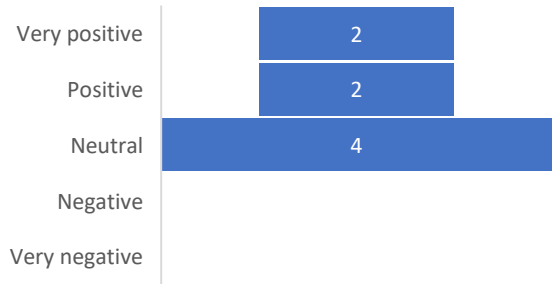


Figure 138: Pre-task perception of AI assistance.

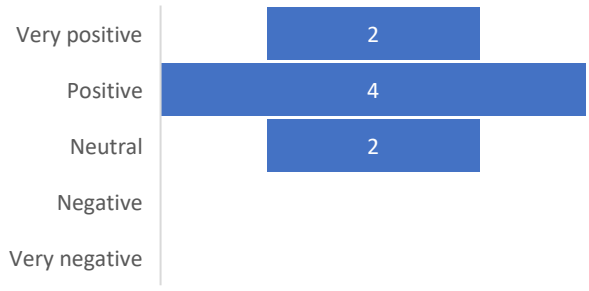


Figure 139: Post-task perception of AI assistance.

The comparison between pre- and post-task perceptions (Figures 138-139) reveals that some initial scepticism about the tool diminished after practical experience. Two participants who initially gave neutral ratings later provided positive feedback.

C. Frustrations

However, six of the eight participants reported frustrations during the task. The most frequently mentioned issues were related to speed (3 participants), followed by specific frustrations such as parts of segments remaining unchanged despite requiring improvements (2 participants), insufficient variation in alternative suggestions (1 participant), and typographic issues, particularly with quotes (1 participant).

D. Output Quality and Stability

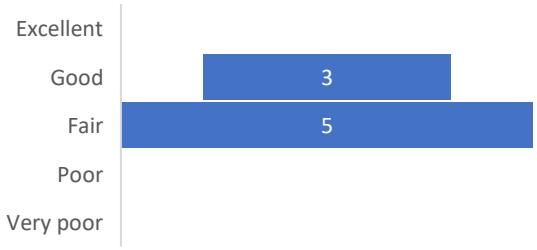


Figure 140: Output quality of the AI Professional Companion for PE2.

Overall, participants had a positive view of the AI Professional Companion's output quality, with five participants rating if as "Fair", and three as "Good" (Figure 140). Nevertheless, several issues were identified. The most commonly reported problems included the outputs perceived as overly literal (three participants), inconsistencies in pronouns (two participants), contextual mismatches, cultural reference inaccuracies and syntax errors.

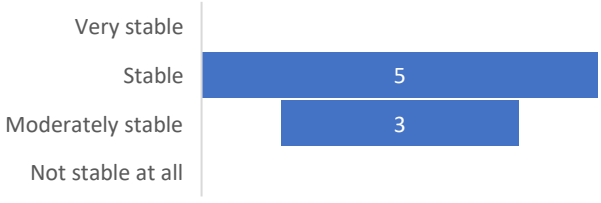


Figure 141: Output quality stability of the AI Professional Companion for PE2.

Participants generally found the output quality to be rather stable, with three considering it "moderately stable" and five "stable" (Figure 141).

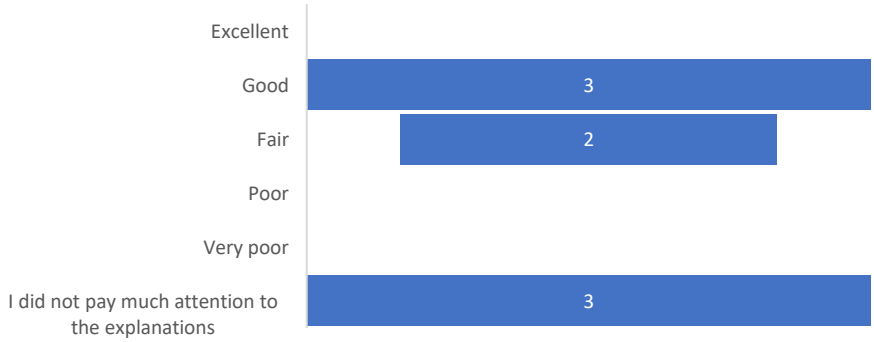


Figure 142: Explanations quality of the AI Professional Companion for PE2.

Regarding the explanations accompanying alternative translations, three participants admitted not having paid attention to them. Among those who did, the quality of explanations was rated as "Fair" by two and "Good" by three (Figure 142). However, one linguist considered the explanations to be unnecessary, stating that "*As a linguist, you are capable of evaluating the quality of the output without reading how and why the tool 'chose' this option*".

E. Efficiency

Six of the eight participants felt that the AI Professional Companion helped them complete their PE tasks more efficiently. Some participants highlighted specific benefits, including getting inspiration (3), receiving suggestions of better quality compared to traditional MT, saving time when searching for synonyms, and the convenience of having the suggestions integrated into the UI. However, two participants expressed mixed feelings, noting that the suggestions were sometimes not useful (2), and that the generation time was too slow. One participant summarised this duality as follows: "*Sometimes it did, sometimes it didn't. Some suggestions were useful and saved me a few extra clicks of looking things up online, or spending additional time trying to think of suitable synonyms myself*."

F. Intuitiveness

Participants generally found the AI Professional Companion intuitive to use, with seven considering it "Very intuitive" and one "Extremely intuitive" (Figure 143).

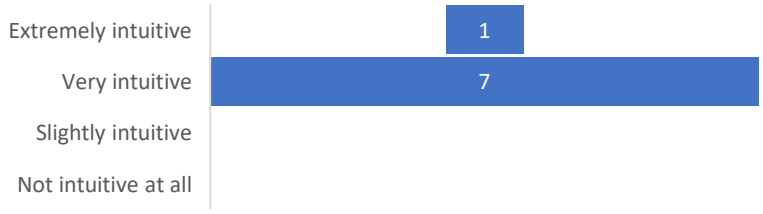


Figure 143: Intuitiveness of the AI Professional Companion for PE2.

However, when asked to elaborate, the participants mentioned that some aspects of the tool could be difficult or confusing, such as its focus on the current segment only, and the inability to copy-paste parts of the output.

G. Additional Features

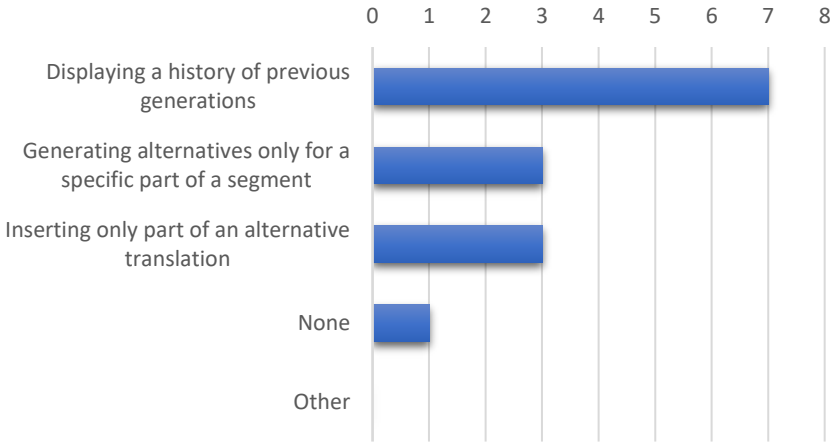


Figure 144: Additional features requested for the AI Professional Companion.

Post-editors were also asked to rate their interest in potential additional features for the AI assistant. As shown in Figure 144, the most popular feature was the history function to display previous generations (7), followed by the ability to generate alternatives for specific parts of a segment (3) and to insert only parts of a suggestion (3). Two participants suggested additional features, including the ability to display all suggestions simultaneously and an interactive feedback dialogue function. The participant who suggested the latter explained that "*it could be useful that the human translator enters a draft or introduces a part of the sentence that the AI Companion can then tweak or review (for instance propose revisions with a richer vocabulary). This could maybe even be in a dialogue style like the web version of ChatGPT. I could have asked the AI to tell me more about the literary works the texts were referring to".* This idea highlights the potential for making AI assistants more interactive.

H. Other Use Cases

When asked for which other use cases AI assistance seemed helpful, participants deemed the following to be the most relevant (Figure 145): asking for synonyms for selected words in the segment (8), verifying that guidelines were correctly applied (7), providing in-context definitions for selected words in the segment (6) and generating alternative translations for a selected fragment inside a segment (4). Additional use cases mentioned by the respondents included ensuring the use of inclusive language and supporting quality assessment tasks.

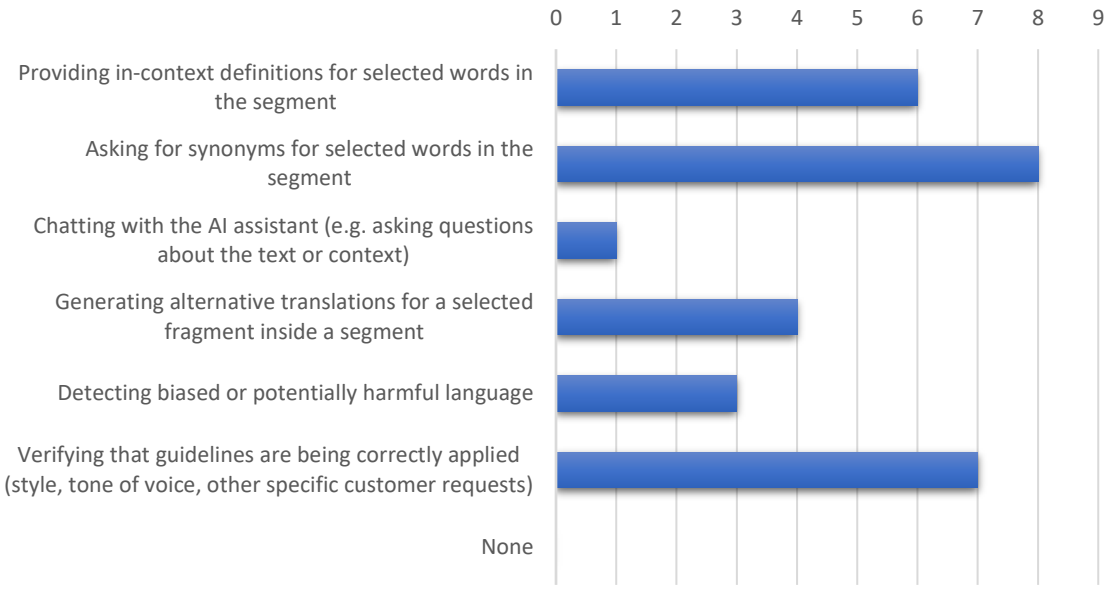


Figure 145: Other uses cases identified for AI assistance.

It should be noted that the comments revealed a high polarisation in user perception. While one participant noted: "*It helped at times, but I still had to do a lot of research and think a lot about the translation of some segments*," another described it as a transformative technology: "*I think this is a game changer when it comes to using MT in creative/marketing texts*."

10.2.5.2 User Perception after Performing PE3

Having covered PE2, the analysis now turns to PE3. The discussion is also based on the prompt usage and questionnaire answers; however, it should be noted that this analysis is based on the participation of five linguists only.

A. Prompts Usage

An overview of the prompts used in PE3 is presented in Figure 146, and the distribution per participant appears in Figure 147.

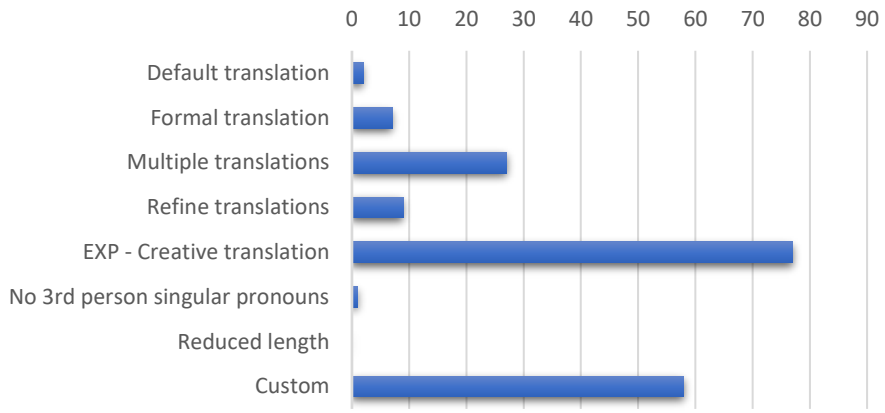


Figure 146: Prompt usage (predefined and custom) in PE3.

Apart from the custom prompt created beforehand for this experiment, the most frequently used predefined prompts in PE3 were "Multiple translations" followed by "Refine translations", "Formal translation," and "Default translation". "No 3rd person singular pronouns" was used only once, and "Reduced length" was not used at all, likely because they were irrelevant to the text type.

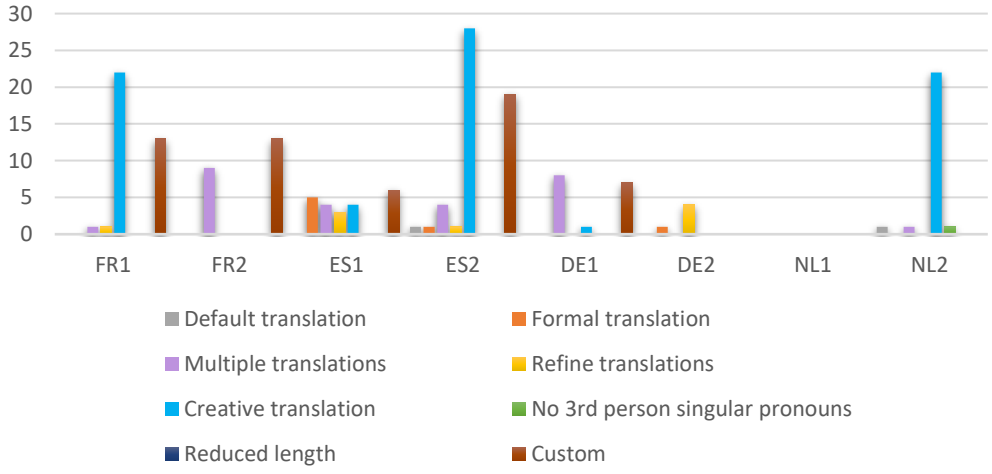


Figure 147: Prompts usage (predefined and custom) per participant in PE3.

As far as individual participant behaviour is concerned, a tendency to rely solely on "creative translation" and to test custom prompts can be noted (e.g. FR1, ES2). The participants who created fewer custom prompts seemed to have tested more predefined options (e.g. ES1, DE1).

B. Predefined Prompts

The quality of outputs generated by predefined prompts was generally viewed positively, with four participants considering it "Fair" and one "Good" (Figure 148).

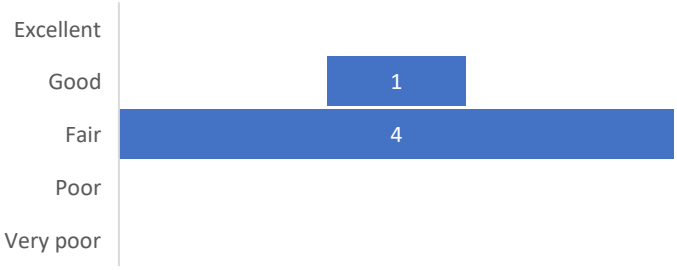


Figure 148: Output quality of predefined prompts in PE3.

However, three of the five participants found the available prompts insufficient for their needs. Suggestions for additional predefined prompts included making the "Creative translation" prompt a default option and adding prompts for inclusive language. Another respondent also expressed interest in combining multiple predefined prompts to enhance flexibility.

C. Custom Prompts

All five participants in PE3 created custom prompts. Examples of prompts created include:

- FR1: *"Perfume: Update this translation using an adequate language to describe a perfume. Make sure to use French quotes (chevrons)"*
- ES2: *"Translate this text using a formal tone, avoiding gender-specific words and with a creative language",* then added: *"Use the formal address"*
- DE1: *"Provide 5 translations of this text creatively playing with the phrasings to create an elegant and appealing text"*

These examples highlight the need to follow the guidelines by prompting the AI assistant to focus on refining the style of the output. The prompt created by DE1 also suggests that this participant valued having multiple alternatives.

