# Working with Pre-translated Texts: Investigating Machine Translation Post-editing and Human Translation Revision at Swiss Corporate In-house Language Services

Girletti, Sabrina

UNIVERSITÉ DE GENÈVE
FACULTY OF TRANSLATION AND INTERPRETING

# Working with Pre-translated Texts: Investigating Machine Translation Post-editing and Human Translation Revision at Swiss Corporate In-house Language Services

A thesis submitted at the Faculty of Translation and Interpreting (FTI) in fulfillment of the requirements for the degree of Doctor of Philosophy

Author

Sabrina Girletti

Jury

- Prof. Pierrette Bouillon, Faculty of Translation and Interpreting, University of Geneva (Supervisor)
- Prof. Annarita Felici, Faculty of Translation and Interpreting, University of Geneva (Jury president)
- Prof. Maarit Koponen, School of Humanities, Foreign Languages and Translation Studies, University of Eastern Finland (External jury member)
- Prof. Alexander Künzli, Faculty of Translation and Interpreting, University of Geneva (Jury member)
- Prof. Marie-Aude Lefer, Louvain School of Translation and Interpreting, UCLouvain (External jury member)

Thesis No. 50

# Declaration of Authorship (fr)

## Déclaration attestant le caractère original du travail effectué

J'affirme avoir pris connaissance des documents d'information et de prévention du plagiat émis par l'Université de Genève et la Faculté de traduction et d'interprétation (notamment la Directive en matière de plagiat des étudiante-s, le Règlement d'études de la Faculté de traduction et d'interprétation ainsi que l'Aide-mémoire à l'intention des étudiants préparant un mémoire de Ma en traduction).

J'atteste que ce travail est le fruit d'un travail personnel et a été rédigé de manière autonome. Je déclare que toutes les sources d'information utilisées sont citées de manière complète et précise, y compris les sources sur Internet.

Je suis consciente que le fait de ne pas citer une source ou de ne pas la citer correctement est constitutif de plagiat et que le plagiat est considéré comme une faute grave au sein de l'Université, passible de sanctions.

Au vu de ce qui précède, je déclare sur l'honneur que le présent travail est original.

Signature electronique : Sabrina Girletti

Lieu et date : Genève, le 29 septembre 2023

# Abstract

With the advent of neural machine translation (NMT), machine-generated texts increasingly resemble human translations. This paradigm shift calls for a comprehensive understanding of the implications associated with working with machine-pre-translated and human-pre-translated texts. The present thesis sets out to explore the similarities and differences between post-editing (PE) machine translation (MT) content and traditional translation revision in the professional context of Swiss corporate in-house language services (CILS).

Two overarching goals frame this investigation. The first goal aims to gather information on the use of MT in Swiss CILS and on how PE and revision tasks are organised in CILS' workflows. The second goal aims to understand how PE and revision compare in the professional context of Swiss CILS from the perspective of in-house linguists. Four key aspects have been outlined to guide this comparative analysis, namely linguists' PE and revision practices, satisfaction, productivity and modifications (edits) made to pre-translated texts during both tasks. To achieve the abovementioned goals, this thesis adopts a multimethod empirical approach, combining data from two different surveys, a field experiment, and a corpus-based analysis of authentic PE and revision assignments.

The data gathered throughout these studies suggest that the boundaries between PE and revision are not as clear-cut as previously perceived. For instance, in the majority of surveyed CILS, PE is performed on fully pre-translated texts, typically as the last step in the workflow, most often resulting in post-edited texts that remain unrevised.

The analysis of PE and revision practices shows that linguists claim to adopt different approaches depending on whether they are working with human-translated or machine-translated texts. However, in practice, they often apply the same reading strategies to these texts. Reading texts segment by segment is the most commonly employed reading strategy in both tasks, emphasising the impact of text display within a computer-assisted translation tool on PE and revision processes. The findings also uncover that post-editors tend to repeat the same reading strategies, while revisers more often adapt them depending on text type, time constraints, and the translator who performed the translation.

According to data on linguists' satisfaction with revision and PE assignments, the latter are less fulfilling and slightly less demanding than revision tasks. However, PE — more than revision — enables linguists to create new content and exert control over the text's final quality. Post-editors who received proper PE training report higher satisfaction levels

with PE tasks than those who did not receive such training. Conversely, revision training seems less effective in influencing overall satisfaction levels in revision tasks.

The results of the field experiment indicate that linguists are faster during revision than during PE. However, the use of NMT seems to enhance processing speed during PE tasks compared to the previous generation of MT systems. The findings also shed light on the potential quality-related risks of bypassing the verification step performed by a second linguist in PE workflows. Indeed, a higher percentage of sentences in post-edited texts requires further correction compared to revised texts.

Lastly, the analysis of PE and revision assignments shows that CILS linguists edit texts more frequently during PE tasks than revision tasks. Furthermore, human-pre-translated sentences require modifications that are narrower in scope compared to machine-pre-translated sentences. Although the percentage breakdown of editing actions is similar between the two tasks, a statistically significant difference indicates that the task influences the distribution of editing actions. In particular, revision assignments include a higher percentage of deletions compared to PE assignments. Edits in PE assignments focus on addressing mistranslations, while revision assignments mainly involve providing synonyms or rephrasing sentences. The majority of edits performed in both PE and revision tasks are considered necessary, although revisers tend to make slightly more optional modifications than post-editors.

In conclusion, this thesis contributes to the growing body of knowledge in translation and interpreting workplace research, enriching our understanding of contemporary professional translation practices and providing valuable insights for translation pedagogy.

# Résumé

Avec l'arrivée de la traduction automatique neuronale (TAN), les textes générés par les machines ressemblent de plus en plus à des traductions humaines. Ce changement de paradigme demande une analyse globale des implications associées au travail avec des textes pré-traduits par une machine et par un humain. Cette thèse explore ainsi les similitudes et les différences entre la post-édition (PE) de la traduction automatique (TA) et la révision traditionnelle de la traduction humaine dans le contexte professionnel des services linguistiques internes aux entreprises suisses (SLIE).

Cette étude s'articule autour de deux objectifs principaux. Premièrement, recueillir des informations sur l'utilisation de la TA au sein des SLIE et sur l'organisation des tâches de PE et de révision dans leur flux de travail. Deuxièmement, examiner la PE et la révision dans ces contextes professionnels selon quatre aspects, à savoir les pratiques, la satisfaction et la productivité des linguistes, ainsi que les modifications apportées aux textes pré-traduits au cours des deux tâches. Pour atteindre les objectifs susmentionnés, cette thèse adopte une approche empirique multiméthode, qui repose sur diverses stratégies de recherche et méthodes de génération de données. Plus précisément, cette thèse triangule des données provenant de deux enquêtes différentes, d'une expérience sur le terrain et d'une analyse fondée sur des corpus de textes révisés et post-édités.

Les données recueillies dans le cadre de ces études confirment que les limites entre la PE et la révision ne sont pas aussi nettes qu'auparavant. Par exemple, la majorité des SLIE qui ont répondu à notre enquête effectuent la PE sur des textes entièrement pré-traduits. Dans ces contextes, la PE constitue généralement la dernière étape du flux de travail, ce qui aboutit le plus souvent à des textes post-édités non-révisés.

L'analyse des pratiques de PE et de révision a montré que les linguistes ont l'impression d'adopter des approches différentes selon qu'ils travaillent sur des textes traduits par des humains ou sur des textes traduits par des machines. Dans la pratique, ils appliquent cependant souvent les mêmes stratégies de lecture. Les stratégies de lecture les plus utilisées dans les deux tâches consistent à lire les textes segment par segment, ce qui montre l'influence de la présentation du texte dans les outils de traduction assistée par ordinateur sur les procédures de PE et de révision. Les résultats ont également montré que les post-réviseurs ont tendance à utiliser systématiquement les mêmes stratégies de lecture, tandis que les

réviseurs les adaptent plus souvent en fonction du type de texte, des contraintes de temps et du traducteur qui a effectué la traduction.

En ce qui concerne la satisfaction des linguistes, les données ont montré que la PE est considérée comme moins gratifiante et légèrement moins exigeante que la révision. La PE, plus que la révision, donne cependant aux linguistes la sensation de créer un nouveau contenu et de pouvoir exercer un certain contrôle sur la qualité finale du texte. Les post-éditeurs qui ont reçu une formation adéquate à la PE sont plus satisfaits que ceux qui n'ont pas reçu une telle formation. Inversement, une formation en révision semble moins efficace pour influencer les niveaux de satisfaction globaux dans les tâches de révision.

Dans l'expérience sur le terrain, nous avons observé que les linguistes étaient plus rapides lors de la révision que lors de la PE, bien que l'utilisation de la TAN semble améliorer la productivité lors des tâches de PE par rapport à la génération précédente de systèmes de TA. Les résultats ont également mis en lumière les risques liés au fait de ne pas passer par une étape de révision après la PE. En effet, un pourcentage plus élevé de phrases dans les textes post-édités semble nécessiter de corrections supplémentaires par rapport aux textes révisés.

Enfin, l'analyse des tâches de PE et de révision a montré que les linguistes effectuent plus de modifications pendant les tâches de PE que pendant les tâches de révision. Une grande partie des phrases du corpus de révision n'ont subi que de légères modifications, alors que la majorité des phrases des textes post-édités ont subi des changements plus importants. Bien que la répartition des actions d'édition soit similaire entre les deux tâches, les résultats indiquent de manière statistiquement significative que la tâche a une influence sur la distribution des actions d'édition. En particulier, les textes révisés comprennent un pourcentage plus élevé de suppressions que les textes post-édités. Les modifications effectuées en PE se concentrent sur les erreurs de traduction, tandis que celles de révision consistent principalement à fournir des synonymes ou à reformuler des phrases. La majorité des modifications effectuées dans les tâches de PE et de révision ont été jugées nécessaires, bien que les réviseurs aient eu tendance à apporter un peu plus de modifications facultatives que les post-éditeurs.

En conclusion, cette thèse apporte une contribution à la recherche dans le domaine de la traduction et de l'interprétation en contexte professionnel. Elle fournit également des informations utiles pour la pédagogie de la traduction.

# Acknowledgements

Here it is, the day that I have been dreaming of for years. This thesis has allowed me to explore a topic in-depth, but most of all it has allowed me to discover many more things about myself and what I am capable of when I choose self-encouragement over self-criticism. I am now ready to thank the people who have played a role in this journey.

First and foremost, I want to extend my deepest gratitude to my supervisor, Professor Pierrette Bouillon. Your guidance, wisdom, and the freedom you afforded me to work in optimal conditions have been instrumental in shaping this thesis.

I am sincerely grateful to the members of my thesis committee, Professor Annarita Felici, Professor Maarit Koponen, Professor Alexander Künzli and Professor Marie-Aude Lefer, who accepted the responsibility of evaluating my work. I also want to thank Anne-Charlotte Bacherot from Swiss Post for her invaluable support and for providing the data for a crucial part of my PhD project.

Writing this thesis has been a long and arduous endeavor, but the support and encouragement of several remarkable people have lightened the load considerably. Elisa, Irene, Paolo, Lucía M., Marianne, Lucía O., Jony and Lise: sharing my PhD with you all has made this journey not only academically enriching but also filled with laughter and warmth. A special mention goes to Marco, who not only provided assistance with statistics but also proved to be the best office mate I could have asked for. I am deeply indebted to Silvia: our fruitful discussions about methodology and your invaluable feedback on some chapters of this thesis have been enlightening. You are a true inspiration as a researcher and as a woman. To all of you, my dear friends, I am extremely grateful. You have been the silver lining of my PhD, and I cherish the moments we have shared.

Lastly, I want to thank my husband. Your unwavering support and encouragement, along with your words during moments of doubt, have been my anchor throughout these years. I know that the simple fact that I am submitting my thesis makes you the happiest person in the world. And this, alone, erases all the struggles I have endured along the way.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| **APE** | Automatic Post-Editing |
| **API** | Application Programming Interface |
| **BLEU** | BiLingual Evaluation Understudy |
| **CAT** | Computer-Assisted Translation |
| **CILS** | Corporate In-house Language Service |
| **CSA** | Common Sense Advisory |
| **DGT** | Directorate-General for Translation |
| **ELIS** | European Language Industry Survey |
| **EMT** | European Master's in Translation |
| **HER** | Human Edit Rate |
| **HTER** | Human-targeted Translation Edit Rate |
| **ISO** | International Organization for Standardization |
| **JDI** | Job Descriptive Index |
| **JDS** | Job Diagnostic Survey |
| **JSS** | Job Satisfaction Survey |
| **LSP** | Language Service Provider |
| **MQM** | Multidimensional Quality Metrics |
| **MSQ** | Minnesota Satisfaction Questionnaire |
| **MT** | Machine Translation |
| **MTPE** | Machine Translation Post-Editing |
| **NMT** | Neural Machine Translation |
| **PBMT** | Phrase-Based Machine Translation |
| **PE** | Post-Editing |
| **POS** | Part Of Speech |
| **QA** | Quality Assurance |
| **RBMT** | Rule-Based Machine Translation |
| **REV** | Revision |
| **SMT** | Statistical Machine Translation |
| **T&I** | Translation and Interpreting |
| **TAUS** | Translation and Automation User Society |
| **TER** | Translation Edit Rate |
| **TR** | Test Round |
| **TTR** | Type-Token Ratio |
| **TM** | Translation Memory |

# Chapter 1

# Introduction

> *La post-édition des traductions automatiques diffère de la tâche de révision à plusieurs égards. Dans l'approche de base, la révision est un processus de découverte, tandis que la post-édition est un exercice permanent d'ajustement. Les deux processus visent principalement à corriger les erreurs, mais la nature des types d'erreurs est quelque peu différente, tout comme leur répartition. Avec la traduction automatique, le post-éditeur a l'assurance que rien n'a été omis ou répété. Des erreurs lexicales seront présentes dans la sortie de la TA, mais elles sont plus prévisibles que celles commises par les traducteurs humains. Comme pour la révision traditionnelle, des erreurs d'interprétation devront peut-être être corrigées, mais elles auront tendance à avoir une portée plus limitée. Les deux processus se ressemblent le plus dans le cas d'une traduction destinée à être publiée.* [1]

The text above is an automated French translation of the abstract from Muriel Vasconcellos' article titled "A comparison of MT post-editing and traditional revision" (1987). In her article, Vasconcellos reports on the use of machine translation (MT) at the Pan American Health Organization, one of the first international organisations to adopt MT systems in the 1980s (Vasconcellos, 1989). The author highlights that "MT post[-]editing differs from traditional revision in terms of both the overall approach and the type of errors that need to be corrected" (Vasconcellos, 1987, p. 145). She argues that one positive aspect of working with a machine, as opposed to a human translator, is that machine errors are local and predictable. Most importantly, when working with MT, "[one] can be confident that nothing has been skipped or repeated" (*ibid.*., p. 411).

Vasconcellos' analysis refers to rule-based machine translation, the first commercially available system architecture. Yet, her considerations remained valid even when statistical machine translation became the dominant paradigm. The latest generation of MT systems, neural machine translation (NMT), emerged around 2015 and quickly gained recognition

1. "Post[-]editing of machine translations differs from the task of revision in several respects. In basic approach, revision is a discovery process, while post[-]editing is an ongoing exercise of adjustment. Both processes are primarily concerned with the correction of errors, but the nature of error types is somewhat different, as is their distribution. With machine translation, the post[-]editor has the assurance that nothing has been skipped or repeated. Lexical errors will be present in the MT output, but they are more predictable than the kind committed by human translators. As with traditional revision, misinterpretations may need to be corrected, but they will tend to be narrower in scope. The two processes are most alike in the case of a translation that is to be published."

as a game-changer (King, 2020). Due to the significantly enhanced quality of its output, considered to be more fluent than ever before (Castilho et al., 2017a, 2018; Klubička, Toral, and Sánchez-Cartagena, 2017; Toral and Sánchez-Cartagena, 2017), NMT has garnered increasing attention from the translation and localisation industry, as well as from the media (ELIA et al., 2019; Nunes Vieira, 2020a). Indeed, NMT systems can generate texts with minimal grammatical and syntactic errors, although the fluency of their output may sometimes mask certain inaccuracies.

The text presented at the beginning of the chapter, generated by DeepL [2], one of the most famous NMT systems freely available on the internet, exemplifies the quality that these systems can deliver. In particular, this text resembles a human translation more than a machine-generated one. With the advent of NMT, the approach to machine-translated texts has undergone a significant transformation: errors are often rare, unpredictable, and not always easy to detect. Consequently, the characteristics that previously marked the differences between post-editing and revision (Vasconcellos, 1987; Wagner, 1985) may no longer hold true. This prompts us to explore ***the extent to which the tasks of correcting MT or human translations differ*** – a question with practical and pedagogical relevance. To address this question, we conducted an empirical investigation of post-editing and revision tasks, involving professional translators working in Swiss corporate in-house language services (CILS).

This first chapter provides an overview of our research endeavours. It encompasses essential background information and fundamental concepts crucial for the reader's understanding of our research (Section 1.1). It also delves into the underlying motivation that drives our work (Section 1.2), outlines the specific goals we aim to achieve and presents the research questions that guide our investigation, along with the methodology adopted to address them (Section 1.3). Lastly, it outlines the structure of the thesis, detailing how the content is organised and presented (Section 1.4).

## 1.1 Background

This section aims to provide the reader with fundamental definitions and theoretical background that will serve as a guiding framework for our research. We will first outline the main characteristics of computer-assisted translation tools and MT systems. Then, we will describe the two tasks that form the core of our investigation, post-editing and revision, and review academic research on these two topics. Finally, we will present the limited number of studies focusing on both tasks [3].

2. `www.deepl.com/en/translator`. Accessed 13 March 2023.

3. The reader should note that this section serves as an introduction to the present thesis. A more extensive literature review of relevant studies will be presented in subsequent chapters, as outlined in Section 1.4.

It is worth noting that in 2021, Koponen, Mossop, Robert and Scocchera edited a book entitled "Translation Revision and Post-Editing: Industry Practices and Cognitive Processes" (Koponen et al., 2021a). This book includes a comprehensive literature review of theoretical and empirical studies on revision and post-editing published until 2019. While this literature review has been valuable in developing the present thesis, we will adopt a distinct perspective by primarily focusing on specific aspects related to the post-editing and revision tasks that have received less attention in previous work [4]. Additionally, our review will encompass studies published up until June 2023.

### 1.1.1 Translation technologies

The era when translators relied on pen and paper dictionaries to translate texts is long gone. Nowadays, translators' workspace is increasingly technologised and translation is considered a form of human-computer interaction (O'Brien, 2020). Industry surveys (Pielmeier and O'Mara, 2020; ELIS, 2023) report that language service companies, translation departments, and independent translators widely adopt computer-assisted translation (CAT) tools. These tools enable translators to streamline the translation process by leveraging the potential of previous content and are, therefore, particularly valuable when working with similar or repetitive texts.