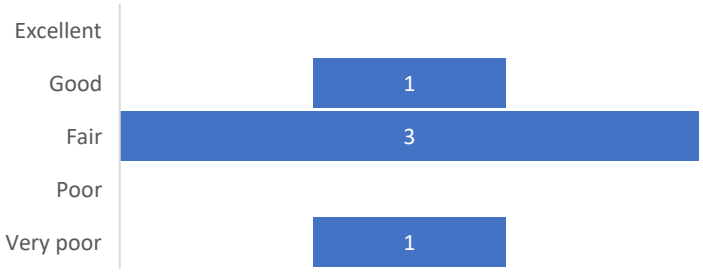


Figure 149: Output quality of custom prompts in PE3.

The quality of outputs generated using custom prompts was mixed (Figure 149) and generally not as high as those produced by predefined prompts. However, the experience seemed to vary significantly from one participant to the other. For example, one considered the output to be "Very poor", yet FR2 explained that they "*created the prompt 'literary translation' and the generated output was better than the standard ones*".

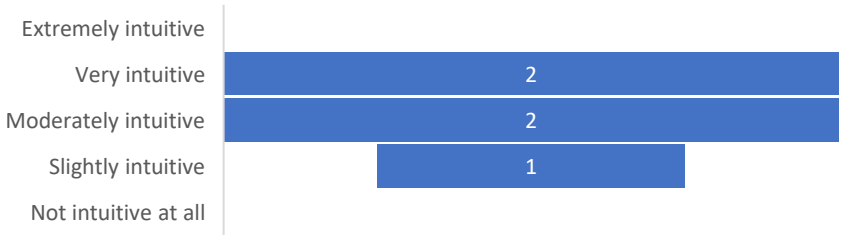


Figure 150: Intuitiveness of custom prompt creation in PE3.

The intuitiveness of working with custom prompts was overall lower than in the case of predefined ones, with only two participants considering it to be "Very intuitive" (Figure 150). Participants reported challenges in creating effective custom prompts due to their lack of familiarity with prompt design. Three participants specifically highlighted such difficulties:

- *"The creation of the prompt, to know how to efficiently communicate with the engine."*
- *"Interface is easy. Finding the right words to create an efficient prompt is not easy. I guess you can easily learn how to do it but without experience of AI it's strange to trust that the tool will 'interpret' what you want. Maybe it's 'too' intuitive."*

- *"The main challenge I see here is that we need to build a relevant prompt, and that requires practice."*

Despite these challenges, four out of five participants believed that training on prompt design would improve their results and make the process easier. Three participants expressed strong interest in receiving such training, mostly to become more efficient, but also to understand how the technology works and what to expect in terms of quality. The other two were open to it but felt that gaining experience through practice could suffice. One also suggested that providing a checklist of customisable parameters (e.g., tone of voice, formal/informal address, sentence length, style) could simplify prompt creation and improve the outcomes.

It should be noted here that NL2 did not fill in the last questionnaire for PE3 (they mentioned that they did not create their own prompt because it seemed to be "*too cumbersome without any training or assistance*", therefore they did not feel comfortable in answering the questionnaire). However, this participant provided relevant feedback separately. Their feedback revolved around the limitations of segment-based translation:

> *"My biggest disappointment is that because we're in the CAT tool environment, we're still bound by segments and sentence-based translation, whereas for something creative this often feels like a limitation, especially because English is a very compact language where you can push a lot of information into one sentence, and Dutch doesn't really like that […]. the real CAT revolution I think will be when segments are not necessary anymore, and you review a translation of the text in its original layout, especially for creative translation".*

This suggests that beyond AI assistance, moving away from the segmentation approach would be beneficial for texts which require a certain degree of creativity. This feedback also confirms the observations of Dragsted (2006) and (Pym, 2011) regarding the decontextualisation resulting from text segmentation.

10.2.5.3 Analysis of Generations

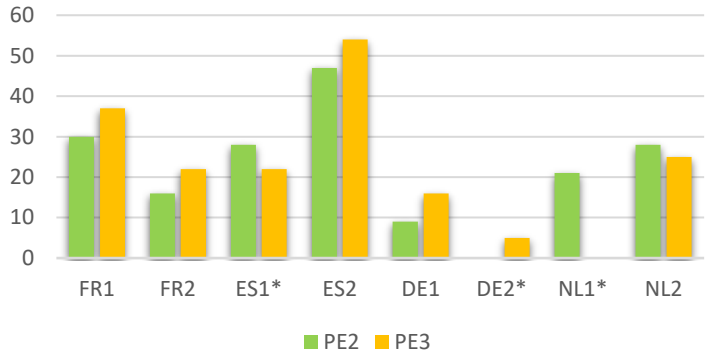
To gain further understanding of the participants' use of the AI assistant, an analysis of the generations was conducted. It revolves around the number of generations per participant, and whether these generations were inserted or not, and for those inserted, whether they were edited or not.

The number of generations per task varied, with participants generating more alternatives on average during PE3 (30.80) than PE2 (22.38), as shown in Table 85.

Task	Average number of generations
PE2	22.38
PE3	30.80*

* For five participants, excluding those who did not complete PE3 entirely

Table 85: Average number of generations per PE task.



* Here, the generations are shown for ES1 and DE2, who completed PE3 partially. Note that NL1 did not perform PE3.

Figure 151: Number of generations in PE2 and PE3 per participant.

This increase in generations for PE3 may be partially attributed to the smaller sample of post-editors considered. However, when analysing the data by language pair, differences emerged (Figure 151). FR1, FR2, ES2, and DE1 generated more alternatives in both PE3 and PE2 compared to PE1. In contrast, NL2 and ES1 (who completed PE3 partially) showed an opposite trend, generating fewer alternatives during PE3. Interestingly, DE2 did not generate any alternatives at all during PE2.

Overall, participants working into FR and ES generated more alternatives than those working into DE and NL, suggesting potential differences in how post-editors across language pairs interact with or perceive the utility of alternative translations.

Furthermore, the number of generations did not seem to have a negative impact on productivity. For example, ES2, who made the highest number of generations, achieved similar completion times as DE1, who relied less on alternative translations (more precisely, ES2 generated 47 times in PE2, and completed this task in 57 minutes, whereas DE1 generated 9 times in the same task, and took 1 hour and 2 minutes to complete it).

It should also be noted that the post-editors did not make the same use of the plug-in, particularly when it comes to regenerating (i.e. using the plug-in more than once for the same segment). Participants were instructed to use the plug-in based on their needs and were allowed to regenerate alternative translations as many times as they deemed necessary. As a result, some behavioural differences could be observed regarding regenerations, as presented in Table 86.

A clear difference can be identified between post-editors who never regenerated alternatives, and those who relied more heavily on regenerations. For instance, DE2 never retriggered the plug-in in any of the tasks, and both FR2 and DE1 did not retrigger it in PE1 and only regenerated a limited number of times in PE2 (4 and 2, respectively). In contrast, a high proportion of the generations of ES2 are regenerations (40% in PE2 and 59% in PE3). Interestingly, this behaviour did not seem to have a significant impact on productivity, as the two post-editors who exhibited diametrically opposed behaviours have similar productivity scores: in PE3, DE1 regenerated twice, and ES2 regenerated 32 times, yet they completed this task in 51 minutes and 53 minutes, respectively.

	PE2	**PE3**
FR1	9 (30%)	16 (43%)
FR2	0 (0%)	4 (18%)
ES1*	14 (50%)	13 (59%)
ES2	19 (40%)	32 (59%)
DE1	0 (0%)	2 (13%)
DE2*	0 (0%)	0 (0%)
NL1*	N/A	N/A
NL2	15 (54%)	11 (44%)
Average	***8.14 (25%)***	***11.14 (34%)***

* Marks participants who did not finish PE3.

Table 86: Number of regenerations per PE task.

It is worth mentioning that while the plug-in offers alternative translations on the fly, its output is not provided immediately, which results in users having to wait a few seconds. To further study the impact of this limitation on the PE experience here, an average response time was calculated based on the observation of 40 generations across different language pairs and tasks. The resulting average response time was 9 seconds per generation. While this value may be considered negligible when the number of

generations is limited, it impacted the experience of post-editors, who tended to rely more on the alternative translations and thus generated alternatives more often. For instance, FR1 used the plug-in 37 times in PE3. As a result, the waiting time for this participant is slightly over 5 minutes, which accounts for 15% of the total time to complete this task (37 minutes), as illustrated in Figure 152. On average, the time spent waiting for the output to be generated is 6% for PE2 and 11% for PE3, with individual values ranging from 2% (DE1 in PE2) to up to 15% (for both FR1 and ES2 in PE3).

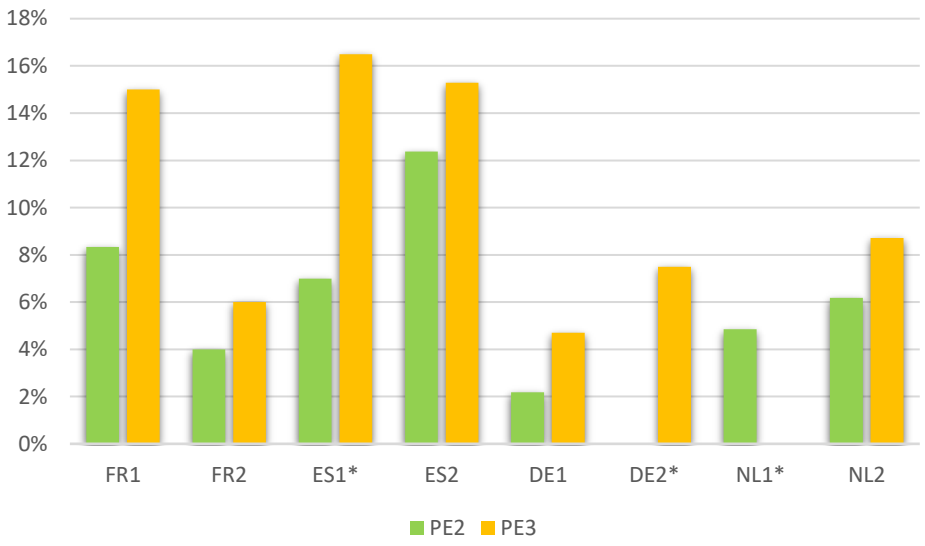


Figure 152: Percentage of time spent waiting for the output to be generated in PE2 and PE3, per participant.

The insertions (happening when a participant chose to insert the alternative translations) are also indicative of the perceived usefulness of the AI assistant.

	PE2	**PE3***
% insertions	19%	14%
% unedited insertions	31%	19%

* For five participants, excluding those who did not complete PE3 entirely

Table 87: Percentage of generations inserted.

The frequency and quality of insertions differed between PE2 and PE3. As shown in Table 87, 19% of the inserted suggestions were edited during PE2, compared to 14% in PE3. Additionally, the percentage

of unedited insertions decreased from 31% in PE2 to 19% in PE3. This could indicate that working with a good, predefined prompt is more efficient than having multiple choices of prompts.

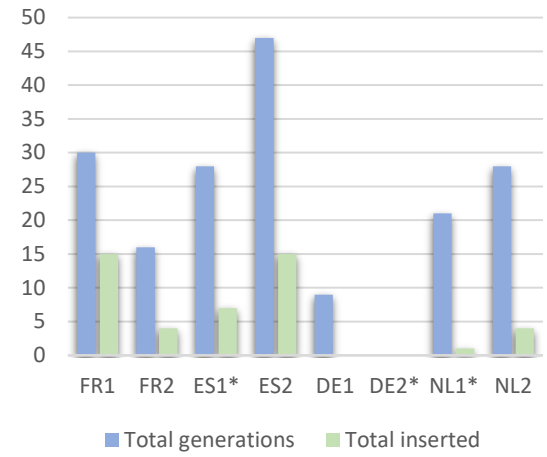


Figure 153: Number of inserted generations per participant in PE2.

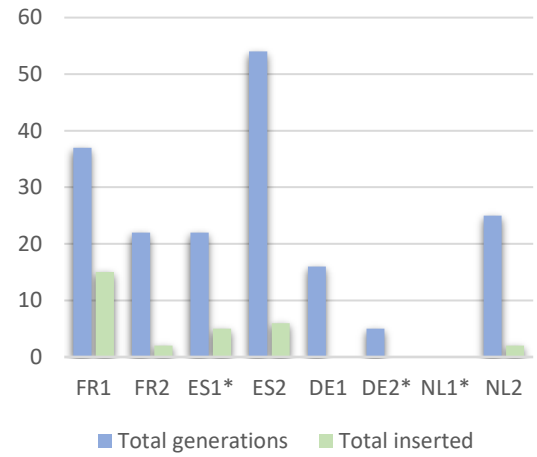


Figure 154: Number of inserted generations per participant in PE3.

* Note that insertions are shown for ES1 and DE2, even though they did not complete PE3.
No insertion is shown for NL1 as this participant did not perform PE3.

When studying insertions by language pair (Figures 153-154), FR1 and ES2 made the highest number of insertions during PE2, though significantly fewer during PE3. All participants inserted a lower number of suggestions during PE3, which could be explained by the lower quality of suggestions due to the lack of training in prompt design. It should also be noted that DE2 did not generate or insert any alternatives during PE2.

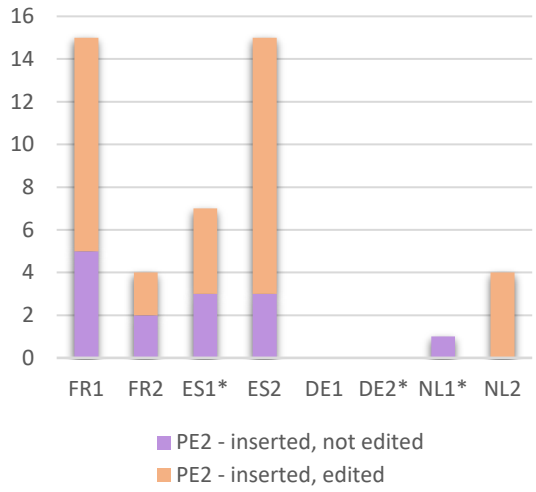


Figure 155: Number of edited and not edited insertions in PE2.

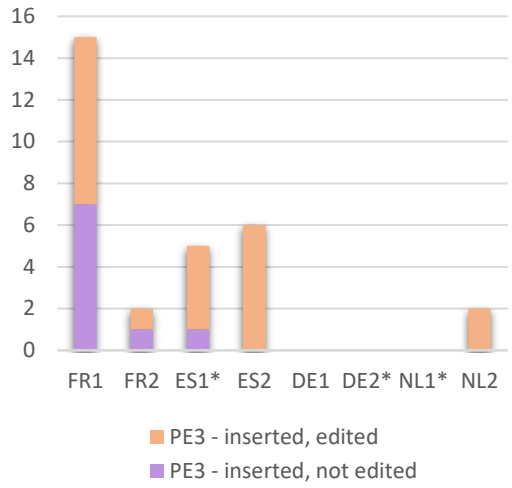


Figure 156: Number of edited and not edited insertions in PE3.

* Note that insertions are shown for ES1 and DE2, even though they did not complete PE3.
No insertion is shown for NL1 as this participant did not perform PE3.