The main feature of CAT tools is the translation memory (TM) [5], a database of previously translated texts aligned with their source version and stored in a sentence-by-sentence format, known as *translation units* (TUs). When a new text is uploaded to the CAT tool interface, it is automatically split into sentences (i.e., *segments*), which are compared against those stored in the TM. When a *match* is found, the tool indicates the degree of similarity, expressed as a percentage, between the segment and one or more TUs in the database. TM suggestions can be categorised as *fuzzy matches*, which exhibit up to 99% similarity between source language segments, or *100% matches*, also known as *exact matches* [6].

CAT tools enable translators to work with various file types, focusing on the content to be translated while the formatting is automatically applied to the final text. Several CAT tool interfaces are adjustable to meet users' preferences, and translators can also determine how TM suggestions should be presented to them. For instance, suggestions can appear in a dedicated window once the translator selects a source segment, or the highest match available can be directly inserted into the editing zone. Users also have the option to generate a pre-translation of the entire source text, for example, using machine translation when no matches from the TM are found. Indeed, in addition to serving as an editing environment

4. See *Key aspects* presented in Section 1.3.

5. It is worth noting that CAT tools are sometimes called "translation memory tools" (Kenny, 2022), and the term "translation memory" is often erroneously used to designate CAT tools. To avoid this confusion, Zetzsche (2007) proposed referring to CAT tools with the more encompassing term "Translation Environment Tools" (TEnTs).

6. Some CAT tools also indicate matches above 100%: these are called *context matches*, when also the preceding and the following segments are exact matches, or *perfect matches* if they come from reference material associated with the translation project, rather than from the TM. For a comprehensive explanation of the functioning of CAT tools, please refer to Rothwell et al. (2023).

and housing a database of previously translated text, CAT tools also incorporate various translation aids, including terminology management tools, quality assurance (QA) tools, and, more recently, machine translation. In this thesis, we will specifically focus on the latter's characteristics.

Machine translation is the automatic translation of text performed by computer software (Rothwell et al., 2023, p. 97). The idea of using computers on natural languages emerged in the 1950s in the United States, drawing inspiration from the achievements in codebreaking during the Second World War. In its early stages, research in MT received substantial funding from the government, until in 1966, a report of a dedicated committee (Automatic Language Processing Advisory Committee, ALPAC) [7] concluded that achieving fully automatic high-quality MT was both unfeasible and unnecessary (National Research Council, 1966). Nonetheless, in the 1970s, international organisations began employing MT systems, starting with the European Commission, while commercial utilisation of MT emerged in the 1980s (Hutchins, 2010).

The earliest MT systems were based on the idea that, to translate, computers needed to acquire knowledge about two languages and the rules for transforming sentences from one language into another. The first generation of MT systems, known as rule-based (RBMT) systems, relied on electronic dictionaries and grammars to perform a more or less in-depth analysis of the source text. Professional linguists were required to develop and maintain the dictionaries and the rules that mapped the words and sentence structures between source and target languages. This process demanded considerable time and resources, which posed a challenge for scaling and adapting RBMT systems to different languages and domains.

Further developments in MT took advantage of translation memories. In the 1990s, statistical machine translation (SMT) began to replace rule-based MT as the dominant approach. SMT falls under the category of data-driven MT, where linguistic knowledge is not manually encoded by professional linguists but instead derived from parallel corpora. These systems calculate the probability that a target sentence is the translation of a source sentence based on the probabilities of equivalence between source and target chunks (referred to as *n-grams*) and the likelihood of a sequence of n-grams being acceptable in the target language. The machine automatically learns these associations during a *training phase*, using TMs as training corpora. SMT systems remained state-of-the-art in MT for over two decades. They began to be integrated into various CAT tools and were successfully applied in commercial contexts (Flournoy and Duran, 2009; Schaefer, Van De Walle, and Van den Bogaert, 2014; Hofmann and Lepan, 2017). The output of these systems was more fluent compared to that of RBMT, but one of their main limitations was the inability to consider larger contexts during the translation process. Indeed, since SMT systems operated on sentence chunks, they often struggled to correctly render the syntactic and semantic relations between constituents. Another limitation of SMT systems was their difficulty in generalising when encountering previously unseen words (i.e. words not present in the training data).

7. For a comprehensive discussion of the ALPAC report and its consequences, the reader can refer to Hutchins (2003) and Poibeau (2017).

These limitations prompted further research in the field, leading to the development of a new approach, which has been considered "the best performing type of machine translation invented so far" (Kenny, 2022, p. 43). Neural MT systems [8], introduced in 2015, are the latest form of data-driven MT. Unlike SMT, NMT systems rely on particular groups of algorithms, called *artificial neural networks*, to create numerical, distributed representations of words based on contextual information. These representations, known as *word embeddings*, enable the systems to consider the entire source sentence, capturing the underlying dependencies between words and producing more fluent and contextually coherent translation. NMT systems have improved over previous approaches, generating translations with fewer errors. However, these errors can be more difficult to detect due to the high fluency of the output. Occasionally, NMT systems may produce incorrect translations or so-called 'hallucinations' (Lee et al., 2018) [9], especially when dealing with out-of-domain or low-quality training data. Gender bias has also been reported as an issue in NMT systems (Vanmassenhove, Shterionov, and Gwilliam, 2021; Savoldi et al., 2021). It is worth noting that, at the time of writing, most commercial NMT systems are limited to sentence-level processing and cannot look beyond individual sentences despite ongoing research efforts in this regard (Maruf, Saleh, and Haffari, 2021). This means that some issues may only be visible when considering the entire text rather than isolated sentences.

Despite these limitations, the introduction of NMT has significantly impacted both MT research and translation practice. Claims of achieving human parity in translation quality have been made (Hassan et al., 2018), although they have also been met with criticism by other researchers (Toral et al., 2018; Läubli, Sennrich, and Volk, 2018; Läubli et al., 2020a) – since it depends on *how* one measures quality. Nonetheless, the growing popularity of NMT systems has also brought MT to the attention of translation scholars. Indeed, as noted by Rico and Pastor (2022, p. 699):

> The fact that MT has gone almost unnoticed in Translation Studies for decades is probably the result of the poor capabilities of computers when it comes to processing languages.[...] It was only when MT research turned to neural networks and output improved in quality, that the machine-translated text finally merited being taken into consideration.

The quality delivered by modern MT systems depends on multiple factors, such as the language pairs considered and the size and quality of the training corpus. The data used to train the systems mark the difference between *generic* and *custom MT*. Customised MT systems [10] are trained with in-domain data to fulfil specific purposes, while generic MT systems aim to cover a wide range of topics (Ramírez-Sánchez, 2022).

8. A detailed explanation of the technical details behind NMT is beyond the scope of the present thesis. For a comprehensive understanding of NMT, including its underlying architectures and training methods, we recommend referring to the work of Pérez-Ortiz, Forcada, and Sánchez-Martínez (2022).

9. Hallucinations are semantically incorrect but grammatically valid translations that are also "completely untethered from the input" (Lee et al., 2018, p. 2).

10. Kenny (2022, pp. 44–45) uses the terms *engine* to refer to customised MT and *system* to refer to generic MT. In the present thesis, we do not subscribe to this distinction and employ the two terms as synonyms.

In professional translation workflows, MT is mainly used within CAT tools (Pielmeier and Lommel, 2019) through a plugin or an application programming interface (API). Zaretskaya, Pastor, and Seghiri (2015) propose to classify methods to integrate MT and TM suggestions into two categories: internal and external integration. The former consists in leveraging MT techniques to improve TM suggestions, while the latter consists in providing an MT suggestion in addition (or in substitution) to the TM one [11]. External integration methods include offline pre-translation of the whole text (i.e., *batch processing*), and real-time processing of the source text in a segment-by-segment fashion (*ibid.*, pp. 81-82). In the latter case, MT suggestions can appear along with TM suggestions or complete in real time the sentence typed by the user in the so-called *autocompletion mode*, which is, however, typical of interactive MT (Peris, Domingo, and Casacuberta, 2017) and implemented by few commercial CAT tools [12].

Before concluding this section, we would like to stress that the seamless integration of MT into CAT tools can create some confusion among translators regarding the functioning of MT and TM. What distinguishes these two technologies is the fact that MT is capable of generating a translation without human input, while the TM is built over time by translators themselves, who add new translation units to the database. Nonetheless, advanced MT leveraging techniques in CAT tools make the distinction between MT and TM less clear-cut (Rothwell et al., 2023, pp. 122–128).

### 1.1.2 Post-editing

Soon after starting research in machine translation in the 1950s, it became clear that post-editing (PE, also abbreviated MTPE) was needed to obtain an acceptable final text in most cases (Koby, 2001; García, 2012). At that time, researchers worked on developing fully automatic, high-quality MT systems; therefore, post-editing, i.e., correcting errors in raw MT output, was seen as an undesirable step (Nunes Vieira, 2020b; Rico and Pastor, 2022). According to Hutchins (1986, p. 31), the term "post-editor" was introduced by Reifler in 1950 and referred to the person whose role is "to select the correct translation from the possibilities found by the computer dictionary and to rearrange the word order to suit the target language". Post-editors were employed in MT research departments and were considered MT's "human partners" (Bar-Hillel, 1951, p. 230). In many cases, they were not required to understand the source language but only the target language and the texts' subject matter (Yngve, 1954; Edmundson and Hays, 1958). These definitions have created a negative reputation for the PE activity and the post-editor as a job position. As Nunes Vieira (2020b, p. 319) pointed out:

> Post-editing [...] came about as part of a paradigm where human editors assisted the machine rather than one where the machine assisted them. Given this machine-centred nature of the activity and the poor results — from today's perspective — that MT was

11. For a comprehensive review of MT-TM integration in commercial CAT tools, the reader can refer to Quintana and Castilho (2022).
12. See, for instance, Lilt (`www.lilt.com/en`).