Furthermore, it is also relevant to analyse whether the inserted generations were used as such or further edited. As shown in Figures 155-156, most of the inserted suggestions were edited. In addition, the participants who made more insertions tended to edit these insertions more frequently. For example, FR1 and ES2 not only inserted more suggestions but also edited them significantly. NL2 systematically edited all their insertions across tasks, and ES2 also edited all their insertions in PE3. This finding shows that the insertions could be used as a basis for PE. Indeed, the generations that were not inserted could still be helpful, notably to give inspiration. In the example below, ES2 did not insert the output directly, but modified the MT output based on the alternative translation:

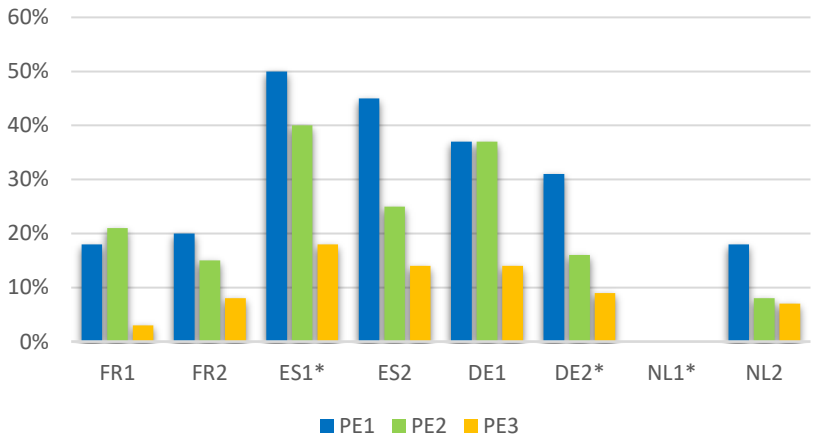
> ***Source**: On the Road is itinerant in nature.*
>
> ***MT**: On The Road es un viaje por la naturaleza.*
>
> ***Alternative translation (AI assistant)**: "En El Camino" es de naturaleza itinerante.*
>
> ***PE**: En El Camino es itinerante por naturaleza.*

In this example, "itinerant in nature" was translated into "viaje por la naturaleza" ("a journey through nature") by the MT system, which is incorrect, as it has a different meaning. Specifically, the source text does not mention "nature" as a physical environment but rather uses "in nature" to describe an inherent characteristic (being itinerant). On the contrary, the alternative translation provided by the AI assistant suggested to use "de naturaleza itinerante" ("of an itinerant nature"), which is closer to the source, both syntactically and semantically. After inserting the alternative translation – probably because

it seemed more accurate – ES2 adjusted the style by changing it to “itinerante por naturaleza” (“itinerant by nature”), which sounds more idiomatic.

10.2.5.4 Analysis of the Post-Editing Process

Apart from the BF pillars and the deep dive into user insights and use of the alternative translations, the annotation of screen recordings provided valuable information as to how working with the AI assistant affected the PE process itself. Notably, it is relevant to study the time spent outside the CAT tool (i.e., to search to further information or definitions) in each scenario. It should be noted here that ES1 and DE2 are excluded from most of the data (as they did not participate in PE3), and no recording data was available for NL1 (due to an issue in the technical setup, only the actions performed in the CAT tool were recorded for this participant).



* Here, the time is shown for ES1 and DE2, who completed PE3 partially. Note that no data was available for NL1.

Figure 157: Percentage of time spent outside the CAT tool per task and participant.

The analysis of the screen recordings showed varying amounts of time spent performing research activities outside the CAT tool, such as searching for information, definitions, synonyms, or other related material. A trend emerged where participants generally spent less time outside the tool as they progressed through the tasks (with time completion following this pattern: PE1>PE2>PE3), in particular for FR2, ES2, and NL2 (Figure 157). This reduction could indicate increasing familiarity with the task or greater reliance on the AI assistant for support. For DE1, there was no significant difference in time spent outside the tool between PE1 and PE2, but a noticeable drop occurred in PE3. Interestingly, FR1 spent the least amount of time outside the tool during PE3, but slightly more time in PE1 than PE2.

These findings suggest that the AI assistant may help reduce the need for external research. However, individual differences and task-specific factors, such as the complexity of the text or familiarity with the prompt system, likely influenced these outcomes.

In order to gain further insight into the types of research activities performed by the post-editors, the screen recordings were annotated to classify the activities conducted outside the CAT tool. The annotation process consisted of first identifying all user actions and classifying them, iteratively refining categories based on observed patterns. Although the intention behind each search activity would be pertinent to find out, this proved challenging to infer from the recordings only. Therefore, the classification system adopted here was based on clearly observable actions, resulting in the following categories:

- **SL search**: looking for a part of the ST in the SL (via a web search or a monolingual dictionary). In this case, the possible intention behind the search was likely to find related content or better understand the source text.
- **TL search**: looking for a part of the TT in the TL (via a web search or a monolingual dictionary). Here, the possible intention could be related to finding context in the target language and gaining inspiration for wording.
- **Translation**: looking for a translation of a word or expression (by explicitly including "translation" or "in [language]" in the search or directly consulting in a bilingual dictionary or concordancer).
- **Synonym**: looking for a synonym (by explicitly specifying "synonym" in the search or directly searching in a dictionary of synonyms).
- **ST**: consulting the source text, possibly to check the structure.
- **Repetition**: returning to a previously performed search.

	PE1	**PE2**	**PE3***
SL search	8.43	6.29	2.80
TL search	13.29	10.14	3.80
Translation	11.43	6.14	1.60
Synonym	1.14	1.29	0.60
ST	3.71	2.00	0.40
Repetition	6.86	1.71	1.40

* Seven for PE1 and PE2, five for PE3

Table 88: Average number of searches per post-editor.

A noticeable pattern emerged across all search types (Table 88), whereby the frequency of searches dropped significantly in PE3 compared to PE1 and PE2. For example, the average number of TL searches went from 13.29 in PE2 to 10.14 in PE2, and to 3.8 in PE3.

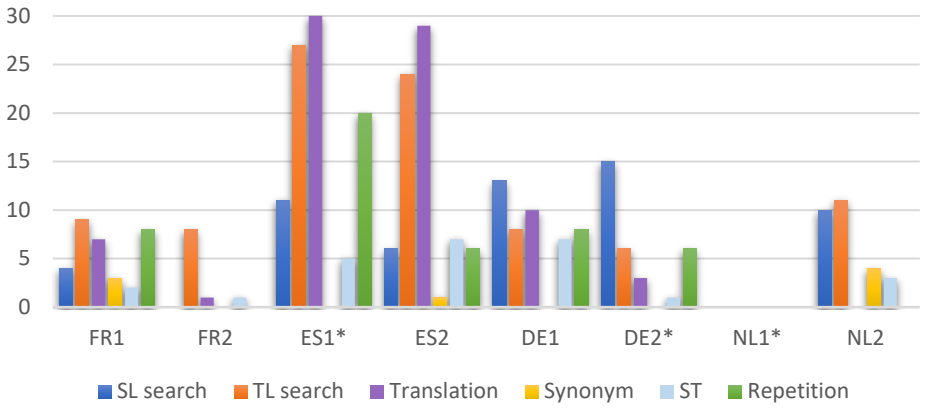


Figure 158: Searches performed per participant in PE1.

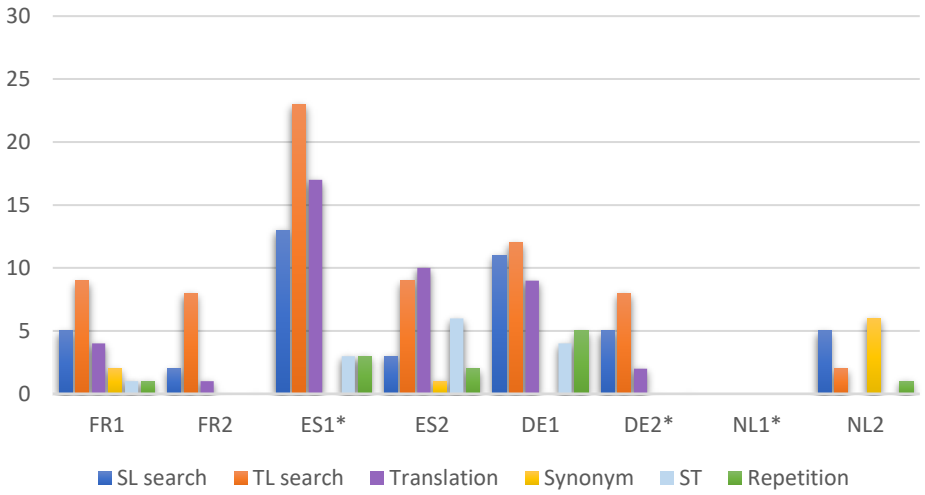


Figure 159: Searches performed per participant in PE2.

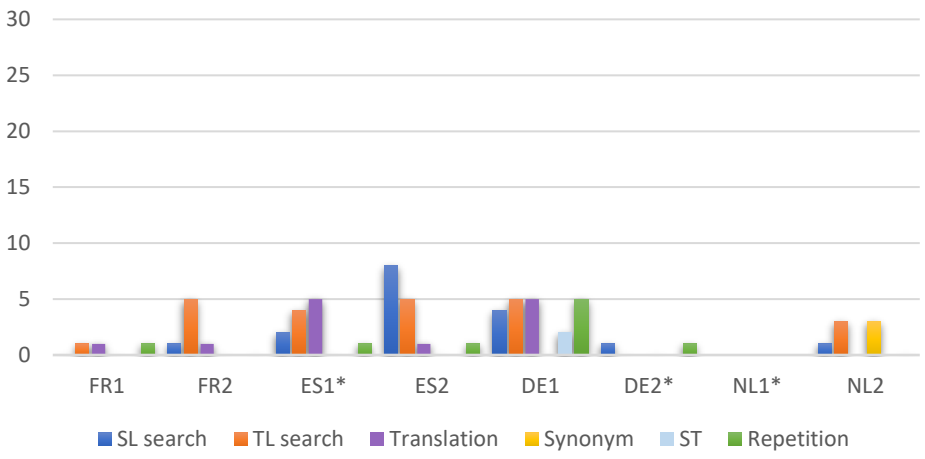


Figure 160: Searches performed per participant in PE3.

The detail of searches by task and post-editor (Figures 158-160) revealed both similarities and differences across participants, depending on the type of activity.

SL searches remained relatively stable between PE1 and PE2. For example, FR1 performed four searches in PE1 and five in PE2, while DE1 had 13 in PE1 and 11 in PE2. However, a sharp drop occurred in PE3, with FR1 registering zero searches and DE1 dropping to four.

TL searches also remained stable between PE1 and PE2 – with a notable exception for ES2, for whom this value dropped from 24 in PE1 to 9 in PE2. ES1 was among the participants performing the highest number of TL searches. The analysis of the recordings showed that this post-editor often searched for perfumes from other brands with descriptions containing the same ingredients, which highlights a need for this person to find similar texts in the TL. As with SL searches, TL searches dropped across all participants in PE3, with a maximum of five searches.

Furthermore, TL searches were generally more frequent compared with SL searches. Although TL searches decreased in PE3, they remained the most widely performed type of search in this task. This suggests that checking fluency in the TL is a common need of post-editors and remains necessary even when alternative translations are available.

Translation searches displayed a gradual decline across tasks, suggesting a reduced reliance on external translation references from PE1 to PE2 and PE3.

Synonym searches followed a mixed trend. For FR1, synonym searches gradually decreased (three in PE1, two in PE2, zero in PE3), likely because the alternative translations provided sufficient alternative phrasings. However, NL2 did not show a clear pattern (four in PE1, six in PE2, three in PE3), suggesting that the AI assistant was not equally effective for all users regarding the need for synonyms.

ST consultation also gradually decreased. This observation is rather concerning, as it may indicate that post-editors tended to refer less to the original text when working with an AI assistant, which could have detrimental consequences in terms of document-level quality. Alternatively, it could also be hypothesised that since the three STs were similar, consulting the original material was not deemed necessary after the first PE task.

Repetitions of previous searches followed a downward trend, which was partly influenced by ES1 being an outlier (with 20 repeated searches in PE1). Despite this decline, repetitions were still present for most post-editors, confirming the idea that the PE process is inherently non-linear (do Carmo, 2017).

Additional observations from the analysis provide further insights into PE behaviours observed in this study. Although all texts were similar product descriptions, T1 may have been slightly more difficult, as noted by two post-editors after performing all tasks. This could also be a phenomenon due to perception, since the fact that all texts share a common structure and terminology resulted in T2 and T3 being perceived as easier (participants were already familiar with this type of content after working on T1). For example, since all texts were about perfumes, they all alluded to the fragrance pyramid (with "top notes, heart notes and based notes"). After performing this search in PE1, participants seem to have remembered the information and did not perform the same search in the subsequent tasks.

In terms of the PE process itself, it was observed from all screen recordings that post-editors systematically returned to the beginning of the text for a self-revision phase (except ES1 and DE2 in PE3, as they did not perform this task). In addition, alternative generations were sometimes triggered during the revision stage, which further emphasises the non-linear nature of PE.

These observations help inform recommendations for AI support tailored to PE needs. Since TL searches were the most frequent across all tasks, AI assistants could integrate examples of common wordings in the TL based on existing documents to help with fluency. The high number of SL and translation searches suggests that embedding these functionalities within the CAT tool could also be beneficial. Additionally, since synonym searches were common for some participants, incorporating automated synonym suggestions could further support post-editors.

10.2.6 Summary & Outcomes

This study provided a comprehensive analysis of the three BF pillars, complemented with direct user insights, and analyses of the generations performed and the impact of the AI assistant on the PE process. Therefore, the answers to the research questions can be formulated as follows.

- Impact on PE: how does working with alternative translations generated via prompts affect PE in terms of
    - Productivity?

The completion times for PE2 and PE3 were generally shorter than PE1, despite the time spent for alternatives to be generated. This suggests that working with alternative translations appeared to have a positive impact on productivity. However, variability across language pairs and participants' feedback highlights the need for further refinement, as not all post-editors experienced a consistent productivity gain.

In addition, the quality of the alternative translations received mixed feedback, which likely also affected productivity. While some participants viewed the quality positively and felt it was better than standard MT, several issues were also noted, involving contextual mismatches, overly literal translations, and occasional linguistic errors. These findings suggest that the AI assistant offers valuable support but that improvements in handling creative translations could yield even higher productivity gains.

The analysis of screen recordings also revealed that post-editors continued to perform searches outside the CAT tool when working with the AI assistant, although the number of searches tended to decrease when alternative translations were available. This finding shows that further AI support could contribute to increasing productivity by minimising the time spent outside the tool.

- UX?

UX scores for the PE environment were slightly lower for PE2 than for PE1, potentially due to the lack of familiarity with the AI assistant. Participants reported notable frustrations with speed and usability and mentioned that they would find additional features valuable. The overall perception of AI assistance was found to be more positive overall after performing all tasks, which suggests that usability and acceptance may improve as users gain more experience with the tool.

However, working with custom prompts was considered less intuitive compared with using the predefined ones. This shows that training in prompt design – which appeared relevant to most participants – could be highly beneficial and contribute to a better UX.

- Quality?

The analysis suggests that working with alternative translations has a modest but observable impact on the final quality of post-edited texts. While no significant overall reduction in errors was found, the use of AI assistance, particularly predefined prompts (PE2), led to fewer language quality errors and a slight decrease in accuracy errors. However, style and terminology errors remained stable or increased minimally, indicating that AI assistance may not uniformly improve all error categories. In addition, the results varied from one linguist to the other, with some benefiting from AI suggestions while others

produced more errors. Thus, while working with AI-generated alternatives seems promising in terms of language quality, the effectiveness depends on user adaptation and task-specific factors. Further research with larger samples would be relevant to validate these trends.

- Post-editors' perceptions: how do post-editors perceive this new PE technology?
    - Do they find it useful?

Overall, participants found the AI assistant useful, with six out of eight indicating that it helped improve efficiency. Several limitations, notably the slow response times and insufficient variation in suggestions, likely hindered its perceived usefulness. Despite those issues, participants seemed to value the inspiration provided by the alternative translations.

- What are its strengths and weaknesses?

Strengths included the convenience of having alternative translations integrated into the UI and an output quality sometimes deemed better than raw MT (for certain language pairs). However, the main benefit of using alternative translations was to gain inspiration, which resulted in less time spent outside the tool looking for common phrasings in the TL.

The main weakness of the AI assistant used in this study is related to its technical performance, as the time spent generating alternatives was often deemed too long by the post-editors. A possible solution to this issue would be to generate alternatives for the selected prompt as soon as the project is opened. While such an approach would not cater for all use cases (notably changing the prompt on the fly), it would certainly enhance the UX. Additional weaknesses are tied to output quality (particularly stability) and usability challenges, including difficulties in prompt design.

- How could it be further tailored to PE needs?

Based on participants' feedback, possible improvements to the AI Professional Companion include:

- Increasing generation speed
- Increasing output quality and stability, notably by enhancing the ability of the model to produce less literal translations, handle pronouns and typography correctly and consistently, and deal better with context and cultural references. Quality improvements could also include more variation (i.e. compared with raw MT).
- Making more features available, in particular:
    - Copy-paste parts of the output (or ability to insert only part of a suggestion)
    - Generating alternatives for part of a segment
    - History of previous generations

- More default prompts (e.g. creative translation, inclusive language)
- Combining different prompts
- Checklist-guided prompt creation (e.g. with customisable parameters)
- Interactivity (e.g. with a chat-like modality to provide feedback on the fly)

In addition, the observation that post-editors still needed to perform various searches (notably TL searches), even when working with the AI assistant, suggests that an embedded feature providing examples of similar texts in the TL would also be beneficial. And beyond AI assistance, revisiting the segment-based approach would also be beneficial for creative texts.

Apart from alternative translations, participants also provided feedback on other possible use cases for AI assistance, and deemed the following to be the most relevant:

- Asking for synonyms for selected words in the segment (8)
- Verifying that guidelines were correctly applied (7)
- Providing in-context definitions for selected words in the segment (6)
- Generating alternative translations for a selected fragment inside a segment (4)

Inclusive language and LQA assistance were also mentioned.

This study examined the impact of AI-assisted PE with alternative translations within Trados Studio, with a particular focus on PE productivity, UX, and translation quality. The findings highlight both the potential benefits and challenges of AI-enhanced PE, showing that while AI can support efficiency and provide useful alternative translations or relevant inspiration, there are also technical performance issues and usability concerns across users and language pairs.

While benchmarking AI-enhanced PE against traditional PE is a relevant and necessary step, other factors such as cognitive load could also be considered in future research. The use of alternative translations requires post-editors to engage with an additional layer of information, potentially increasing cognitive effort. A promising next step would thus lie in examining the correlation between the results obtained herein and the cognitive effort, assessing whether AI assistance adds to or alleviates the complexity of the PE task.

Looking ahead, numerous enhancements could improve AI-enhanced PE workflows. Potential developments could include, for instance, automatic retrieval of similar texts in the TL to provide examples, or more interactive features, which provide real-time feedback and multimodal support, including voice-based interaction. However, apart from AI-based assistance, other changes could be

introduced in CAT tools to further adapt them to the task of PE, such as moving away from the traditional segment-based layout.

The willingness of linguists to learn and adapt to recent technologies is an encouraging finding in the present study. However, ensuring that translation tools meet the actual needs of their users requires continued research. More studies of this kind are essential to refine AI-driven solutions and develop CAT tools that truly enhance the PE experience.

V. CONCLUSIONS

As the technological transformation in the language industry is fundamentally reshaping the work of language professionals, novel approaches are needed to assess and improve translation tools, notably from the perspective of PE. This research brought together the industry and academia via an industrial collaboration with LanguageWire in an effort to develop and apply a benchmarking method to evaluate the PE environment. The approach was grounded in early conceptual underpinnings behind translation tools, which originally envisaged a symbiosis between humans and machines, whereby technology would extend the capabilities of language professionals.

A thorough analysis of the literature allowed for the identification of three core pillars – namely, productivity, UX and quality – as essential dimensions for evaluating PE environments. Benchmarking was selected as a suitable method for the evaluation of PE environments, and special attention was placed on addressing gaps in existing assessment models. First, an exploratory user survey allowed for further contextualising the present work based on concrete feedback and real needs expressed by respondents. Then, based on the three core pillars, a BF was designed to provide a structured and user-centred approach to assess translation tools for PE. While the productivity dimension was defined as the number of words post-edited per hour and thus did not require further investigation, the other pillars required to be meticulously defined in order to establish assessment processes. Therefore, two key research activities were conducted.

First, a PE UX questionnaire was designed based on the importance attributed by post-editors to various UX dimensions. The exploratory user survey allowed for identifying the UX elements that post-editors value the most. Based on this finding, a PE UX questionnaire was designed following the SUS model to provide a practical tool for evaluating the UX of translation tools in a PE context.

Then, MT quality was explored as a potential predictor of the effort (and potentially productivity) through a dedicated experiment. The outcomes of this experiment revealed significant variability in the PE experiences of individual language professionals when it comes to error identification, severity assessment, and effort perception. The initial hypothesis, according to which specific MT errors would systematically result in a greater PE effort, could not be confirmed. This finding therefore challenged the feasibility of designing an MT evaluation model based primarily on the PE effort. Nevertheless, it shed light on subjectivity in PE and set the ground for further research on that topic.

Once the BF was fully defined, it was then applied to two real-world scenarios in order to test its effectiveness in evaluating PE environments.

First, the BF was applied to Smart Editor at LanguageWire. This analysis revealed key trends in terms of productivity, UX, and quality over 2024. MT quality was found to directly influence both UX and productivity. Furthermore, high productivity did not always correlate with high satisfaction. Quality remained consistently high throughout the year, even during periods of increased workload or reduced satisfaction. A mid-year decline in satisfaction, linked to higher productivity, was observed, but the situation stabilised in October, which coincided with improved MT quality and increased user engagement. Based on these findings, coupled with a detailed analysis of the main pain points expressed by users, a roadmap for 2025 development was established. It included notably improvements to spellchecking and QA features, as well as continued investment in MT quality to boost both productivity and satisfaction.

The second application of the BF consisted of an experiment to evaluate prompt-based alternative translations for PE using the AI Professional Companion plug-in within Trados Studio. The outcomes of this experiment highlighted the potential benefits and challenges of AI-augmented PE. While AI assistance improved efficiency and provided useful alternatives, technical and usability issues were observed across the different participants and language pairs. Here, the identified enhancements included automatic retrieval of similar texts and interactive features with real-time feedback. Beyond AI, changes to CAT tools, such as moving away from the segment-based visualisation, could also contribute to further adapting existing tools to PE needs. While the willingness of linguists to adapt to modern technologies is encouraging, continued research is essential to ensure tools meet user needs and enhance the PE experience.

While this research provides valuable insights into the evaluation of PE environments, it is important to acknowledge its limitations. In particular, the experiments were conducted with a limited number of participants. Therefore, larger-scale studies would be necessary to strengthen the findings and identify more robust patterns. Moreover, the rapid pace of advances, particularly in AI, means that the landscape of translation technology is constantly evolving. As a result, the present work may soon need to be updated to reflect new paradigms and emerging tools.

This research opens several promising avenues for future work. First, exploring which factors primarily influence the BF results (e.g. language pair, user background, etc.) and attempting to quantify to which extent appears relevant. A deeper exploration of MT quality evaluation from the perspective of the PE effort is also needed. This type of research is particularly necessary to create new compensation models that are better aligned with the experiences of language professionals. Investigating the influence of the root cause of MT errors on perceived effort and quantifying subjectivity in PE tasks could yield valuable insights. Additional research could also explore the cognitive load associated with AI assistance in PE settings, including alternative translations as well as other potential use cases. For example, alternative

translations may result in an increased cognitive effort since users need to read more text, which could counterbalance its benefits, particularly in terms of productivity.

Applying the BF to other tools and contexts could further validate its utility and versatility. For example, technological advances are also transforming the field of media localisation, in particular with the emergence of new tools for PE of speech for AI dubbing. Such tools appear to be ideal candidates for further application of the BF. Finally, benchmarking should be seen as a continuous effort, with regular evaluations serving to assess whether the latest advancements genuinely benefit end users. In that regard, subsequent research could also build on the BF and adapt it to evaluate the tools that language professionals will be using in the future.

VI. BIBLIOGRAPHY

Adelson, J. L., and McCoach, D. B. (2010). Measuring the Mathematical Attitudes of Elementary Students: The Effects of a 4-point or 5-point Likert-Type Scale. *Educational and Psychological Measurement*, *70*(5), 796–807. https://doi.org/10.1177/0013164410366694

Ahrenberg, L. (2017). Comparing Machine Translation and Human Translation: A Case Study. In I. Temnikova, C. Orăsan, G. C. Pastor, and S. Vogel (Eds.), *Proceedings of the Workshop Human-Informed Translation and Interpreting Technology* (pp. 21–28). Association for Computational Linguistics, Shoumen, Bulgaria. https://doi.org/10.26615/978-954-452-042-7_003

Alonso, E., and Nunes Vieira, L. (2017). The Translator's Amanuensis 2020. *The Journal of Specialised Translation*, 345–361. https://doi.org/10.26034/cm.jostrans.2017.245

Alotaibi, H. M. (2020). Computer-Assisted Translation Tools: An Evaluation of Their Usability Among Arab Translators. Applied Sciences, 10(18), 6295. https://doi.org/10.3390/app10186295

ALPAC. (1966). Language and Machines: Computers in Translation and Linguistics. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, DC. National Academies Press. https://doi.org/10.17226/9547

Alva-Manchego, F., Specia, L., Szoc, S., Vanallemeersch, T., and Depraetere, H. (2021). Validating Quality Estimation in a Computer-Aided Translation Workflow: Speed, Cost and Quality Trade-off. In J. Campbell, B. Huyck, S. Larocca, J. Marciano, K. Savenkov, and A. Yanishevsky (Eds.), *Proceedings of Machine Translation Summit XVIII: Users and Providers Track* (pp. 306–315). Association for Machine Translation in the Americas. https://aclanthology.org/2021.mtsummit-up.22

Alves, F. (2003). Foreword: Triangulation in Process-Oriented Research in Translation. In F. Alves (Ed.), *Triangulating Translation: Perspectives in Process-Oriented Research* (pp. vii–x). John Benjamins Publishing Company. https://doi.org/10.1075/btl.45

Alves, F., and Gonçalves, J. L. (2013). Investigating the Conceptual-Procedural Distinction in the Translation Process: A Relevance-Theoretic Analysis of Micro and Macro Translation Units. *Target. International Journal of Translation Studies*, *25*(1), 107–124. https://doi.org/10.1075/target.25.1.09alv

Alves, F., Koglin, A., Mesa-Lao, B., Martínez, M. G., de Lima Fonseca, N. B., de Melo Sá, A., Gonçalves, J. L., Szpak, K. S., Sekino, K., and Aquino, M. (2016). Analysing the Impact of Interactive

Machine Translation on Post-editing Effort. In M. Carl, S. Bangalore, and M. Schaeffer (Eds.), *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB* (pp. 77–94). Springer. https://doi.org/10.1007/978-3-319-20358-4_4

Andersen, B., and Pettersen, P. G. (1995). *Benchmarking Handbook*. Chapman & Hall. https://link.springer.com/book/9780412735202

Arnold, D., Balkan, L., Meijer, S., Humphreys, R. L., and Sadler, L. (1994). *Machine Translation: An Introductory Guide*. https://aclanthology.org/www.mt-archive.info/Arnold-1994-0.pdf

Arthern, P. J. (1978). Machine Translation and Computerized Terminology Systems—A Translator's Viewpoint. *Proceedings of Translating and the Computer*. TC Conference, London, UK. https://aclanthology.org/1978.tc-1.5

ASTM. (2023). *ASTM F2575—Standard Practice for Language Translation*. ASTM International. https://doi.org/10.1520/F2575-23E01

Austermühl, F. (2001). *Electronic Tools for Translators*. St Jerome Publishing. https://doi.org/10.4324/9781315760353

Aziz, W., Castilho, S., and Specia, L. (2012). PET: a Tool for Post-Editing and Assessing Machine Translation. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA). https://aclanthology.org/L12-1587/

Baker, M. (1993). Corpus Linguistics and Translation Studies—Implications and Applications. In M. Baker, G. Francis, and E. Tognini-Bonelli (Eds.), *Text and Technology. In honour of John Sinclair* (pp. 233–252). Routledge. https://doi.org/10.1075/z.64.15bak

Bar-Hillel, Y. (1960). The Present Status of Automatic Translation of Languages. In F. L. Alt (Ed.), *Advances in Computers* (Vol. 1, pp. 91–163). Elsevier. https://doi.org/10.1016/S0065-2458(08)60607-5

Bassnett, S., and Lefevere, A. (1995). Translation, History, and Culture. Cassell. http://archive.org/details/translationhisto0000unse_q4b7

Béchara, H., Orăsan, C., Parra Escartín, C., Zampieri, M., and Lowe, W. (2021). The Role of Machine Translation Quality Estimation in the Post-Editing Workflow. *Informatics*, *8*(3), Article 3. https://doi.org/10.3390/informatics8030061

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural Versus Phrase-Based Machine Translation Quality. In J. Su, K. Duh, and X. Carreras (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 257–267). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/d16-1025

Bevan, N. (2009). What is the Difference Between the Purpose of Usability and User Experience Evaluation Methods. *Proceedings of UXEM'09 Workshop, INTERACT 2009*. https://www.semanticscholar.org/paper/What-is-the-difference-between-the-purpose-of-and-Bevan/cba74036995821ca560d31bf397c695a460a63a5

Bhutta, K. S., and Huq, F. (1999). Benchmarking – Best Practices: An Integrated Approach. *Benchmarking: An International Journal*, *6*(3), 254–268. https://doi.org/10.1108/14635779910289261

Bisbey, R., and Kay, M. (1972). *The MIND Translation System: A Study in Man-Machine Collaboration*. RAND Corporation. https://www.rand.org/pubs/papers/P4786.html

Bizzoni, Y., Juzek, T. S., España-Bonet, C., Dutta Chowdhury, K., van Genabith, J., and Teich, E. (2020). How Human is Machine Translationese? Comparing Human and Machine Translations of Text and Speech. In M. Federico, A. Waibel, K. Knight, S. Nakamura, H. Ney, J. Niehues, S. Stüker, D. Wu, J. Mariani, and F. Yvon (Eds.), *Proceedings of the 17th International Conference on Spoken Language Translation* (pp. 280–290). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.iwslt-1.34

Blagodarna, O. (2018). Insights into Post-Editors' Profiles and Post-Editing Practices. *Tradumàtica*, 35–51. https://doi.org/10.5565/rev/tradumatica.198

Blain, F., Senellart, J., Schwenk, H., Plitt, M., and Roturier, J. (2011). Qualitative Analysis of Post-Editing for High Quality Machine Translation. *Proceedings of Machine Translation Summit XIII: Papers*. MT Summit 2011, Xiamen, China. https://aclanthology.org/2011.mtsummit-papers.17

Boisen, E., Bygholm, A., and Hejlesen, O. K. (2002). Activity Theory and Medical Informatics: Usability, Utility, and Copability. In G. Surján, R. Engelbrecht, and P. McNair (Eds.), *Health Data in the Information Society* (pp. 826–831). IOS Press. https://doi.org/10.3233/978-1-60750-934-9-826

Bota, L., Schneider, C., and Way, A. (2013). COACH: Designing a New CAT Tool with Translator Interaction. In A. Way, D. Grasmick, and H. Depaetere (Eds.), *Proceedings of the Machine Translation Summit XIV: User Track* (pp. 313–220). https://aclanthology.org/2013.mtsummit-user.3

Bowker, L., and Fisher, D. (2010). Computer-Aided Translation. In Y. Gambier and L. Van Doorslaer (Eds.), *Handbook of Translation Studies* (pp. 60–65). John Benjamins. https://doi.org/10.1075/hts.1.comp2

Briva-Iglesias, V. (2024). Fostering Human-Centered, Augmented Machine Translation: Analysing Interactive Post-Editing [Ph.D. Thesis]. Dublin City University. https://doras.dcu.ie/30182/

Briva-Iglesias, V., Cavalheiro Camargo, J. L., and Dogru, G. (2024). Large Language Models "Ad Referendum": How Good Are They at Machine Translation in the Legal Domain? Preprint. https://arxiv.org/pdf/2402.07681

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

Briva-Iglesias, V., and O'Brien, S. (2022). Translators' Pre-Task Perceptions of MT and Their Impact on Final Translation Quality: Implications for Training. Presentation at Translating and the Computer (TC 44). https://asling.org/tc44/slides/TC44-Briva-Iglesias-Pre-task_preconceptions(v3).pdf

Briva-Iglesias, V., and O'Brien, S. (2023). Measuring Machine Translation User Experience (MTUX): A Comparison between AttrakDiff and User Experience Questionnaire. In M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, T. Ranasinghe, E. Vanmassenhove, S. Alvarez Vidal, N. Aranberri, M. Nunziatini, C. Parra Escartín, M. Forcada, M. Popovic, Scarton, and H. Moniz (Eds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation* (pp. 335–344). https://doi.org/10.1075/tcb.00077.bri

Briva-Iglesias, V., O'Brien, S., and Cowan, B. R. (2023). The Impact of Traditional and Interactive Post-Editing on Machine Translation User Experience, Quality, and Productivity. Translation, *Cognition & Behavior*, 6(1), 60–86. https://doi.org/10.1075/tcb.00077.bri

Brooke, J. (1996). SUS – A Quick and Dirty Usability Scale. In P. W. Jordan, B. Thomas, I. Lyall McClelland, and B. Weerdmeester (Eds.), *Usability Evaluation in Industry* (pp. 189–194). Taylor & Francis. https://doi.org/10.1201/9781498710411-35

Bundgaard, K., Paulsen Christensen, T., and Schjoldager, A. (2016). Translator-Computer Interaction in Action: An Observational Process Study of Computer-Aided Translation. *Journal of Specialised Translation*, *25*, 106–130. https://www.jostrans.org/article/view/7729

Camp, R. C. (1989). *Benchmarking: The Search for Industry Best Practices that Lead to Superior Performance*. ASQC Quality Press.