> able to deliver then, it is not surprising that post-editing has developed an unfavourable reputation throughout the history of MT.

Despite the negative comments listed in the ALPAC report [13], PE activity paved its way within large corporations as a support for technical translation, and within institutions such as the European Commission as a solution to translate urgent texts not intended for publication (García, 2012). Capitalising on the advancements in MT research, PE became a service that language service providers (LSPs) could offer to some of their customers to reduce translation costs and delivery time. In 2017, the International Organization for Standardization (ISO) published a specific standard on the requirements that LSPs should meet to offer post-editing services to their customers. Here, *post-editing* is defined as "edit[ing] and correct[ing] machine translation output" (ISO, 2017, p. 2). In this document, it is noted that PE helps translation service providers improve translation productivity and turn-around times, allowing them to "remain competitive in an environment where clients show an increasing demand for using MT in translation" (*ibid.*, p. v).

Two main levels of PE have been defined, depending on the aim of the final text (Wagner, 1985; Allen, 2003; O'Brien, 2022). *Light PE* should be applied if the translation only provides a general idea of what the source text contains, i.e., it is used for gisting purposes. On the other hand, if the translation is to be published, i.e., MT is used for dissemination purposes, a *full PE* is needed. Other authors have suggested more than two PE levels. For instance, Egdom and Pluymaekers (2020) present up to four levels, namely *minimal, light, moderate* and *full*, which are inspired by degrees of revision set out by Mossop (2020). Post-editing levels are associated with specific guidelines (Hu and Cadwell, 2016), which are often difficult for linguists to follow (Flanagan and Christensen, 2014). The ISO standard mentions only the two main levels, light and full PE, and provides guidelines similar to those of the Translation and Automation Users Society (TAUS, Massardo et al., 2016). These guidelines detail the issues that post-editors should look for when post-editing and set out different levels of final quality, namely "good-enough" quality for light PE and "human-like" translation quality for full PE. Nevertheless, since NMT architectures have improved raw MT output quality (Bentivogli et al., 2016), many of the major issues to be corrected during light PE are not in the output anymore. Therefore, it has been claimed that the notion of light PE itself should be reviewed, as well as its corresponding guidelines (Nunes Vieira, 2020b).

Research has explored different aspects of the PE activity, such as the ability to carry out PE without the source text (*monolingual PE*; Mitchell, Roturier, and O'Brien, 2013; Schwartz, 2014; Koponen and Salmi, 2015; Nitzke, 2016), the quality of post-edited texts (Depraetere, De Sutter, and Tezcan, 2014; Koponen and Salmi, 2017), the correlation between PE performance and translation experience (Fiederer and O'Brien, 2009; de Almeida and O'Brien, 2010; Guerberof Arenas, 2014b), translators' productivity gains (Plitt and Masselot, 2010; Guerberof Arenas, 2012; Läubli et al., 2019) and the effort involved in the

13. The report clearly stated that PE "took slightly longer to do and was more expensive than conventional human translation" (National Research Council, 1966, p. 19).

PE task (Krings, 2001; O'Brien, 2006a, 2011; Carl et al., 2011; Koponen, 2016). The latter, in particular, has been extensively investigated.

Krings (2001) has defined three main sources of effort, namely *temporal, technical*, and *cognitive* effort. Temporal effort refers to the time spent post-editing and is often the most used measure of effort (Popović et al., 2014). Technical effort includes edit operations performed on the keyboard to carry out the PE task and is usually measured through keylogging tools. Cognitive effort refers to the cognitive load required to detect and correct issues in raw MT output, and it has been mainly quantified through measurement of pauses in keystroke logs (O'Brien, 2006b; Lacruz and Shreve, 2014) or gaze fixations using an eye-tracker (Vieira, 2014; Daems et al., 2017a; Moorkens, 2018). In general, research has shown that PE requires less temporal and technical effort than from-scratch translation and does not lower final-text quality (Plitt and Masselot, 2010; Green, Heer, and Manning, 2013). At the same time, results for cognitive effort are mixed and do not necessarily correspond to perceived effort (Koponen, 2012; Moorkens et al., 2015). In addition, some scholars have shown that using MT influences cognitive processes and can interfere with syntax and lexical choices (Elming, Winther Balling, and Carl, 2014; Čulo et al., 2014).

The advent of neural machine translation has opened the floor for new research. NMT is reported to "[have] the potential to change certain aspects of how post-editing tasks are carried out" (Nunes Vieira, 2020b, p. 325). Emerging issues include the increased difficulty in detecting errors in raw output, as shown in Yamada (2019), and a possibly higher cognitive effort (Jia, Carl, and Wang, 2019b). NMT has also played a pivotal role in expanding the adoption of MT within professional translation contexts and accelerating the evolution of MTPE as a practice. As a result, PE is currently "in a state of terminological flux" (Nunes Vieira, 2020b, p. 320) since it can be seen as encompassing various procedures, including the use of interactive MT and the use of MT suggestions as a source of inspiration. Nunes Vieira (2020b) proposes categorising different types of PE based on the degree of agency [14] granted to the post-editor (Figure 1.1).

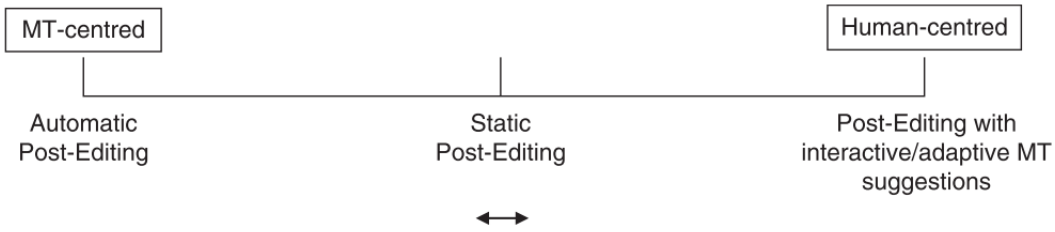



FIGURE 1.1 – Spectrum of human agency in the post-editing process; retrieved from Nunes Vieira (2020b, p. 328).

14. Kinnunen and Koskinen (2010, p. 6) define the concept of agency as the "willingness and ability to act".

On one hand of the continuum, the author places automatic PE (APE, do Carmo et al., 2021), a process carried out without any human intervention, which can be thus categorised as an MT-centred activity. On the opposite end of the spectrum, we find interactive PE (Sanchis-Trilles et al., 2014; Knowles and Koehn, 2016; Knowles, Sanchez-Torron, and Koehn, 2019), where the MT engine adjusts on the fly to the sentence typed by the human, in a configuration that portrays MTPE as a human-centred activity. In between these extremes, static PE involves human editing of a fixed MT suggestion. Within this framework, the use of MT as a source of inspiration can be placed in the space between static PE and interactive PE, as an activity that grants a higher degree of agency to the human, compared to, e.g., monolingual PE [15]. This thesis primarily focuses on static, bilingual PE performed by professional in-house translators. Automatic PE and interactive PE are excluded from the scope of our analysis.

### 1.1.3 Revision

As established by the ISO 17100:2015 process standard, which provides requirements for translation services, revision is a mandatory step in the translation process. It is defined as "bilingual examination of target language content against source language content for its suitability for the agreed purpose" (International Organization for Standardization, 2015, p. 2) made by "the reviser, who shall be a person other than the translator" (*ibid.*, p.10). The task itself consists in verifying the same aspects listed for translation, such as semantic accuracy, appropriate syntax, lexical cohesion and phraseology. Revision is deemed "essential[,] because every translator makes mistakes" (Mossop, 2020, p. xii), and its purpose is mainly to detect errors and suggest corrections to the translators, depending on the setting in which revision is carried out (*ibid.*). As Mossop (2020, p. xiii) points out:

> Reviser' is not the name of a profession; the activity or function of translation revision has developed historically as part of the profession of translator, though some translators may spend much or even all of their time revising. The relationship between writer and editor is therefore different from the relationship between translator and reviser, which might be better described as a relationship between the drafting translator and the revising translator.

Many scholars and practitioners have lamented terminological confusion around terms associated with revision in Translation Studies (Parra Galiano, 2005; Martin, 2007; Schjoldager, Rasmussen, and Thomsen, 2008; Hernández Morin, 2009b; Künzli, 2014; Scocchera, 2015; Robert et al., 2017; Mossop, 2020). As Robert et al. (2017, p. 3) note, the lack of consensus in revision-related terminology is mainly due to the fact that (i) Translation Studies is "a relatively young discipline" and (ii) the concept of revision is also used in related fields, such

15. Nunes Vieira (2020b, pp. 328–329) places monolingual PE and PE carried out by non-professional translators on the left-hand side of the spectrum, thus shifting the focus from the degree of human agency to the profile of the post-editor. We hold a different perspective and prefer to describe PE as an activity influenced by the dynamics of interaction between humans and MT, as well as by the contextual framework within which it takes place.

as Writing Studies. To begin, we should differentiate between *self-revision*, i.e., the examination of one's own translation, and *other-revision*, the examination of a translated text produced by someone else [16]. Other terms have been used to designate other-revision, such as *rereading*, *checking*, *proofreading*, *reviewing*, and *editing*. However, Scocchera (2017) notes that, for revisers, rereading is not synonymous with revision since they indeed read the translation for the first time (in this sense, rereading would be more a synonym for self-revision). Checking and proofreading involve the examination of the target text only (International Organization for Standardization, 2015): the former is carried out by a translator, while the latter is carried out on the revised text (International Organization for Standardization, 2015, p. 3) and often by a non-translator (Mossop, 2020, p. 249). Lastly, reviewing is often used to designate the revision by a subject-matter expert, while editing [17] usually refers to the revision of texts that are not translations, although this terminology is far from being used consistently (see, for instance, in Bisiada, 2018a).