Candel-Mora, M. Á. (2015a). Comparable Corpus Approach to Explore the Influence of Computer-Assisted Translation Systems on Textuality. *Procedia - Social and Behavioral Sciences*, *198*, 67–73. https://doi.org/10.1016/j.sbspro.2015.07.420

Candel-Mora, M. Á. (2015b). Evaluation of English to Spanish MT Output of Tourism 2.0 Consumer-Generated Reviews with Post-Editing Purposes. *Proceedings of Translating and the Computer 37*. TC37 Conference, London, UK. https://aclanthology.org/2015.tc-1.7

Candel-Mora, M. Á. (2016). Translator Training and the Integration of Technology in the Translator's Workflow. In M. L. Carrió-Pastor (Ed.), *Technology Implementation in Second Language Teaching and Translation Studies: New Tools, New Approaches* (pp. 49–69). Springer. https://doi.org/10.1007/978-981-10-0572-5_4

Candel-Mora, M. Á. (2017). Criteria for the Integration of Term Banks in the Professional Translation Environment. *Sendebar: Revista de La Facultad de Traducción e Interpretación*, *28*, 243–260. https://revistaseug.ugr.es/index.php/sendebar/article/download/5353/5648/15926

Candel-Mora, M. Á. (2021). Futuro pasado de las tecnologías de la lengua para la traducción: Visiones, tendencias y realidad. In C. Vargas-Sierra and A. B. Martínez López (Eds.), *Investigación traductológica en la enseñanza y práctica profesional de la traducción y la interpretación* (pp. 37–48). Comares. https://dialnet.unirioja.es/servlet/articulo?codigo=8235100

Candel-Mora, M. Á. (2023). Transferencia de sentimientos en contenido generado por el usuario procesado con traducción automática. In C. Rico Pérez and M. del M. Sánchez Ramos (Eds.), *Traducción automática en contextos especializados, 2023, ISBN 9783631888841, págs. 71-91* (pp. 71–91). Peter Lang. https://dialnet.unirioja.es/servlet/articulo?codigo=9415005

Candel-Mora, M. Á., and Borja-Tormo, C. (2017). Desarrollo de la aplicación Post-editing Calculeffort para la estimación del esfuerzo en posedición. *Tradumàtica*, *15*, 1–9. https://doi.org/10.5565/rev/tradumatica.190

Candel-Mora, M. Á., and Ramírez Polo, L. (2015). Translation Technology in Institutional Settings: A Decision-Making Framework for the Implementation of Computer-Assisted Translation Systems. In P. Sánchez-Gijón, O. Torres-Hostench, and B. Mesa-Lao (Eds.), *Conducting Research in Translation Technologies* (pp. 71–92). Peter Lang. https://doi.org/10.3726/978-3-0353-0732-0

Carl, M., Bangalore, S., and Schaeffer, M. (2016). *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*. Springer. https://doi.org/10.1007/978-3-319-20358-4

Carl, M., Dragsted, B., and Jakobsen, A. L. (2011). On the Systematicity of Human Translation Processes. Session 2—Translation as a Profession. *Proceedings of Tralogy I. Métiers et Technologies de La Traduction : Quelles Convergences Pour l'Avenir ?* https://hal.science/hal-02495638

Caro Quintana, R. (2021). Integration of Machine Translation and Translation Memory: Post-Editing Efforts. In R. Mitkov, V. Sosoni, J. C. Giguère, E. Murgolo, and E. Deysel (Eds.), *Proceedings of the Translation and Interpreting Technology Online Conference* (pp. 161–166). https://doi.org/10.26615/978-954-452-071-7_018

Castano, A., and Casacuberta, F. (1997). A Connectionist Approach to Machine Translation. *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, 91–94. https://doi.org/10.21437/Eurospeech.1997-50

Castilho, S., Doherty, S., Gaspari, F., and Moorkens, J. (2018). Approaches to Human and Machine Translation Quality Assessment. In J. Moorkens, S. Castilho, F. Gaspari, and S. Doherty (Eds.), *Translation Quality Assessment: From Principles to Practice* (pp. 9–38). Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7_2

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., and Way, A. (2017). Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics*, *108*(1), 109–120. https://doi.org/10.1515/pralin-2017-0013

Castilho, S., O'Brien, S., Alves, F., and O'Brien, M. (2014). Does Post-Editing Increase Usability? A Study with Brazilian Portuguese as Target Language. In M. Cettolo, M. Federico, L. Specia, and A. Way (Eds.), *Proceedings of the 17th Annual Conference of the European Association for Machine Translation* (pp. 183–190). European Association for Machine Translation. https://aclanthology.org/2014.eamt-1.40/

Chereji, R., Wiesinger, C., Brockmann, J., Secară, A., and Ciobanu, D. (2022). *The Use of Speech Technologies in Translation, Revision, and Post-Editing Machine Translation*. https://asling.org/tc44/slides/TC44-Chereji+Brockmann-The_use_of_speech_technologies_UniVie(v5).pdf

Chu, C., and Wang, R. (2018). A Survey of Domain Adaptation for Neural Machine Translation. In E. M. Bender, L. Derczynski, and P. Isabelle (Eds.), *Proceedings of the 27th International Conference on*

Computational Linguistics (pp. 1304–1319). Association for Computational Linguistics. https://aclanthology.org/C18-1111

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Cooper, A. (2004). *Why High-Tech Products Drive Us Crazy and How to Restore the Sanity*. Sams Publishing.

Coppers, S., Van den Bergh, J., Luyten, K., Coninx, K., van der Lek, I., Vanallemeersch, T., and Vandeghinste, V. (2018). Intellingo: An Intelligible Translation Environment. In R. L. Mandryk, M. Hancock, M. Perry, and A. L. Cox (Eds.), *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). https://doi.org/10.1145/3173574.3174098

Corpas Pastor, G. (2006). Translation Quality Standards in Europe: An Overview. In E. Miyares Bermúdez and L. Ruiz Miyares (Eds.), *Linguistics in the Twenty First Century: Vol. Linguistics in the Twenty First Century* (pp. 47–57). Cambridge Scholars Publishing. https://www.cambridgescholars.com/product/9781904303862

Corpas Pastor, G., Mitkov, R., Afzal, N., and Pekar, V. (2008). Translation Universals: Do They Exist? A Corpus-Based NLP Study of Convergence and Simplification. *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas: Research Papers*, 75–81. https://aclanthology.org/2008.amta-papers.5

Costa, Â., Ling, W., Luís, T., Correia, R., and Coheur, L. (2015). A Linguistically Motivated Taxonomy for Machine Translation Error Analysis. *Machine Translation, 29(2)*, 127–161. https://doi.org/10.1007/s10590-015-9169-0

Cronin, M. (2003). *Translation and Globalization*. Routledge. https://www.routledge.com/Translation-and-Globalization/Cronin/p/book/9780415270656?srsltid=AfmBOoq82QvlW3lLdbFpwZQLNACerlEYwauYdXp_6RwNI3G6X6gXvltj

Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16, 297–334. https://doi.org/10.1007/BF02310555

CSA Research. (2024). *The Post-Localization Era*. https://www.linkedin.com/pulse/post-localization-era-csa-research-nofrc/

Daems, J. (2022). Dutch Literary Translators' Use and Perceived Usefulness of Technology: The Role of Awareness and Attitude. In J. L. Hadley, K. Taivalkoski-Shilov, C. S. C. Teixeira, and A. Toral (Eds.), *Using Technologies for Creative-Text Translation* (pp. 40–65). Routledge. https://doi.org/10.4324/9781003094159-3

Daems, J., and Macken, L. (2019). Interactive Adaptive SMT Versus Interactive Adaptive NMT: A User Experience Evaluation. *Machine Translation*, *33*(1), 117–134. https://doi.org/10.1007/s10590-019-09230-z

Daems, J., De Clercq, O., and Macken, L. (2017). Translationese and Post-Editese: How Comparable is Comparable Quality? *Linguistica Antverpiensia New Series-Themes in Translation Studies*, *16*, 89–103. https://doi.org/10.52034/lanstts.v16i0.434

Daems, J., Vandepitte, S., Hartsuiker, R. J., and Macken, L. (2017). Identifying the Machine Translation Error Types with the Greatest Impact on Post-editing Effort. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01282

Davis, F. D. (1987). *User Acceptance of Information Systems: The Technology Acceptance Model: TAM*. University of Michigan, School of Business Administration.

de Rosnay, J. (2016). *Je cherche à comprendre… Les codes cachés de la nature*. Les Liens Qui Libèrent. https://www.editionslesliensquiliberent.fr/livre-Je_cherche_%C3%A0_comprendre-9791020903952-1-1-0-1.html

de Saussure, F. (1971). *Cours de linguistique générale*. Études et documents Payot.

Denkowski, M., and Lavie, A. (2012). Challenges in Predicting Machine Translation Utility for Human Post-Editors. *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers*. AMTA 2012, San Diego, California, USA. https://aclanthology.org/2012.amta-papers.6

DePalma, D. A. (2017). *Augmented Translation Powers up Language Services*. CSA Research. https://csa-research.com/

Dinu, G., Mathur, P., Federico, M., and Al-Onaizan, Y. (2019). Training Neural Machine Translation to Apply Terminology Constraints. In A. Korhonen, D. Traum, and L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3063–3068). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1294

Djabri, S., and Caro Quintana, R. (2021). Introducing Linguistic Transformation to Improve Translation Memory Retrieval. Results of a Professional Translators' Survey for Spanish, French and Arabic. In S. Djabri, D. Gimadi, T. Mihaylova, and I. Nikolova-Koleva (Eds.), *Proceedings of the Student Research Workshop Associated with RANLP 2021* (pp. 44–50). https://doi.org/10.26615/issn.2603-2821.2021_007

do Carmo, F. E. M. (2017). *Post-Editing: A Theoretical and Practical Challenge for Translation Studies and Machine Learning* [Ph.D. Thesis, Universidade do Porto]. https://repositorio-aberto.up.pt/handle/10216/107518

do Carmo, F., Shterionov, D., Moorkens, J., Wagner, J., Hossari, M., Paquin, E., Schmidtke, D., Groves, D., and Way, A. (2021). A Review of the State-of-the-Art in Automatic Post-Editing. *Machine Translation*, *35*(2), 101–143. https://doi.org/10.1007/s10590-020-09252-y

do Carmo, F., Trigo, L., and Maia, B. (2016). From CATs to KATs. *Proceedings of Translating and the Computer 38*. TC38 Conference, London, UK. https://aclanthology.org/2016.tc-1.15

Doherty, S., and O'Brien, S. (2009). Can MT Output Be Evaluated Through Eye Tracking? *Proceedings of Machine Translation Summit XII: Posters*. MT Summit, Ottawa, Canada. https://aclanthology.org/2009.mtsummit-posters.5

Doherty, S., and O'Brien, S. (2014). Assessing the Usability of Raw Machine Translated Output: A User-Centered Study Using Eye Tracking. *International Journal of Human-Computer Interaction*. https://www.tandfonline.com/doi/abs/10.1080/10447318.2013.802199

Domingo, M., García-Martínez, M., Peris, Á., Helle, A., Estela, A., Bié, L., Casacuberta, F., and Herranz, M. (2020). A User Study of the Incremental Learning in NMT. In A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, and M. L. Forcada (Eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 319–328). European Association for Machine Translation. https://aclanthology.org/2020.eamt-1.34

Dragsted, B. (2006). Computer-Aided Translation as a Distributed Cognitive Task. In I. E. Dror and S. Harnad (Eds.), *Cognition Distributed. How Cognitive Technology Extends our Minds* (Vol. 14, pp. 443–464). John Benjamins. https://doi.org/10.1075/bct.16.16dra

EAGLES Evaluation Working Group. (1996). *EAGLES Evaluation of Natural Language Processing Systems: Final Report*. Center for Sprogteknologi, Denmark.

EAGLES Evaluation Working Group. (1999). *The EAGLES 7-Step Recipe*. https://www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html

Ehrensberger-Dow, M., and O'Brien, S. (2015). Ergonomics of the Translation Workplace: Potential for Cognitive Friction. *Translation Spaces*, *4*(1), 98–118. https://doi.org/10.1075/ts.4.1.05ehr

EMT Board. (2017). *European Master's in Translation Competence Framework*. https://commission.europa.eu/system/files/2018-02/emt_competence_fwk_2017_en_web.pdf

Engelbart, D. C. (1962). Augmenting Human Intellect: A Conceptual Framework. *Summary Report AFOSR-3223 under Contract AF 49(638)-1024, SRI Project 3578 for Air Force Office of Scientific Research*.

Englund Dimitrova, B. (2010). Translation Process. In Y. Gambier and L. van Doorslaer (Eds.), *Handbook of Translation Studies* (Vol. 1, pp. 406–411). John Benjamins. https://doi.org/10.1075/hts.5

Escribe, M., and Candel-Mora, M. Á. (2024). Optimising Translation Tools for Post-Editing: Results of a User Survey. In C. Orăsan, T. Ranasinghe, G. Corpas Pastor, R. Mitkov, M. Kunilovskaya, V. Sosoni, and M. Escribe (Eds.), *Proceedings of the International Conference on New Trends in Translation and Technology Conference 2024* (pp. 63–75). Incoma Ltd. Shoumen, Bulgaria. https://doi.org/10.26615/issn.2815-4711.2024_006

Escribe, M., and Mitkov, R. (2023). Applying Incremental Learning to Post-Editing Systems: Towards Online Adaptation for Automatic Post-Editing Models. In Pan, J., Laviosa, S. (Eds.), *Corpora and Translation Education: Advances and Challenges* (pp. 35-62). Singapore: Springer. https://doi.org/10.1007/978-981-99-6589-2_3

Farrell, M. (2023). Current Evidence of Post-Editese: Differences Between Post-Edited Neural Machine Translation Output and Human Translation Revealed Through Human Evaluation. In C. Orăsan, R. Mitkov, G. Corpas Pastor, and J. Monti (Eds.), *Proceedings of the Human-Informed Translation and Interpreting Technology International Conference* (pp. 52–63). https://doi.org/10.26615/issn.2683-0078.2023_005

Fırat, G. (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

Fomicheva, M., Sun, S., Fonseca, E., Zerva, C., Blain, F., Chaudhary, V., Guzmán, F., Lopatina, N., Specia, L., and Martins, A. F. T. (2022). MLQE-PE: A Multilingual Quality Estimation and Post-Editing Dataset. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara,

B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 4963–4974). European Language Resources Association. https://aclanthology.org/2022.lrec-1.530

Font Llitjós, A., Carbonell, J. G., and Lavie, A. (2005). A Framework for Interactive and Automatic Refinement of Transfer-Based Machine Translation. *Proceedings of the 10th EAMT Conference: Practical Applications of Machine Translation*, 87–96. https://aclanthology.org/2005.eamt-1.13

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

Foster, G., Langlais, P., and Lapalme, G. (2002). User-Friendly Text Prediction for Translators. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, 148–155. https://doi.org/10.3115/1118693.1118713

GALA. (2022). *The MTPE Training Protocol—Interview with Viveta Gene (Intertranslation) and Lucía Guerrero (CPSL)*. GALA Global. https://www.gala-global.org/knowledge-center/professional-development/articles/mtpe-training-protocol-interview-viveta-gene

García Castro, R. (2008). *Benchmarking Semantic Web Technology* [Ph.D. Thesis, Universidad Politécnica de Madrid]. https://oa.upm.es/2234/

Garcia, I. (2014). Computer-Aided Translation: Systems. In C. Sin-wai (Ed.), *Routledge Encyclopedia of Translation Technology* (pp. 68–87). Routledge. https://doi.org/10.4324/9781315749129

García-Martínez, M., Singla, K., Tammewar, A., Mesa-Lao, B., Thakur, A., Carl, M., and Bangalore, S. (2014). SEECAT: ASR & Eye-tracking Enabled Computer-Assisted Translation. In M. Cettolo, M. Federico, L. Specia, and A. Way (Eds.), *Proceedings of the European Association for Machine Translation Conference* (pp. 81–88). https://aclanthology.org/2014.eamt-1

Gene, V., and Guerrero, L. (2021). *A Report on the TC43 Workshop: Drafting Effective Machine Translation Post-Editing Guidelines*. Translating and the Computer 43. https://www.tradulex.com/varia/TC43-OnTheWeb2021.pdf#page=43

Ginovart Cid, C. (2020). The Professional Profile of a Post-Editor According to LSCs and Linguists: A Survey-Based Research. *HERMES: Journal of Language and Communication in Business*, *60*, 171–190. https://doi.org/10.7146/hjlcb.v60i0.121318

Ginovart Cid, C. (2021). The Need for Practice in the Acquisition of the Post-Editing Skill-Set: Lessons Learned from the Industry [Ph.D. Thesis, Universitat Pompeu Fabra]. In *TDX (Tesis Doctorals en Xarxa)*. https://repositori.upf.edu/handle/10230/48823

Göpferich, S. (2009). Towards a Model of Translation Competence and its Acquisition: The Longitudinal Study TransComp. In S. Göpferich, A. L. Jakobsen, and I. M. Mees, *Behind the Mind: Methods, Models and Results in Translation Process Research* (pp. 11–37). Samfundslitteratur.