Revision has a *business function*, as it is intended to "[prepare] the text for delivery to the client, and perhaps writing performance appraisals for the personnel department" (Mossop, 2020, p. 198). It also has a *training function*, since it aims at helping students and (novice) translators improve their work. These two functions are often combined in business settings (*ibid.*, p.202). The reviser's job consists in reading the translation and spotting issues, following the so-called *revision parameters*. Horguelin and Brunette (1998) consider five of these: accuracy, correct usage, readability, functional adaptation and profitability. The latter assesses the feasibility of revision, i.e., whether revising is more cost- and time-effective than re-translating the text. Horváth (2009) proposes six parameters, namely editing, equivalence, spelling, style, terminology and word order. The most complete categorisation is the one by Mossop (2020), who defines five groups of parameters dealing with transfer, content, language, presentation and specification issues. Each group includes specific parameters (12 in total), such as accuracy, completeness and layout.

Different *revision degrees* can be defined depending on how many parameters the reviser has to check in the time allocated to the revision task. Robert (2013, p. 94) differentiates between *full*, *loyal*, *functional* and *minimal* revision. The first degree takes into account all quality parameters, while a loyal revision focuses on accuracy. Functional revision checks "linguistic coding and appropriateness", whereas minimal revision takes into account "only some accuracy and linguistic coding items" (*ibid.*). Similarly, Parra Galiano (2005) defines four revision degrees, which depend on translation QA procedures, how much of the target text is revised, and whether the target text is compared to the source text. Mossop (2020) proposes four degrees of revision based on accuracy and writing quality requirements, namely intelligible, informative, publishable, and polished. Additionally, revision can be applied to the whole text (*full revision*) or just to some parts of it (*partial revision* or *partial check*).

16. While this distinction is nowadays widely accepted in Translation Studies and translation practice alike, it is worth noting that in Translation Process Research, self-revision is simply referred to as "revision phase", contributing to the terminological confusion mentioned earlier.

17. Further information on this term will be provided in Chapter 6.

The decision to apply a specific degree should be explained in the revision brief and based on pre-defined criteria. For instance, Prioux and Rochard (2007) set up a model to decide whether a second translator should revise a text, and which translator should carry out the task based on their experience.

Research on translation revision does not have a long history. The first academic publication was a PhD thesis (Brunette, 1995), but research on the topic has been growing since (as shown by Robert, 2018). Scholars have investigated the process and product of revision, both in theoretical and conceptual articles and in observational and experimental studies, using qualitative and quantitative methods (Künzli, 2005; Lee, 2006; Martin, 2007; Künzli, 2007; Laflamme, 2009; Bertaccini and Di Nisio, 2011; Robert and Van Waes, 2014; Chakhachiro, 2015; Robin, 2018, 2019). These studies have focused on the various aspects of revision mentioned so far, but also on revision teaching (Künzli, 2006; Schjoldager, Rasmussen, and Thomsen, 2008; Shreve, Angelone, and Lacruz, 2014; Robert, 2016; Hagemann, 2019; Scocchera, 2020), revision competence (Horváth, 2009; Terryn et al., 2017; Robert et al., 2017; Liang, Li, and Sang, 2023), revision practices (Hernández Morin, 2009a; Rasmussen and Schjoldager, 2011), and the profile of revisers or their expertise (Rensburg, 2017; Scocchera, 2017; Schaeffer et al., 2019). Additionally, a few researchers have focused on revision procedures (Brunette, Gagnon, and Hine, 2005; Robert, 2012; Robert and Van Waes, 2014) and on the reading order of the texts during bilingual revision (Künzli, 2009; Ipsen and Dam, 2016; Volkart et al., 2022; Riondel, 2022).

Across the various chapters of the thesis, we will look at some of these studies in greater depth. In the next section, we will start by exploring existing studies that considered PE and revision in the same investigation. It is important to clarify that the present thesis focuses on revision as defined by the ISO standard 17100 (2015) – therefore, as *other-revision*. Whenever other types of revision (e.g., *self-revision*) are considered, we will explicitly state so.

### 1.1.4 Previous comparisons of PE and revision

Comparisons between PE and revision tasks in different settings have been scarce. While initial contributions were made in the 1980s, the topic remained overlooked until NMT emerged. This section will mainly focus on theoretical comparisons between PE and revision tasks, providing an overview of the existing literature. Since empirical studies involving PE and revision will be thoroughly discussed in each chapter, in this section we will only point out which aspects were already investigated through an empirical approach and how our research will contribute to the current state of the art.

#### Theoretical comparisons

Wagner (1985) was one of the first authors to address the differences between translation, revision and PE. She was a reviser at the English Translation Division of the European Commission, where translators were using Systran as MT system. The author highlighted the translator's "freedom in selecting the words" in contrast to the post-editor's "restricted

choice", as post-editing involves a pre-translated text. The task was therefore deemed more similar to the revision of human translations, in the sense that "[working] by correction rather than creation comes as more of a shock to translators than to revisers" (Wagner, 1985, pp. 1–2). At the same time, the main difference with revision lay "in the intellectual level of the text [the post-editor] has to correct", since the output of the machine contained "errors which no human, even a small child or a nonnative speaker, would ever make" (*ibid.*, p. 2).

The author also reported on Commission translators' and revisers' reactions to the PE task, talking of an evolution in three stages. The first one is the "anthropomorphic stage", where translators and revisers find raw MT errors irritating or amusing, but at the same time "enjoy the chance to wield the red pen" on a badly translated text. The second is called the "objective stage", when the linguist finally understands why post-editing differs from translation and revision. Here, the Wagner makes a comparison between rapid PE and playing Scrabble: the post-editor has to "rearrange" words produced by Systran to obtain a comprehensible translation, while using as much raw output as possible. The last one is the "discerning stage", where the post-editor needs to decide whether MT is helpful, depending on the quality of the raw output and the final text requirements (full or rapid PE).

From the end-user's point of view, content that had been fully post-edited *and* revised was "indistinguishable from human translation" (*ibid.*, p. 4). However, the author reported that revisers working on fully post-edited content – which should already be indistinguishable from human translation – had realised that translators sometimes forgot to correct MT errors (mostly terminological errors) because they trusted the system too much.

While Wagner's considerations on the differences between post-editing and revision are summarised in one paragraph, Vasconcellos (1987) tackles one of the first thorough comparisons of MT post-editing and translation revision. Drawing on her experience as a translator, reviser and researcher at the Translation service for the Pan American Health Organization, the author stated that PE is a "highly specialized activity" that is best suited to revisers or experienced translators with good keyboard skills. According to Vasconcellos, both tasks involved the correction of errors, but there were "differences in the types of errors and in the approach to dealing with them [...]" (*ibid.*, p. 409). The aim of translation revision was "to catch and correct any errors that the first-stage translator may have made and, as required, to 'polish' the text so that it is appropriate to the setting in which it is to be used" (*ibid.*, p. 410). Revision was seen as "a discovery process" of errors made by translators, and revisers (Vasconcellos uses the term *reviewers*) were expected to work "two or three times as fast as the translator" (*ibid.*).

The errors that a reviser could encounter were mostly omissions, repetitions or misspellings, but also errors in formatting, punctuation or lack of consistency in terminology. Thus, formal corrections were the most frequent ones, while substantive errors were "rather infrequent" and often due to issues in the source text. A reviser had to ensure consistency and cohesion of the text, which could not result in "a congeries of isolated phrases and sentences" (*ibid.*, p. 411).

While revision was considered an [exciting adventure] because errors to find were "creative and unpredictable", PE was seen as "an ongoing exercise of adjusting relatively predictable difficulties". In post-editing, "[the] passages that clearly require corrections, though many of them are minor and local, [were] more frequent than in traditional revision" (*ibid.*). Vasconcellos reported on a study by Cressey (1987) on the types of corrections made during post-editing. Although the study did not deal with revision, assumptions were made about the fact that the distribution of errors and their proportion (particularly for word-order changes) in PE was higher than in revision. She stressed the fact that errors in "syntactic category, relative clause, number and agreement" would have been "rare or non-existent" in a human translation, while lexical substitutions remained the most frequent correction in both revision and post-editing (Vasconcellos, 1987, p. 414).

When reporting positive aspects of the machine translation system, Vasconcellos stated that "[one] can be confident that nothing has been skipped or repeated" (p. 411). Omissions and spelling errors were non-existent, unless there were errors in the dictionary. The correction of "unfortunate lexical choices", mainly due to the polysemous and ambiguous nature of natural language, occurred via changes in the dictionary (the author reports on several examples in the translation between Spanish and English). In PE, mistranslations were mostly local, and most MT errors were predictable. In McElhaney and Vasconcellos (1988), it was also argued that the system would have never made "wild guesses". Therefore, there was no danger of creative mistranslations. Post-editors had to solve MT issues with their own extralinguistic knowledge, which the machine lacked.

Post-editing and revision became more similar when the final text was to be published (i.e., in full PE), and thus, the responsibility of the linguist accrued. This interesting point was also raised by Wagner (1985) and is still very relevant (see Canfora and Ottmann, 2020). However, while Wagner speculated on the need to shift the responsibility from the linguist to the machine, Vasconcellos claimed that post-editors had "the same level of responsibility" as revisers and had, therefore, to apply the same criteria as revisers.

According to Vasconcellos, another difference between the two tasks was the medium, since post-editors worked on screen while revisers usually worked on paper. Post-editors were more concerned with "the need of saving time and keystrokes", while reviewers were not. At the same time, novice post-editors were usually prone to make unnecessary changes to valid raw output and had to be revised. This latter point – Vasconcellos claimed – "neutralizes the advantage of [using] machine translation" (*ibid.*, p. 414). Nowadays, some of these statements are not valid anymore. Research shows that NMT systems make omissions and quite unpredictable mistranslations (Castilho et al., 2018; Moorkens, 2018; Álvarez-Vidal, Oliver, and Badia, 2021), and the post-editor cannot intervene in a dictionary to alter the system's output. Furthermore, revisers work mainly on screen, as CAT tools are widely used to streamline the translation workflow (Robert, 2008; Scocchera, 2017).

Other authors have occasionally spent a few words on the parallel between PE and revision, mostly focusing on their differences. Laurian (1984, p. 237) maintained that PE was a different activity and not comparable to revision. It was "a new way of working on [a text], for a new aim". Similarly, McElhaney and Vasconcellos (1988, p. 141) argued that

"[t]he job of the post-editor is to do the interpretation, not to correct the interpretation of a colleague". In the same vein, Somers (1997) highlighted the sociological aspect of the difference between the tasks, namely that it was easier to correct MT output than colleagues' hard work – a point raised by Krings (2001) as well. Koby (2001) stated that translators and revisers shared the same cultural background and extra-textual knowledge, which was not the case with MT. Post-editors had to correct repetitive errors and "be aware of the quality expectations of the end user, which may be lower than those of the translator" (*ibid.*, p. 7).

More recently, Jakobsen (2019) examined the blurring boundaries between translation, PE and revision, acknowledging that one thing that differentiates PE from revision is that the machine does not have the ability "to construct meaning from [words]", as humans do (*ibid.*, p. 76). Do Carmo and Moorkens (2021) have published another interesting theoretical contribution on the differences between post-editing and revision. The authors maintain that nowadays, MT should be considered one of the resources contributing to the translation decision process, just like TM matches. Therefore, post-editing should be viewed as more similar to *translating* than revising. According to the authors, two strong arguments support the traditional view that PE is similar to revision:

— firstly, some studies have demonstrated that both PE and revision entail more reading than writing (Koehn, 2009; Ortiz-Martínez et al., 2016), while translation is essentially a writing activity, "a process of creation" (do Carmo and Moorkens, 2021, p. 36); and

— secondly, PE occurs after an initial translation step (performed by MT).

The authors argue that the second point, in particular, constitutes the notion behind the views of PE as a task that can be performed without extensive language skills and even without access to the source text (Krings, 2001; Schwartz, 2014). Additionally, PE assignments often receive lower compensation than translation assignments. For example, Lommel (2018) reports 60%-65% of the full word (translation) rate. This is a direct consequence of the view that associates the PE task with the revision task, which is paid less than translation tasks.

In their article, however, do Carmo and Moorkens list a number of arguments in favour of viewing PE as a form of translation that outperform arguments that identify PE as a form of revision. These include the following:

— PE jobs often imply more translating than revising (translators themselves recognise when a revision job hides raw MT);

— MT does not produce a finalised translation, but rather "a set of 'suggestions' or 'hypotheses' for the translation" (*ibid.*, p. 40);

— The post-editor is responsible for the final translation, and PE does not consist solely in looking for any errors in the text;

— It does not make sense to consider *translation* the editing of TM matches, and *revision* the editing of MT, as previous research has found these processes to be quite similar (see Sánchez-Gijón, Moorkens, and Way, 2019);

— Lastly, PE texts are often revised in commercial contexts, and since the industry is not inclined to redundancy, this is another sign that PE is seen as a form of translation.

The authors conclude (do Carmo and Moorkens, 2021, p. 42):

> Taking into consideration all that happens at professional translators' workbenches during PE, we propose that it should be considered a type of translation. Not only because PE represents an evolution of industrial translation processes and because it fulfils the same purpose as translation (to produce a good target text in an efficient and effective way), but also because it requires advanced writing and reading skills in two different languages.

We argue, however, that the "advanced writing and reading skills" put forward by these authors also pertain to professional revisers. We only partly agree with their view: the authors talk of PE and revision in general terms without contextualising their claims and without acknowledging the existence of contexts where PE and revision may appear more similar than in others – one such case is when the two tasks are both the last step in the workflow. Additionally, they state that post-edited texts are also revised in many contexts. However, their claim does not appear to be founded on empirical data.

At this point, it is also worth mentioning the work of Konttinen, Salmi, and Koponen (2021), who do not share the same view as do Carmo and Moorkens (2021) since they consider post-editing as a QA activity, just like revision. Their suggestion is to include theoretical content for revision in translation courses, while this same type of content for PE could be dealt with in translation technology courses. Indeed, the authors state (*ibid.*, p. 202):

> For curriculum planning, a crucial difference between training in revision and training in post-editing lies in how to teach the students to communicate any requirements and suggestions for changes in human translations, on the one hand, and in MT and TM output, on the other.

To establish "an efficient and pedagogically effective model" for translator education (Konttinen, Salmi, and Koponen, 2021, p. 188), the authors review existing models of revision and PE competence, finding that they both include several common sub-competences, which can be grouped into three main categories (*ibid.*, p. 194):

— *Strategic sub-competences*, i.e., detecting, identifying and evaluating errors, plus information seeking;

— *Interpersonal, psycho-physiological or attitudinal sub-competences*, i.e. adopting a "revising" frame of mind as opposed to "re-translating";

— *Instrumental sub-competences*, i.e., using appropriate tools (MT and CAT tools in particular).

Additionally, the authors cite previous studies that pointed out a different perception of revisers and post-editors' backgrounds: in these studies, revision is seen as a task more suited to experienced translators, while PE is sometimes considered a task for novice translators or even subject-matter experts without experience in translation. This happens because PE is often associated with light PE, but the wide adoption of NMT challenges this view (*ibid.*,

p. 195). Indeed, in professional contexts, PE tasks are assigned to novice *and* experienced translators.

Lastly, Robert, Ureel, and Schrijver (2022) analysed existing translation, revision and PE competence models, supporting the hypothesis that these are different, but share a common core. The authors concluded by stating that it remains to be verified, through further empirical research, the *extent* to which the three competences differ.

#### Empirical comparisons

Experimental studies involving PE and revision tasks together emerged in the late 2010s. One of the first studies of this kind was conducted by Temizöz (2016, 2017). It aimed to compare two workflows in technical translation involving PE of SMT content by professional translators or subject-matter experts (engineers) followed by a revision step carried out by the opposite cohort.

In the field of Translation Process Research, Huang (2016, 2018) investigated the working style of student translators in self-revision, other-revision and PE. The study defines "working style" as "the coordination of physical and mental activities" (Huang, 2018, p. 146), intended as sequential reading and typing activities and their underlying purposes. Four distinct working styles were identified triangulating data from eye-tracking, keylogging and cue-based retrospection. Additionally, the study revealed that students primarily focused on checking accuracy rather than fluency, in all tasks. However, methodological concerns could be raised since it was reported that participants did not perform full PE, but rather re-translated the source texts in most cases. The researcher attributed this misunderstanding to participants' lack of awareness regarding the differences between translation and PE. We could argue that some PE knowledge and a better-quality NMT output could have led to more similar results in the PE and other-revision tasks.

With the emergence of NMT and its integration into professional workflows, empirical studies on PE and revision have multiplied. Vardaro, Schaeffer, and Hansen-Schirra (2019a,b) conducted a corpus-based study of translations produced in an institutional context, comparing SMT and NMT raw output, post-edited segments and their final, revised version. Daems and Macken (2021) investigated professional translators' trust towards MT and human translation, examining the influence of assumed production method (PE or revising) on edit rate, revision quality and intervention optimality. Macken et al. (2022) analysed modifications made to the texts in a professional workflow including PE of a literary text, followed by revision. A similar study was conducted by Casas (2021), who compared the performance of MA students in revision and PE tasks with that of a professional translator, across different text types. Finally, Robert, Schrijver, and Ureel (2023) empirically validated common sub-competences between revision and PE. In a small-scale exploratory study conducted in the context of a translation course, the authors investigated the extent to which revision and PE share certain sub-competences. The students' performance in translating, revising and post-editing into their L2 was captured using a keylogging software, and the products of their tasks were analysed. The findings revealed that problem

detection is a shared sub-competence between revision and PE, but participants performed better in revision than in PE. We argue, however, that the quality of pre-translated texts in PE and revision tasks may have influenced the results, as previously shown in Temizöz (2017). In a companion paper (Robert, Schrijver, and Ureel, 2022), the authors showed the importance of carefully choosing the method to assess revision competence, since different measuring instruments may lead to different results.

In synthesis, theoretical contributions that compare the PE and revision activities in a professional setting are either outdated (as in Wagner, 1985 and Vasconcellos, 1987), or do not delimit the scope and the context in which the comparison is conducted (Jakobsen, 2019; do Carmo and Moorkens, 2021). At the same time, none of the empirical studies considering PE and revision has compared the two tasks in a professional setting. In these studies, PE and revision have been either considered two subsequent steps in the same workflow (as in Temizöz, 2017; Vardaro, Schaeffer, and Hansen-Schirra, 2019b; and Macken et al., 2022), or experiments have been carried out with students (as in Huang, 2018; Casas, 2021 and in Robert, Schrijver, and Ureel, 2023).