Göpferich, S., and Jääskeläinen, R. (2009). Process Research into the Development of Translation Competence: Where Are We, and Where Do We Need to Go? *Across Languages and Cultures*, 10(2), 169–191. https://doi.org/10.1556/Acr.10.2009.2.1

Görög, A. (2014). Quantifying and Benchmarking Quality: The TAUS Dynamic Quality Framework. *Tradumàtica*, *12*, 443–454. https://doi.org/10.5565/rev/tradumatica.66

Gould, J. D., and Lewis, C. (1985). Designing for Usability: Key Principles and What Designers Think. *Communications of the ACM*, *28*(3), 300–311. https://doi.org/10.1145/3166.3170

Green, S., Heer, J., and Manning, C. D. (2015). Natural Language Translation at the Intersection of AI and HCI. *Communications of the ACM*, *58*(9), 46–53. https://doi.org/10.1145/2767151

Guerberof Arenas, A. (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. https://doras.dcu.ie/23672/

Guo, F. (2012). *More Than Usability: The Four Elements of User Experience, Part I*. https://www.uxmatters.com/mt/archives/2012/04/more-than-usability-the-four-elements-of-user-experience-part-i.php

Gutt, E.-A. (2014). *Translation and Relevance: Cognition and Context*. Routledge. https://doi.org/10.4324/9781315760018

Halliday, M. A. K., and Hasan, R. (1985). *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective*. OUP/ Deakin University Press.

Hamm, J., and Klein, J. (2023). How STAR Transit NXT Can Help Translators Measure and Increase their MT Post-Editing Efficiency. In M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, T. Ranasinghe, E. Vanmassenhove, S. A. Vidal, N. Aranberri, M. Nunziatini, C. P. Escartín, M. Forcada, M. Popovic, C. Scarton, and H. Moniz (Eds.), *Proceedings of the 24th Annual Conference*

of the European Association for Machine Translation (pp. 513–514). European Association for Machine Translation. https://aclanthology.org/2023.eamt-1.59

Harris, B. (1987). Bi-Text, a New Concept in Translation Theory. *Language Monthly*, *54*, 8–10.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang D., Zhang Z., and Zhou, M. (2018). *Achieving Human Parity on Automatic Chinese to English News Translation* (arXiv:1803.05567). arXiv. https://doi.org/10.48550/arXiv.1803.05567

Hassenzahl, M., Burmester, M., and Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In G. Szwillus and J. Ziegler (Eds.), *Mensch & Computer 2003: Interaktion in Bewegung* (pp. 187–196). Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-322-80058-9_19

Herbig, N., Düwel, T., Pal, S., Meladaki, K., Monshizadeh, M., Krüger, A., and van Genabith, J. (2020). MMPE: A Multi-Modal Interface for Post-Editing Machine Translation. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1691–1702). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.155

Herbig, N., Pal, S., van Genabith, J., and Krüger, A. (2019). *Integrating Artificial and Human Intelligence for Efficient Translation.* (arXiv:1903.02978). arXiv. https://arxiv.org/pdf/1903.02978

Heyn, M. (1996, August 29). Integrating Machine Translation into Translation Memory Systems. *Proceedings of EAMT Machine Translation Workshop*. EAMT 1996, Vienna, Austria. https://aclanthology.org/1996.eamt-1.12

Holmes, J. S. (1988). *Translated! Papers on Literary Translation and Translation Studies*. Rodopi. https://brill.com/display/title/27760

House, J. (2014). *Translation Quality Assessment: Past and Present*. Routledge. https://doi.org/10.4324/9781315752839

Hovy, E., King, M., and Popescu-Belis, A. (2002). Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, *17*, 43–75. https://doi.org/10.1023/A:1025510524115

Hu, K., and Cadwell, P. (2016). A Comparative Study of Post-editing Guidelines. *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, 34206–34353. https://aclanthology.org/W16-3420

Hutchins, J. (1998). The Origins of the Translator's Workstation. *Machine Translation*, *13*(4), 287–307. https://doi.org/10.1023/A:1008123410206

Hutchins, J. (2005). *The First Public Demonstration of Machine Translation: The Georgetown-IBM System, 7th January 1954*.

Hutchins, J. (2014). *The History of Machine Translation in a Nutshell*. https://aclanthology.org/www.mt-archive.info/10/Hutchins-2014.pdf

Hutchins, J., and Lovtskii, E. (2000). Petr Petrovich Troyanskii (1894–1950): A Forgotten Pioneer of Mechanical Translation. *Machine Translation*, *15*(3), 187–221. https://doi.org/10.1023/A:1011653602669

Hutchins, W. J., and Somers, H. L. (1992). *An Introduction to Machine Translation*. Academic Press.

ISO. (2015a). *ISO 17100:2015 Translation Services—Requirements for Translation Services*. ISO. https://www.iso.org/standard/59149.html

ISO. (2015b). *ISO 9000:2015 Quality Management Systems — Fundamentals and Vocabulary*. ISO. https://www.iso.org/standard/45481.html

ISO. (2017). *ISO 18587:2017 Translation Services—Post-Editing of Machine Translation Output—Requirements*. ISO. https://www.iso.org/standard/62970.html

ISO. (2018). *ISO 9241-11:2018 Ergonomics of Human-System Interaction—Part 11: Usability: Definitions and Concepts*. https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

ISO. (2023). *ISO 639:2023 Code for Individual Languages and Language Groups*. ISO. https://www.iso.org/standard/74575.html

ISO. (2024). *ISO 11669:2024 Translation Projects—General Guidance*. https://www.iso.org/standard/79089.html

ISO. (2024). *ISO 5060:2024 Translation Services—Evaluation of Translation Output—General Guidance*. ISO. https://www.iso.org/standard/80701.html

ISO/IEC. (1991). *ISO/IEC 9126:1991 Software Engineering—Product Quality*. ISO. https://www.iso.org/standard/16722.html

ISO/IEC. (2001). *ISO/IEC 14598-6:2001 Software Engineering—Product Evaluation—Part 6: Documentation of Evaluation Modules*. https://www.iso.org/obp/ui/#iso:std:iso-iec:14598:-6:ed-1:v1:en

ISO/IEC. (2011). *ISO/IEC 25010:2011 Systems and Software Engineering—Systems and Software Quality Requirements and Evaluation (SQuaRE)—System and Software Quality Models.* ISO. https://www.iso.org/standard/35733.html

ISO/IEC. (2023). *ISO/IEC 25010:2023—Systems and Software Engineering—Systems and Software Quality Requirements and Evaluation (SQuaRE)—Product Quality Model*. https://www.iso.org/standard/78176.html

ISO/IEC. (2023). *ISO/IEC 25019:2023 Systems and Software Engineering—Systems and Software Quality Requirements and Evaluation (SQuaRE)—Quality-In-Use Model*. ISO. https://www.iso.org/standard/78177.html

ISO/IEC. (2024). *ISO/IEC 25002:2024 Systems and Software Engineering—Systems and Software Quality Requirements and Evaluation (SQuaRE)—Quality Model Overview and Usage*. ISO. https://www.iso.org/standard/78175.html

Ive, J., Specia, L., Szoc, S., Vanallemeersch, T., Van den Bogaert, J., Farah, E., Maroti, C., Ventura, A., and Khalilov, M. (2020). A Post-Editing Dataset in the Legal Domain: Do We Underestimate Neural Machine Translation Quality? In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3692–3697). European Language Resources Association. https://aclanthology.org/2020.lrec-1.455

Jia, Y., Carl, M., and Wang, X. (2019). How Does the Post-Editing of Neural Machine Translation Compare with From-Scratch Translation? A Product and Process Study. *Journal of Specialised Translation*, *31*, 60–86. https://doi.org/10.26034/cm.jostrans.2019.177

Johnson, W. (1944). A Program of Research. *Psychological Monographs*, 56(2), 1–15. https://doi.org/10.1037/h0093508

Kanavos, P., and Kartsaklis, D. (2010). Integrating Machine Translation with Translation Memory: A Practical Approach. In V. Zhechev (Ed.), *Proceedings of the Second Joint EM+/CNGL Workshop:*

Bringing MT to the User: Research on Integrating MT in the Translation Industry (pp. 11–20). https://aclanthology.org/2010.jec-1

Karakanta, A., Bentivogli, L., Cettolo, M., Negri, M., and Turchi, M. (2022). Post-Editing in Automatic Subtitling: A Subtitlers' Perspective. In H. Moniz, L. Macken, A. Rufener, L. Barrault, M. R. Costa-jussà, C. Declercq, M. Koponen, E. Kemp, S. Pilos, M. L. Forcada, C. Scarton, J. Van den Bogaert, J. Daems, A. Tezcan, B. Vanroy, and M. Fonteyne (Eds.), *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation* (pp. 261–270). European Association for Machine Translation. https://aclanthology.org/2022.eamt-1.29/

Kappus, M., and Ehrensberger-Dow, M. (2020). The Ergonomics of Translation Tools: Understanding when Less is Actually More. *The Interpreter and Translator Trainer*, *14*(4), 386–404. https://doi.org/10.1080/1750399x.2020.1839998

Karimova, S., Simianer, P., and Riezler, S. (2018). A User-Study on Online Adaptation of Neural Machine Translation to Human Post-Edits. *Machine Translation*, *32*(4), 309–324. https://doi.org/10.1007/s10590-018-9224-8

Kay, M. (1980). *The Proper Place of Men and Machines in Language Translation. Xerox PARC CSL-80-11 Working Paper.* https://aclanthology.org/www.mt-archive.info/70/Kay-1980.pdf

King, M., and Maegaard, B. (1998). Issues in Natural Language Systems Evaluation. *Proceedings of the First International Conference on Language Resources & Evaluation*. First International Conference on Language Resources & Evaluation, Granada, Spain. https://aclanthology.org/www.mt-archive.info/LREC-1998-King.pdf

Koehn, P. (2009). A Process Study of Computer-Aided Translation. *Machine Translation*, *23*(4), 241–263. https://doi.org/10.1007/s10590-010-9076-3

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815829

Koehn, P., and Knowles, R. (2017). Six Challenges for Neural Machine Translation. In T. Luong, A. Birch, G. Neubig, and A. Finch (Eds.), *Proceedings of the First Workshop on Neural Machine Translation* (pp. 28–39). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3204

Kolko, J. (2010). *Thoughts on Interaction Design*. Morgan Kaufmann. https://doi.org/10.1016/c2009-0-61347-7

Koponen, M. (2012). Comparing Human Perceptions of Post-Editing Effort with Post-Editing Operations. In C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia (Eds.), *Proceedings of the Seventh Workshop on Statistical Machine Translation* (pp. 181–190). Association for Computational Linguistics. https://aclanthology.org/W12-3123

Koponen, M., Sulubacak, U., Vitikainen, K., and Tiedemann, J. (2020). MT for Subtitling: Workshop on Post-Editing in Modern-Day Translation. *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas October 6 - 9, 2020*, 79–92. https://aclanthology.org/2020.amta-pemdt.6/

Krings, H. P. (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Kent State University Press. https://doi.org/10.1075/target.15.2.15jak

Krollmann, F. (1971). Linguistic Data Banks and the Technical Translator. *Meta : Journal Des Traducteurs / Meta: Translators' Journal*, *16*(1–2), 117–124. https://doi.org/10.7202/003352ar

Krüger, R. (2016). Contextualising Computer-Assisted Translation Tools and Modelling Their Usability. *Trans-Kom-Journal of Translation and Technical Communication Research*, *9*(1), 114–148. https://www.trans-kom.eu/bd09nr01/trans-kom_09_01_08_Krueger_CAT.20160705.pdf

Krüger, R. (2019). A Model for Measuring the Usability of Computer-Assisted Translation Tools. In H. E. Jüngst, L. Link, K. Schubert, and C. Zehrer (Eds.), *Challenging Boundaries – New Approaches to Specialized Communication* (pp. 93–118). https://www.frank-timme.de/en/programme/product/challenging_boundaries

Lagoudaki, E. (2006). Translation Memories Survey 2006: User's Perceptions Around TM Usage. *Proceedings of Translating and the Computer 28*. TC28 Conference, London, UK. https://aclanthology.org/2006.tc-1.2

Lagoudaki, P. M. (2008). *Expanding the Possibilities of Translation Memory Systems: From the Translators Wishlist to the Developers Design* [Ph.D. Thesis, Imperial College London]. http://spiral.imperial.ac.uk/handle/10044/1/7879

Landwehr, F., Steinmann, T., and Mascarell, L. (2023). OpenTIPE: An Open-Source Translation Framework for Interactive Post-Editing Research. In D. Bollegala, R. Huang, and A. Ritter (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)* (pp. 208–216). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-demo.19

Language Monthly. (1986). New ALPS system running on IBM PC AT. *Language Monthly*, 10.

Läubli, S., Amrhein, C., Düggelin, P., Gonzalez, B., Zwahlen, A., and Volk, M. (2019). Post-Editing Productivity with Neural Machine Translation: An Empirical Assessment of Speed and Quality in the Banking and Finance Domain. In M. Forcada, A. Way, B. Haddow, and R. Sennrich (Eds.), *Proceedings of Machine Translation Summit XVII: Research Track* (pp. 267–272). European Association for Machine Translation. https://aclanthology.org/W19-6626/

Läubli, S., Sennrich, R., and Volk, M. (2018). Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4791–4796). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1512

Laugwitz, B., Held, T., and Schrepp, M. (2008). Construction and Evaluation of a User Experience Questionnaire. In A. Holzinger (Ed.), *HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society* (pp. 63–76). Springer. https://doi.org/10.1007/978-3-540-89350-9_6

Lavie, A., and Agarwal, A. (2007). METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In C. Callison-Burch, P. Koehn, C. S. Fordyce, and C. Monz (Eds.), *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 228–231). Association for Computational Linguistics. https://aclanthology.org/W07-0734

Lee, D., Ahn, J., Park, H., and Jo, J. (2021). IntelliCAT: Intelligent Machine Translation Post-Editing with Quality Estimation and Translation Suggestion. In H. Ji, J. C. Park, and R. Xia (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations* (pp. 11–19).

Lee, K., Firat, O., Agarwal, A., Fannjiang, C., and Sussillo, D. (2018). Hallucinations in Neural Machine Translation. *Proceedings of the Interpretability and Robustness in Audio, Speech, and Language Workshop, Conference on Neural Information Processing Systems*. NeurIPS 2018, Montreal, Canada. https://clarafy.github.io/research/neurips_irasl_2018.pdf

Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., Koo, S., and Lim, H. (2023). A Survey on Evaluation Metrics for Machine Translation. *Mathematics*, *11*(4), 1006. https://doi.org/10.3390/math11041006

Leiter, C., Lertvittayakumjorn, P., Fomicheva, M., Zhao, W., Gao, Y., and Eger, S. (2024). Towards Explainable Evaluation Metrics for Machine Translation. *Journal of Machine Learning Research*. https://www.jmlr.org/papers/volume25/22-0416/22-0416.pdf

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 707–710.

Levý, J. (1967). Translation as a Decision Process. *To Honor Roman Jakobson: Essays on the Occasion of His Seventieth Birthday* (Vol. 2, pp. 1171–1182). De Gruyter Mouton. https://doi.org/10.1515/9783111349121-031

Licklider, J. C. R. (1960). Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics*, *HFE-1*(1), 4–11. https://doi.org/10.1109/THFE2.1960.4503259

Lippmann, E. O. (1971). An Approach to Computer-Aided Translation. *IEEE Transactions on Engineering Writing and Speech*, *14*(1), 10–33. IEEE Transactions on Engineering Writing and Speech. https://doi.org/10.1109/TEWS.1971.4322455

Liu, J., Wong, C. K., and Hui, K. K. (2003). An Adaptive User Interface Based on Personalized Learning. *IEEE Intelligent Systems*, *18*(2), 52–57. https://doi.org/10.1109/mis.2003.1193657

Lo, C., Knowles, R., and Goutte, C. (2023). Beyond Correlation: Making Sense of the Score Differences of New MT Evaluation Metrics. In M. Utiyama and R. Wang (Eds.), *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track* (pp. 186–199). Asia-Pacific Association for Machine Translation. https://aclanthology.org/2023.mtsummit-research.16/

Lommel, A., Görög, A., Melby, A., Uszkoreit, H., Burchardt, A., and Popović, M. (2015). *QT21 Deliverable 3.1. Harmonised Metric.* https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5a0f575d3&appId=PPGMS

Lommel, A., Uszkoreit, H., and Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, 455–463. https://doi.org/10.5565/rev/tradumatica.77

Maheshwari, A., Ravindran, A., Subramanian, V., and Ramakrishnan, G. (2023). UDAAN - Machine Learning based Post-Editing Tool for Document Translation. *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)*, 263–267. https://doi.org/10.1145/3570991.3571068

Makoushina, J. (2007, November 29). Translation Quality Assurance Tools: Current State and Future Approaches. *Proceedings of Translating and the Computer 29*. TC29 Conference, London, UK. https://aclanthology.org/2007.tc-1.8

Marick, B. (2008). A Manglish Way of Working: Agile Software Development. In A. Pickering, K. Guzik, B. Herrnstein Smith, and E. R. Weintraub (Eds.), *The Mangle in Practice*. Duke University Press. https://doi.org/10.1215/9780822390107-008

Massardo, I., van der Meer, J., O'Brien, S., Hollowood, F., Aranberri, N., and Drescher, K. (2016). *MT Post-Editing Guidelines*. https://info.taus.net/mt-post-editing-guidelines

Maxwell, K. D., and Forselius, P. (2000). Benchmarking Software Development Productivity. *IEEE Software*, *17*(1), 80–88. IEEE Software.

McCarthy, P. M., and Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment. *Behavior Research Methods*, 42(2), 381–392. https://doi.org/10.3758/brm.42.2.381

Měchura, M. (2022). A Taxonomy of Bias-Causing Ambiguities in Machine Translation. In C. Hardmeier, C. Basta, M. R. Costa-Jussà, G. Stanovsky, and H. Gonen (Eds.), *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)* (pp. 168–173). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.gebnlp-1.18

Melby, A. K. (1982). Multi-Level Translation Aids in a Distributed System. *Proceedings of the Ninth International Conference on Computational Linguistics*. COLING 1982. https://aclanthology.org/C82-1034

Melby, A. K. (1992). The Translator Workstation. In J. Newton (Ed.), *Computers in Translation: A Practical Appraisal* (pp. 147–165). Routledge. https://doi.org/10.4324/9780203128978

Melby, A., and Kurz, C. (2021). *Data: Of Course! MT: Useful or Risky. Translators: Here to Stay! | MultiLingual Nov/Dec 2021*. https://multilingual.com/issues/november-december-2021/data-of-course-mt-useful-or-risky-translators-here-to-stay/

Miłkowski, M. (2012). Translation Quality Checking in LanguageTool. In P. Pezik (Ed.), *Corpus Data across Languages and Disciplines* (pp. 213–223). Peter Lang. https://doi.org/10.3726/978-3-653-02465-4

Moorkens, J. (2012). *Measuring Consistency in Translation Memories: A Mixed-Methods Case Study* [Ph.D. Thesis, Dublin City University]. https://core.ac.uk/reader/147602800

Moorkens, J. (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

Moorkens, J., and O'Brien, S. (2013). User Attitudes to the Post-Editing Interface. In S. O'Brien, M. Simard, and L. Specia (Eds.), *Proceedings of the 2nd Workshop on Post-editing Technology and Practice*. https://aclanthology.org/2013.mtsummit-wptp.3

Moorkens, J., and O'Brien, S. (2017). Assessing User Interface Needs of Post-Editors of Machine Translation. In D. Kenny (Ed.), *Human Issues in Translation Technology* (pp. 109–130).

Moorkens, J., O'Brien, S., da Silva, I. A. L., de Lima Fonseca, N. B., and Alves, F. (2015). Correlations of Perceived Post-Editing Effort with Measurements of Actual Effort. *Machine Translation*, *29*(3), 267–284. https://doi.org/10.1007/s10590-015-9175-2

Mossop, B. (2006). Has Computerization Changed Translation? *Meta : Journal Des Traducteurs / Meta: Translators' Journal*, *51*(4), 787–805. https://doi.org/10.7202/014342ar

Mossop, B. (2020). *Revising and Editing for Translators* (4th ed.). Routledge. https://doi.org/10.4324/9781315158990

Müller, H., and Sedley, A. (2014). HaTS: Large-Scale In-Product Measurement of User Attitudes & Experiences with Happiness Tracking Surveys. *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design*, 308–315. https://doi.org/10.1145/2686612.2686656

Mutal, J., Volkart, L., Bouillon, P., Girletti, S., and Estrella, P. (2019). Differences between SMT and NMT Output—A Translators' Point of View. In I. Temnikova, C. Orăsan, G. Corpas Pastor, and R. Mitkov (Eds.), *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)* (pp. 75–81). Incoma Ltd., Shoumen, Bulgaria. https://doi.org/10.26615/issn.2683-0078.2019_009

Ñeco, R. P., and Forcada, M. L. (1997). Asynchronous Translations with Recurrent Neural Nets. *Proceedings of the International Conference on Neural Networks (ICNN'97)*, *4*, 2535–2540 vol. 4. https://doi.org/10.1109/ICNN.1997.614693

Newmark, P. (2003). *A Textbook of Translation* (3rd ed.). Longman.

Nida, E. A., and Taber, C. R. (1969). *The Theory and Practice of Translation* (Vol. 8). Brill Archive. https://brill.com/display/title/341

Nielsen, J. (1994). *Usability Engineering*. Morgan Kaufmann. https://doi.org/10.1016/c2009-0-21512-1

Niessen, S., Och, F. J., Leusch, G., and Ney, H. (2000). An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhauer (Eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)* (pp. 39–45). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2000/pdf/278.pdf

Nimdzi. (2023). *Nimdzi Language Technology Atlas 2023*. https://www.nimdzi.com/language-technology-atlas/

Nimdzi. (2024). *The Nimdzi 100*. https://www.nimdzi.com/nimdzi-100/04-market/

Nitzke, J., and Hansen-Schirra, S. (2021). *A Short Guide to Post-Editing* (Vol. 16). Language Science Press. https://doi.org/10.5281/zenodo.5646896

Nunes Vieira, L. (2019). Post-Editing of Machine Translation. In M. O'Hagan (Ed.), *The Routledge Handbook of Translation and Technology* (pp. 319–335). https://doi.org/10.4324/9781315311258

Nunes Vieira, L., and Specia, L. (2011). A Review of Translation Tools from a Post-Editing Perspective. In Zhechev (Ed.), *Proceedings of the Third Joint EM+/CNGL Workshop Bringing MT to the User: Research Meets Translators (JEC 2011)* (pp. 33–42). https://aclanthology.org/www.mt-archive.info/JEC-2011-Vieira.pdf

Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.). McGraw-Hill.

O'Brien, S. (2002). Teaching Post-Editing: A Proposal for Course Content. *Proceedings of the 6th EAMT Workshop: Teaching Machine Translation*. EAMT 2002, Manchester, England. https://aclanthology.org/2002.eamt-1.11

O'Brien, S. (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, S. (2012). Translation as Human–Computer Interaction. *Translation Spaces*, *1*, 101–122. https://doi.org/10.1075/ts.1.05obr

O'Brien, S. (2020). Translation, Human–Computer Interaction and Cognition. In F. Alves and A. L. Jakobsen (Eds.), *The Routledge Handbook of Translation and Cognition* (pp. 376–388). Routledge. https://doi.org/10.4324/9781315178127-25

O'Brien, S., Ehrensberger-Dow, M., Hasler, M., and Connolly, M. (2017). Irritating CAT Tool Features that Matter to Translators. *Hermes: Journal of Language and Communication in Business*, *56*, 145–162. https://doi.org/10.7146/hjlcb.v0i56.97229

O'Brien, S., Moorkens, J., and Vreeke, J. (2014). Kanjingo – A Mobile App for Post-Editing. In M. Cettolo, M. Federico, L. Specia, and Way (Eds.), *Proceedings of the 17th Annual Conference of the European Association for Machine Translation* (pp. 137–141). https://aclanthology.org/2014.eamt-1.33

O'Brien, S., O'Hagan, M., and Flanagan, M. (2010). Keeping an Eye on the UI Design of Translation Memory: How Do Translators Use the 'Concordance' Feature? *Proceedings of the 28th Annual European Conference on Cognitive Ergonomics*, 187–190. https://dl.acm.org/doi/pdf/10.1145/1962300.1962338?casa_token=T8XlhG3B0tIAAAAA:IkrYsSg3pR55xYJawwdwzpwqbPNGW0YuAhByt4BtMFO3NuWe8UCOKIxfAX6wqJdVA56Lk6Ek8Za0

Olohan, M. (2011). Translators and Translation Technology: The Dance of Agency. *Translation Studies*, *4*(3), 342–357. https://doi.org/10.1080/14781700.2011.589656

Ou, O., and Kleiner, B. (2015). Excellence in Benchmarking. *Industrial Management*, *57*(6), 20–24.

Ovchinnikova, I. (2020). Impact of New Technologies on the Types of Translation Errors. *Proceedings of the Linguistic Forum 2020: Language and Artificial Intelligence*. https://ceur-ws.org/Vol-2852/paper8.pdf

P., Leiva, L., Mesa-Lao, B., Ortiz, D., Saint-Amand, H., Sanchis, G., and Tsoukala, C. (2013). CASMACAT: An Open Source Workbench for Advanced Computer Aided Translation. *The Prague Bulletin of Mathematical Linguistics*, *100*(1), 101–112. https://doi.org/10.2478/pralin-2013-0016

PACTE Group. (2005). Investigating Translation Competence: Conceptual and Methodological Issues. *Meta : Journal Des Traducteurs / Meta: Translators' Journal*, *50*(2), 609–619. https://doi.org/10.7202/011004ar

Paggio, P., and Underwood, N. L. (1998). Validating the TEMAA LE Evaluation Methodology: A Case Study on Danish Spelling Checkers. *Natural Language Engineering*, *4*(3), 211–228. https://doi.org/10.1017/S1351324998001995

Pal, S., Naskar, S. K., Zampieri, M., Nayak, T., van Genabith, J., and Watanabe, H. (2016). CATaLog Online: A Web-Based CAT Tool for Distributed Translation with Data Capture for APE and Translation Process Research. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, 98–102. https://aclanthology.org/C16-2021

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In P. Isabelle, E. Charniak, and D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Paulsen Christensen, T., Bundgaard, K., Schjoldager, A., and Dam Jensen, H. (2022). What Motor Vehicles and Translation Machines Have in Common—A First Step Towards a Translation Automation Taxonomy. *Perspectives*, *30*(1), 19–38. https://doi.org/10.1080/0907676X.2021.1900307

Peris, Á., and Casacuberta, F. (2019). Online Learning for Effort Reduction in Interactive Neural Machine Translation. *Computer Speech & Language*, *58*, 98–126. https://doi.org/10.1016/j.csl.2019.04.001

Pickering, A. (1995). *The Mangle of Practice: Time, Agency, and Science*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226668253.001.0001

Pigott, I. M. (1986). Machine Translation as an Integral Part of the Electronic Office Environment. *Proceedings of Translating and the Computer 7*. TC7 Conference, London, UK. https://aclanthology.org/1985.tc-1.16.pdf

Popović, M. (2015). chrF: Character N-gram F-score for Automatic MT Evaluation. In O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, and P. Pecina (Eds.), *Proceedings of the Tenth Workshop on Statistical Machine Translation* (pp. 392–395). Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3049

Popović, M. (2018). Error Classification and Analysis for Machine Translation Quality Assessment. In J. Moorkens, S. Castilho, F. Gaspari, and S. Doherty (Eds.), *Translation Quality Assessment: From Principles to Practice* (pp. 129–158). Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7_7

Popović, M., Lommel, A., Burchardt, A., Avramidis, E., and Uszkoreit, H. (2014). Relations Between Different Types of Post-Editing Operations, Cognitive Effort and Temporal Effort. In M. Cettolo, M. Federico, L. Specia, and A. Way (Eds.), *Proceedings of the 17th Annual Conference of the European Association for Machine Translation* (pp. 191–198). https://aclanthology.org/2014.eamt-1.41

Popović, M., Lommel, A., Burchardt, A., Avramidis, E., and Uszkoreit, H. (2014). Relations Between Different Types of Post-Editing Operations, Cognitive Effort and Temporal Effort. In M. Cettolo, M. Federico, L. Specia, and A. Way (Eds.), *Proceedings of the 17th Annual Conference of the European Association for Machine Translation* (pp. 191–198). https://aclanthology.org/2014.eamt-1.41

Porto Veloso, P. (2013). *Escriba, an Adaptive Web CAT Tool* [MSc Dissertation]. University of Dublin. https://publications.scss.tcd.ie/theses/diss/2013/TCD-SCSS-DISSERTATION-2013-083.pdf

Pym, A. (2011). What Technology Does to Translating. *The International Journal of Translation and Interpreting Research*, *3*(1), 1–9. https://doi.org/10.3316/informit.409249876138991

Pym, A. (2013). Translation Skill-Sets in a Machine-Translation Age. *Meta : Journal Des Traducteurs / Meta: Translators' Journal*, *58*(3), 487–503. https://doi.org/10.7202/1025047ar

Quesenbery, W. (2003). The Five Dimensions of Usability. In M. J. Albers and M. Beth Mazur (Eds.), *Content and Complexity: Information Design in Technical Communication* (pp. 81–102). Lawrence Erlbaum. https://www.taylorfrancis.com/chapters/edit/10.4324/9781410607409-5/five-dimensions-usability-whitney-quesenbery

Rabadán, R. (2008). Refining the Idea of 'Applied Extensions'. In A. Pym, M. Shlesinger, and D. Simeoni (Eds.), *Beyond Descriptive Translation Studies* (pp. 103–118). John Benjamins. https://doi.org/10.1075/btl.75

Rahman, M. S., Shuhidan, S. M., Masrek, M. N., and Baharuddin, M. F. (2022). Validity and Reliability Testing of Geographical Information System (GIS) Quality and User Satisfaction towards Individual Work Performance. *Proceedings of the International Academic Symposium of Social Science 2022*, *82*, 68. https://doi.org/10.3390/proceedings2022082068

Ranasinghe, T., Orăsan, C., and Mitkov, R. (2020). Intelligent Translation Memory Matching and Retrieval with Sentence Encoders. In A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, and M. L. Forcada (Eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. https://aclanthology.org/2020.eamt-1

Raunak, V., Menezes, A., Post, M., and Hassan, H. (2023). Do GPTs Produce Less Literal Translations? In A. Rogers, J. Boyd-Graber, and N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 1041–1050). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-short.90