## 1.2 Motivation

In addition to the research gaps discussed so far, a need for more empirical research on the comparison between PE and revision tasks has also been put forward by translation scholars and practitioners. For instance, Brunette and O'Brien (2011, p. 4) expressed concern over the limited availability of data on the differences between these two tasks:

> To those who want to know the differences between post-editing and revision, we can say nothing other than: 1° the errors to be corrected are not the same; 2° the resulting text can reproduce the formal structure of the [source text] in the case of PE; 3° PE brings those who do it into a world of extreme productivity unfamiliar to traditional translators. This is not our point, but this kind of explanation does nothing to advance science. [*Our translation*] [18]

More than a decade later, despite the emergence of a few additional studies in recent years, there remains a notable lack of knowledge regarding the characteristics that differentiate PE and revision. Relevant aspects for professional practice, such as the strategies employed and the individual satisfaction experienced when performing these tasks, have never been explored. Koponen et al. (2021b) have highlighted that with the advent of NMT, the boundaries between revision and PE have become increasingly blurred. Recognising the importance of investigating the implications of this phenomenon, Zapata and Polikar (2023) explicitly called for further research on the subject:

18. *À ceux qui veulent savoir ce qui différencie postédition et révision, on ne sait rien dire d'autre que : 1° les erreurs à corriger ne sont pas les mêmes ; 2° le texte résultant peut reproduire la structure formelle du [texte de départ] dans le cas de la PE ; 3° la PE transporte ceux qui la font dans un univers de productivité extrême étranger aux traducteurs traditionnels. Ce n'est pas notre propos, mais ce genre d'explications ne fait guère avancer la science.*

> It is important to continue to distinguish between revision and [PE], and to reflect this distinction in translation practice and teaching. Advanced scientific research into this issue is highly desirable. The various methods of collecting empirical data in research into translation processes (recording typing and on-screen activities, eye-tracking, interviews and questionnaires, among others) would make it possible to highlight the actions of editing and text production, information retrieval and translation from scratch that are carried out in the context of post-editing, by comparing them with the actions of revision. These studies would also make it possible to measure productivity gains, the quality of texts (post-edited and then revised), and the satisfaction and well-being of members of the language professions. Such research could lead us to see post-editing as an effective translation strategy requiring specialised skills – and not to confuse it with revision, which remains [indispensable]. [*Our translation*] [19]

Our work directly responds to the call for more empirical studies shedding light on the similarities and differences between PE of NMT content and translation revision. We conducted our research in a well-defined professional context, touching upon multiple aspects of the PE and revision activities and the relationship between the two tasks. In the upcoming sections, we will delve further into the relevance of this topic and the rationale behind the choice of our particular context of investigation.

### 1.2.1 Working with pre-translated texts

As previously discussed (Section 1.1.2), contemporary translators rarely begin their work from scratch when translating written texts. During translation and PE tasks, translators heavily rely on suggestions provided by TMs and MT systems, thus working on a text that has been *pre-translated* to some extent.

As opposed to *interactive* PE, where the system displays new suggestions as the post-editor modifies the existing text, *static* PE consists of reading and modifying a fixed MT suggestion (Nunes Vieira, 2020b). When MT is applied to the whole text using a batch pre-translate function, we get a fully pre-translated text that the post-editor has to check for accuracy and fluency. It could be argued that this is the same situation as that experienced by a reviser, except that for the latter, the pre-translated text is produced by a human.

Considering the current situation, we contend that the interest in an investigation of PE and revision, as two activities performed on pre-translated texts, has multiple justifications:

— For didactic purposes, it is crucial to identify and be able to describe the fundamental differences between the two activities. Students must be aware of these distinctions to

19. *Il s'avère important de continuer de distinguer la révision de la postédition, ainsi que de refléter cette distinction tant dans la pratique que dans l'enseignement de la traduction. Des recherches scientifiques poussées qui se penchent sur la question sont plus que souhaitables. Les différentes méthodes de récolte de données empiriques de la recherche en processus traductionnels (l'enregistrement de la frappe et des activités à l'écran, les mesures oculométriques, les entretiens et les questionnaires, entre autres) permettraient de mettre en évidence les actions d'édition et de production textuelle, de recherche informationnelle et de traduction à partir de zéro qui sont effectuées dans le contexte de la postédition en les comparant avec les gestes de la révision. Ces études permettraient aussi de mesurer les gains de productivité, la qualité des textes (postédités, puis révisés), ainsi que la satisfaction et le bienêtre des membres des professions langagières. De telles recherches pourraient nous amener à voir la postédition comme une stratégie de traduction efficace nécessitant des compétences spécialisées – et à ne pas la confondre avec la révision, qui demeure primordiale et indispensable.*

approach each task appropriately. Additionally, providing insights from professional practice can better prepare translation trainees for the evolving market, enabling them to navigate its dynamics more effectively.

— Research in MT can benefit from a deeper understanding of the nature of the errors made by NMT systems as opposed to those made by humans. Insights into error patterns can inform the improvement of NMT systems.

— For management purposes, it is helpful to understand how professionals devise their strategies when performing PE or revision. It allows for the formulation of best practices and could guide decision-making. Furthermore, understanding the factors that contribute to job satisfaction in performing PE and revision tasks can lead to the design of more rewarding workflows and foster a positive work environment. This knowledge can ultimately enhance the overall quality and efficiency of translation processes.

Throughout this thesis, the term *pre-translated text* will be used to refer to two distinct concepts: the raw output of machine translation (referred to as *machine-pre-translated text*) and the draft translation produced by a human translator before the revision process (referred to as *human-pre-translated text*). We acknowledge that some researchers and practitioners may raise concerns about the term "pre-translation", since it could be argued that human translations do not have the same status of translation hypotheses that do Carmo and Moorkens (2021) assign to raw MT output but are finalised translations. However, we deem it useful to introduce this terminology to clarify the differences between texts at various stages of the translation process.

### 1.2.2 Swiss corporate in-house language services

Due to its multilingual nature, Switzerland stands out as a country where translation holds significant importance. Translators find employment both in-house and as freelancers within various institutional settings. These include government departments at the national and cantonal levels, as well as bilingual city administrations. Numerous non-governmental and international organisations, such as the United Nations, primarily concentrated in the Geneva area [20], own in-house translation and interpreting departments. Furthermore, several Swiss-based companies have their own corporate in-house language services to ensure effective communication within the different language-speaking regions of Switzerland. In the present thesis, we focus precisely on this cohort that is underrepresented in language industry surveys and has received limited attention compared to institutional (Cadwell et al., 2016; Rossi and Chevrot, 2019; Riondel, 2021a) and freelance translators (Gaspari, Almaghout, and Doherty, 2015; Zaretskaya, Pastor, and Seghiri, 2015; Piecychna, 2019).

While the term *corporate* denotes the nature of the language service provider (i.e., pertaining to an enterprise), we explicitly mention that this service is *in-house* to indicate that

20. Geneva is also home to one of the oldest translation and interpreting schools, now a faculty, originally established to cater to the needs of these organisations `https://www.unige.ch/fti/fr/`.

the language provider employs at least one translator in-house, as opposed to companies with an internal communication function that regularly entrust translations to an external language service provider. Our emphasis on CILS translators is particularly justified in the context of exploring the use of MT by language professionals. Unlike freelance translators, who often express dissatisfaction with the low rates offered for PE tasks (Álvarez-Vidal, Oliver, and Badia, 2020; Pérez Macías, 2020; Nunes Vieira, 2018; SFT, 2022), CILS translators are salaried employees, and their compensation remains unaffected by the use of MT. This allows them to assess the advantages and disadvantages of MT without external factors affecting their judgment.

Additionally, in-house translators fully benefit from TM suggestions, which facilitate consistency in sentence structures and terminology and improve efficiency. In commercial settings, TMs are considered valuable assets contributing to cost control. Indeed, many LSPs offer discounts to clients for sentences retrieved from the TM, which can impact translators' remuneration. In contrast, such considerations do not apply to in-house translation departments.

Another significant aspect that justifies the focus on CILS translators compared to institutional and agency translators is their emphasis on *productivity*. Unlike commercial language service providers, such as agencies, whose primary business revolves around translation and related services, CILS represent specialised departments within companies with a distinct core business. Establishing an in-house language service constitutes a noteworthy cost for the company, often rendering outsourcing to external LSPs a more favourable option. Consequently, the imperative for CILS to consistently validate their relevance is heightened as they navigate corporate dynamics that often prioritise efficiency, dismissing unnecessary or low-performing functions. Given these circumstances, the ability of CILS to operate within stringent time constraints becomes paramount and MT is primarily used to expedite processes.

This perspective, rooted in the pressure to meet corporate demands, aligns closely with prevailing market realities, and differs from the conception of translation in institutional and governmental translation departments, where MT is often offered to linguists as an additional tool. Within these settings, translation is above the mere commercial interest and serves a dual purpose – not only a communicative service but also a mechanism for promoting social equity and access to information. While pressure on institutional translators to enhance their productivity is present (as reported in Cadwell, O'Brien, and Teixeira, 2017), we argue that it could be less acutely felt than in CILS contexts.