Recski, G., and Kádár, F. (2023). Language Complexity in Human and Machine Translation: A Preliminary Study. In C. Orăsan, R. Mitkov, G. Corpas Pastor, and J. Monti (Eds.), *Proceedings of the Human-informed Translation and Interpreting Technology International Conference* (pp. 268–281). Incoma Ltd. https://doi.org/10.26615/issn.2683-0078.2023_022

Rei, R., Guerreiro, N. M., Pombal, J., van Stigt, D., Treviso, M., Coheur, L., C. de Souza, J. G., and Martins, A. (2023). Scaling up CometKiwi: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task. In P. Koehn, B. Haddow, T. Kocmi, and C. Monz (Eds.), *Proceedings of the Eighth Conference on Machine Translation* (pp. 841–848). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.73

Reiss, K. (2000). Type, Kind and Individuality of Text: Decision Making in Translation. In L. Venuti (Ed.), *The Translation Studies Reader* (Vol. 2, pp. 168–181). https://doi.org/10.2307/1772491

Reiss, K., and Vermeer, H. J. (2014). *Towards a General Theory of Translational Action: Skopos Theory Explained*. Routledge. https://doi.org/10.4324/9781315759715

Reuneker, A. (2017). Lexical Diversity Measurements. https://www.reuneker.nl/files/ld/index.php

Rico, C. (2001). Reproducible Models for CAT Tools Evaluation: A User-Oriented Perspective. *Proceedings of Translating and the Computer 23*. TC23 Conference, London, UK. https://aclanthology.org/2001.tc-1.8

Risku, H. (2002). *Situatedness in Translation Studies*. Cognitive Systems Research, 3(3), 523–533. https://doi.org/10.1016/S1389-0417(02)00055-4

Rodríguez Vázquez, S., Fitzpatrick, D., and O'Brien, S. (2018). Is Web-Based Computer-Aided Translation (CAT) Software Usable for Blind Translators? In K. Miesenberger and G. Kouroupetroglou (Eds.), *Computers Helping People with Special Needs: Proceedings of the 16th International Conference, ICCHP 2018* (Vol. 10896, pp. 31–34). Springer International Publishing. https://doi.org/10.1007/978-3-319-94277-3_6

Rusu, C., Rusu, V., Roncagliolo, S., and González, C. (2015). Usability and User Experience: What Should We Care About? *International Journal of Information Technologies and Systems Approach (IJITSA)*, *8*(2), 1–12. https://doi.org/10.4018/IJITSA.2015070101

Sadiku, M., Ashaolu, T. J., Ajayi-Majebi, A., and Musa, S. (2021). Augmented Intelligence. *International Journal Of Scientific Advances*, *2*(5). https://doi.org/10.51542/ijscia.v2i5.17

SAE. (2016). *J2450_201608: Translation Quality Metric - SAE International*. https://www.sae.org/standards/content/j2450_201608/

Sager, J. C. (1994). *Language Engineering and Translation: Consequences of Automation*. John Benjamins. https://doi.org/10.1075/target.8.1.17sch

Santy, S., Dandapat, S., Choudhury, M., and Bali, K. (2019). INMT: Interactive Neural Machine Translation Prediction. In S. Padó and R. Huang (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations* (pp. 103–108). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-3018

Sauro, J., and Lewis, J. R. (2016). *Quantifying the User Experience: Practical Statistics for User Research*. Morgan Kaufmann.

Schrepp, D. M. (2023). *User Experience Questionnaire Handbook*. https://www.ueq-online.org/Material/Handbook.pdf

Schrepp, M., Thomaschewski, and Kollmorgen, J. (2023). A Comparison of SUS, UMUX-LITE, and UEQ-S - JUX. *Journal of User Experience*, *18*(2), 86–104.

Schrepp, M., Thomaschewski, J., and Hinderks, A. (2017). Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*, 103–108. https://doi.org/10.9781/ijimai.2017.09.001

Segarra Beneyto, M., and Huertas Abril, C. A. (2017). Automated Quality Assurance Tools: A Step Forward Towards Achieving Quality Translations. In A. Bécart, V. Merola, and R. López-Campos Bodineau (Eds.), *New Technologies Applied to Translation Studies: Strategies, Tools and Resources*. Editorial Bienza. https://www.researchgate.net/publication/323342184_New_Technologies_Applied_to_Translation_Studies_Strategies_Tools_and_Resources

Sharou, K. A., and Specia, L. (2022). A Taxonomy and Study of Critical Errors in Machine Translation. In H. Moniz, L. Macken, A. Rufener, L. Barrault, M. R. Costa-Jussà, C. Declercq, M. Koponen, E. Kemp, S. Pilos, M. L. Forcada, C. Scarton, J. Van den Bogaert, J. Daems, A. Tezcan, B. Vanroy, and M. Fonteyne (Eds.), *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation* (pp. 171–180). European Association for Machine Translation. https://aclanthology.org/2022.eamt-1.20

Shimanaka, H., Kajiwara, T., and Komachi, M. (2018). RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi, and K. Verspoor (Eds.), *Proceedings of the Third Conference on Machine Translation: Shared Task Papers* (pp. 751–758). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6456

Simard, M., and Foster, G. (2013). PEPr: Post-Edit Propagation Using Phrase-based Statistical Machine Translation. In A. Way, K. Sima'an, and M. L. Forcada (Eds.), *Proceedings of Machine Translation Summit XIV: Papers*. https://aclanthology.org/2013.mtsummit-papers.24/

Slator. (2022). *Slator Machine Translation Expert-in-the-Loop Report*. https://slator.com/machine-translation-expert-in-the-loop-report/

Slator (2024). *Language Industry Market Report*. https://slator.com/2024-language-industry-market-report-language-ai-edition/

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 223–231. https://aclanthology.org/2006.amta-papers.25

Sommerville, I. (2011). *Software Engineering*. Pearson.

Specia, L., and Wilks, Y. (2016). Machine Translation. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (2nd ed., p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199573691.013.26

Specia, L., Blain, F., Fomicheva, M., Zerva, C., Li, Z., Chaudhary, V., and Martins, A. F. T. (2021). Findings of the WMT 2021 Shared Task on Quality Estimation. In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-Jussa, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita, and C. Monz (Eds.), *Proceedings of the Sixth Conference on Machine Translation* (pp. 684–725). Association for Computational Linguistics. https://aclanthology.org/2021.wmt-1.71

Speerstra, N. (2018). A Comparison of Statistical and Neural MT in a Multi-Product and Multilingual Software Company—User Study. In J. A. Pérez-Ortiz, F. Sánchez-Martínez, M. Esplà-Gomis, M. Popović, C. Rico, A. Martins, J. Van den Bogaert, and M. L. Forcada (Eds.), *Proceedings of the 21st Annual Conference of the European Association for Machine Translation* (pp. 335–341). https://aclanthology.org/2018.eamt-main.34

Stasimioti, M. (2022). *How Fast Can You Post-Edit Machine Translation?* Slator. https://slator.com/how-fast-can-you-post-edit-machine-translation/

Stasimioti, M., Sosoni, V., Kermanidis, K., and Mouratidis, D. (2020). Machine Translation Quality: A Comparative Evaluation of SMT, NMT and Tailored-NMT Outputs. In A. Martins, H. Moniz, S.

Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, and M. L. Forcada (Eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 441–450). European Association for Machine Translation. https://aclanthology.org/2020.eamt-1.47

Surveypal. (2023). 10 Customer Service Benchmarks. https://surveypal.com/wp-content/uploads/2023/08/Customer-Service-Benchmarks-1.pdf

TAUS. (2016). *TAUS Quality Dashboard*.

Tavakol, M., and Dennick, R. (2011). Making Sense of Cronbach's Alpha. *International Journal of Medical Education*, 2, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Teixeira, C. (2011). Knowledge of Provenance and its Effects on Translation Performance in an Integrated TM/MT Environment. In B. Sharp, M. Zock, M. Carl, and A. L. Jakobsen (Eds.), *Proceedings of the 8th International NLPCS Workshop – Special Theme: Human-Machine Interaction in Translation. Copenhagen Studies in Language* (pp. 107–118). https://aclanthology.org/www.mt-archive.info/NLPSC-2011-Teixeira.pdf

Teixeira, C. S. C., Moorkens, J., Turner, D., Vreeke, J., and Way, A. (2019). Creating a Multimodal Translation Tool and Testing Machine Translation Integration Using Touch and Voice. *Informatics*, *6*(1), Article 1. https://doi.org/10.3390/informatics6010013

Temnikova, I. (2010). Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/437_Paper.pdf

Terribile, S. (2024a). AREA: Automating Repetitive Editing Actions in Post-Editing. *Proceedings of Translating and the Computer 46*, 87–100. https://asling.org/tc46/wp-content/uploads/2025/03/TC46-proceedings.pdf#page=87

Terribile, S. (2024b). Is Post-Editing Really Faster Than Human Translation? Translation Spaces, 13(2), 171–199. https://doi.org/10.1075/ts.22044.ter

Toral, A. (2019). Post-Editese: An Exacerbated Translationese. In M. Forcada, A. Way, B. Haddow, and R. Sennrich (Eds.), *Proceedings of the Machine Translation Summit XVII: Research Track* (pp. 273–281). European Association for Machine Translation. https://aclanthology.org/W19-6627

Toral, A. (2023). *Dig & Rewrite: NLP to Assist Translators.* Presentation at Convergence 2023.

Toral, A., and Sánchez-Cartagena, V. M. (2017). A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. In M. Lapata, P. Blunsom, and A. Koller (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 1063–1073). Association for Computational Linguistics. https://aclanthology.org/E17-1100

Toury, G. (2012). *Descriptive Translation Studies – and Beyond*. John Benjamins. https://doi.org/10.1075/btl.100

Ulitkin, I. (2013). Human Translation vs. Machine Translation. *Translation Journal*, *17*(1). https://translationjournal.net/journal/63mtquality.htm

Ustaszewski, M. (2019). Exploring Adequacy Errors in Neural Machine Translation with the Help of Cross-Language Aligned Word Embeddings. In I. Temnikova, C. Orăsan, G. Corpas Pastor, and R. Mitkov (Eds.), *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)* (pp. 122–128). Incoma Ltd., Shoumen, Bulgaria. https://doi.org/10.26615/issn.2683-0078.2019_015

Van Brussel, L., Tezcan, A., and Macken, L. (2018). A Fine-Grained Error Analysis of NMT, SMT and RBMT Output for English-to-Dutch. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA). https://aclanthology.org/L18-1600

van der Meer, J. (2021). *Translation Economics of the 2020s | MultiLingual July/August 2021*. https://multilingual.com/issues/july-august-2021/translation-economics-of-the-2020s/

Vanmassenhove, E., Shterionov, D., and Way, A. (2019). Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. In M. Forcada, A. Way, B. Haddow, and R. Sennrich (Eds.), *Proceedings of Machine Translation Summit XVII: Research Track* (pp. 222–232). European Association for Machine Translation. https://aclanthology.org/W19-6622

Vanroy, B., Tezcan, A., and Macken, L. (2023). MATEO: MAchine Translation Evaluation Online. In M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, T. Ranasinghe, E. Vanmassenhove, S. Alvarez Vidal, N. Aranberri, M. Nunziatini, C. Parra Escartín, M. Forcada, M. Popovic, C. Scarton, and H. Moniz (Eds.), *Proceedings of the 24th Annual Conference of the European*

Association for Machine Translation (pp. 499–500). European Association for Machine Translation (EAMT). https://aclanthology.org/2023.eamt-1.52

Vardaro, J., Schaeffer, M., and Hansen-Schirra, S. (2019). Translation Quality and Error Recognition in Professional Neural Machine Translation Post-Editing. *Informatics*, *6*(3), Article 3. https://doi.org/10.3390/informatics6030041

Vargas-Sierra, C. (2019). Usability Evaluation of a Translation Memory System. *Quaderns de Filologia: Estudis Lingüístics XXIV*, 119–146. https://doi.org/10.7203/QF.24.16302

Vela-Valido, J. (2021). Translation Quality Management in the AI Age: New Technologies to Perform Translation Quality Assurance Operations. *Tradumàtica*, *19*, 0093–0111. https://doi.org/10.5565/rev/tradumatica.285

Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. (2006). Error Analysis of Statistical Machine Translation Output. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, and D. Tapias (Eds.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)* (pp. 697–702). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf

Vinay, J.-P., and Darbelnet, J. (1972). *Stylistique comparée du français et de l'anglais: Méthode de traduction*. Didier.

Volkart, L., and Bouillon, P. (2023). Are Post-Editese Features Really Universal? In C. Orăsan, R. Mitkov, G. Corpas Pastor, and J. Monti (Eds.), *Proceedings of the Human-Informed Translation and Interpreting Technology International Conference* (pp. 294–304). https://doi.org/10.26615/issn.2683-0078.2023_025

Vygotsky, L. S. (1978). *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4

Wachowicz, M., Cui, L., Vullings, W., and Bulens, J. (2008). The Effects of Web Mapping Applications on User Satisfaction: An Empirical Study. In M. P. Peterson (Ed.), *International Perspectives on Maps and the Internet* (pp. 397–415). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72029-4_25

Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., and Tu, Z. (2023). Document-Level Machine Translation with Large Language Models. In Bouamor, H. Pino, J., and Bali, K. (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 16646–16661). 10.18653/v1/2023.emnlp-main.1036

Wei, Y., Utsuro, T., and Nagata, M. (2022). *Extending Word-Level Quality Estimation for Post-Editing Assistance* (arXiv:2209.11378). Preprint. https://doi.org/10.48550/arXiv.2209.11378

White, J. S., O'Connell, T. A., and O'Mara, F. E. (1994, October 5). The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. *Proceedings of the First Conference of the Association for Machine Translation in the Americas*. AMTA 1994, Columbia, Maryland, USA. https://aclanthology.org/1994.amta-1.25

Xiao, K., and Muñoz Martín, R. (2020). Cognitive Translation Studies: Models and Methods at the Cutting Edge. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, *19*. https://doi.org/10.52034/lanstts.v19i0.593

Zaretskaya, A. (2016). A Quantitative Method for Evaluation of CAT Tools Based on User Preferences. In M. F. Litzler, J. García Laborda, and C. Tejedor Martínez (Eds.), *Beyond the Universe of Languages for Specific Purposes: The 21st Century Perspective* (pp. 153–158). Servicio de Publicaciones de la Universidad de Alcalá. https://dialnet.unirioja.es/servlet/articulo?codigo=5694699

Zaretskaya, A. (2017). *Translators' Requirements for Translation Technologies: User Study on Translation Tools* [Ph.D. Thesis]. Universidad de Málaga. https://riuma.uma.es/xmlui/bitstream/handle/10630/16246/TD_ZARETSKAYA_Anna.pdf?sequence=1

Zaretskaya, A. M., Vela, G., Seghiri, M., and Corpas Pastor, G. (2016a). Comparing Post-Editing Difficulty of Different Machine Translation Errors in Spanish and German Translations from English. *International Journal of Language and Linguistics*, *3*(3), 91–100. https://doi.org/10.30845/ijll

Zaretskaya, A., Corpas Pastor, G., and Seghiri, M. (2017). User Perspective on Translation Tools: Findings of a User Survey. In G. Corpas Pastor and I. Durán-Muñoz (Eds.), *Trends in E-Tools and Resources for Translators and Interpreters* (pp. 37–56). Brill. https://doi.org/10.1163/9789004351790_004

Zaretskaya, A., Pastor, G. C., and Seghiri, M. (2015). Integration of Machine Translation in CAT Tools: State of the Art, Evaluation and User Attitudes. *Skase Journal of Translation and Interpretation*, *8*(1), 76–89. http://www.skase.sk/Volumes/JTI09/pdf_doc/04.pdf

Zaretskaya, A., Vela, M., Corpas Pastor, G., and Seghiri, M. (2016b). Measuring Post-editing Time and Effort for Different Types of Machine Translation Errors. *New Voices in Translation Studies*, *15*, 63–91. https://doi.org/10.14456/nvts.2016.21

Zeng, J., Meng, F., Yin, Y., and Zhou, J. (2024). Improving Machine Translation with Large Language Models: A Preliminary Study with Cooperative Decoding. In L.-W. Ku, A. Martins, and V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 13275–13288). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.786