## 1.3 Thesis goals, research questions and methods

This section provides an overview of the methodological approach adopted for our research project [21]. The broader objective of this thesis is to offer a comprehensive description of the

21. The methodological details of each phase will be discussed in individual chapters.

similarities and differences between post-editing of NMT output and translation revision, from the perspective of in-house professional translators (hereafter *linguists*[22]).
Two specific goals define the scope of our research:

> **Goal 1** – To gather information on the use of MT in Swiss CILS and on how PE and revision tasks are organised in CILS' workflows.
>
> **Goal 2** – To understand how PE and revision tasks compare in the professional context of Swiss CILS from the perspective of in-house linguists.

The goals are interconnected and form a cohesive framework. Goal 1 defines the context of our study, profiling our cohort and allowing us to identify eligible participants for the subsequent steps of our work, which is then guided by Goal 2. To attain Goal 2, the tasks of PE and revision will be compared by looking into the following key aspects:

— Linguists' **PE and revision practices**

— Linguists' **satisfaction** in performing PE and revision tasks

— Linguists' **productivity** in performing PE and revision tasks

— **Modifications (*edits*)** made by linguists to pre-translated texts in PE and revision tasks

Specifically, we will try to answer the following research questions (RQs)[23]:

**RQ1**: *In the professional workflow of Swiss CILS, how do linguists' PE and revision practices differ?*

**RQ2**: *In the professional workflow of Swiss CILS, how do PE and revision compare in terms of task satisfaction for the linguist who performs these tasks?*

**RQ3**: *When PE and revision tasks are carried out under the same conditions, how does linguists' productivity compare in these tasks?*

**RQ4**: *When PE and revision tasks are carried out under the same conditions, how do PE and revision compare in terms of edits made to pre-translated texts?*

As previously mentioned (Sections 1.1.2 and 1.1.3), in this thesis, we understand PE as the work on a machine pre-translated text (and we focus especially on NMT output), while revision is understood as a bilingual examination of human pre-translated texts, carried out by a different linguist than the one who translated the text in the first place. In two research questions (RQ3 and RQ4), we explicitly compare PE and revision "carried out under the same conditions". More specifically, we refer to static PE of fully machine-pre-translated texts, which involves both MT and TM suggestions.

Our research project subscribes to the relatively young field of *translation and interpreting workplace research*, which, according to Risku, Rogl, and Milošević (2020, p. 37):

22. We use *linguist* as a hypernym that encompasses various roles that a language professional can assume, including translator, post-editor and reviser.
23. Research sub-questions and related hypotheses will be presented in each chapter.

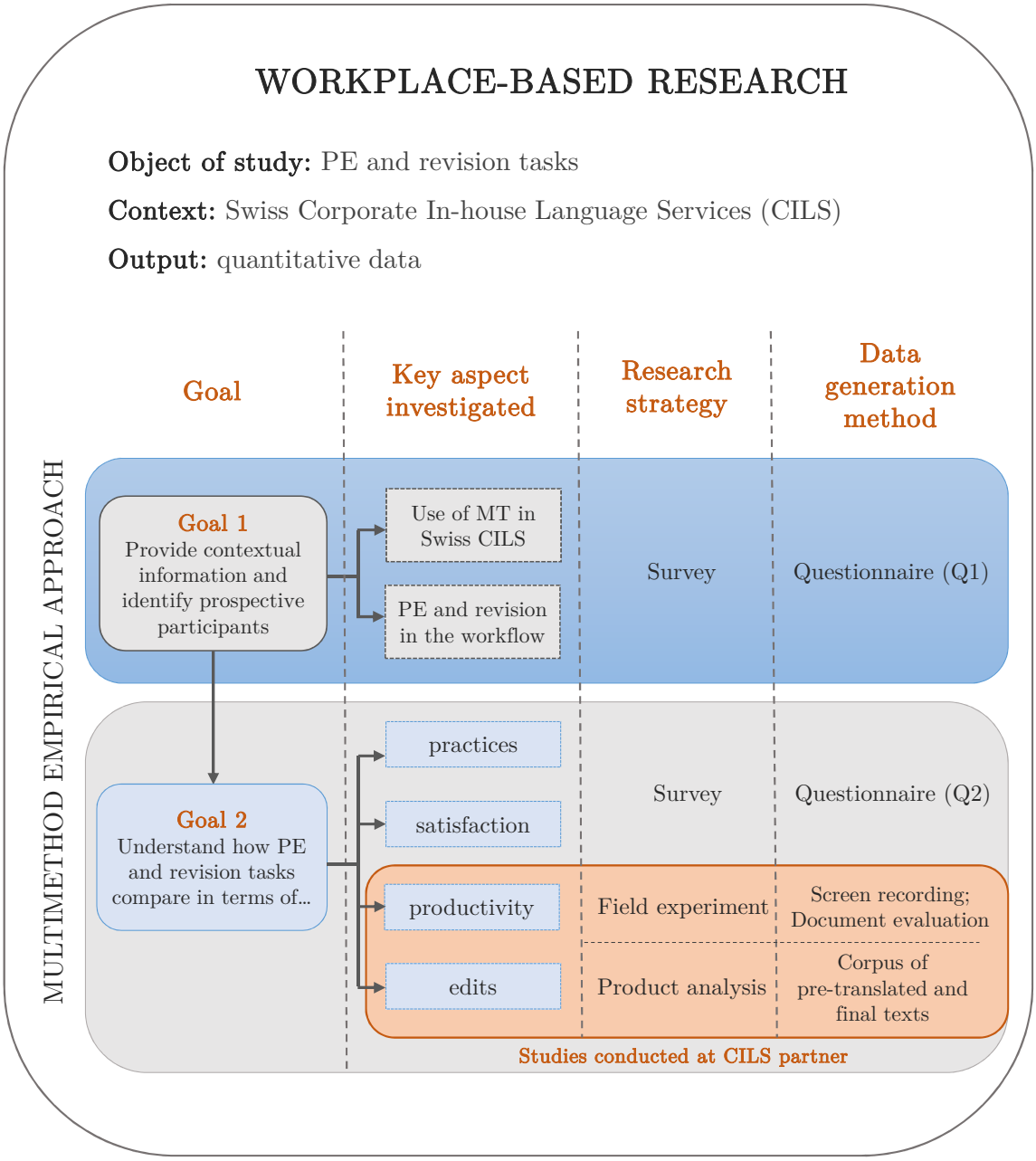



FIGURE 1.2 – Overview of the methodological approach adopted in the present thesis.

> takes the workplaces of translators, interpreters and other language service providers as analytical units and endeavours to retrace their work-related activities, interactions and working conditions as carried out and experienced in day-to-day professional practice. Current key areas in translation and interpreting workplace research include, for example, work organization and routines; cooperation and social dynamics; cognitive, organizational and physical ergonomics; and the use and implications of (collaborative) technology in the workplace.

To investigate the tasks of PE and revision in professional workplaces, we employed a *multimethod empirical approach*, combining various research strategies and quantitative data collection methods, as shown in Figure 1.2. Hunter and Brewer (2015, p. 187) defined multimethod research [24] as follows:

> [The] practice of employing two or more different methods or styles of research within the same study or research program [...]. Unlike mixed method research, it is not restricted to combining qualitative and quantitative methods but rather is open to the full variety of possible methodological combinations.

Mark (2015, p. 40) explains that researchers should employ multiple methods "in whatever way optimizes the quality of one's research, the strength of inference it allows, and the likelihood of informing and influencing the relevant audiences". Therefore, to attain Goal 2, our primary strategy has been to triangulate data from different independent sources to provide a more complete picture of the topic investigated. In particular, the present thesis is the product of three studies:

1. A survey-based study with a threefold aim, namely (i) to map the use of MT in Swiss CILS; (ii) to investigate PE and revision practices in Swiss CILS; and (iii) to investigate CILS linguists' satisfaction in performing PE and revision. The study involved two different questionnaires (Q1 and Q2, as shown in Figure 1.2). The first questionnaire was aimed at CILS directors and project managers, while the second questionnaire was sent to in-house linguists working at CILS that employ MT in their workflows.
2. A field experiment carried out in a Swiss CILS to measure linguists' productivity, defined in terms of temporal effort and adequacy of the final product, in PE and revision tasks.
3. A corpus-based study of authentic PE and revision assignments carried out in the same Swiss CILS to investigate the modifications applied to pre-translated texts.

Following a distinction introduced by Ehrensberger-Dow and Massey (2020), our research is both conducted *on* the workplace, i.e., investigating work-related activities of language professionals working in a well-defined professional context, and *in* the workplace, with two studies conducted in a specific CILS.

24. While 'mixed methods research' usually implies the use of quantitative and qualitative methods, 'multimethod research' involves mixing two or more methods, regardless of their nature (Hunter and Brewer, 2015).

## 1.4 Thesis structure

Since this thesis encompasses several studies, we have structured it in a way that facilitates reading and comprehension. Each chapter is dedicated to the investigation of a specific key aspect: it includes relevant previous work related to the aspect under investigation, as well as the rationale behind the study and the methodological framework adopted. Results are presented and discussed, together with the limitations of each specific research method. A summary of the key findings concludes each chapter.
After this introductory chapter, the thesis is structured as follows:

**Chapter 2** focuses on **Goal 1**: it presents a survey-based study that delves into the use of MT among Swiss CILS, as well as how PE and revision tasks are organised in CILS workflows. It begins by reviewing previous studies on the use of MT in professional contexts and then narrows down the focus to the Swiss professional translation landscape. The chapter outlines the structure of the questionnaire Q1 and the criteria used for its distribution. It proceeds to present the findings, which are then compared with similar studies conducted in other geographical areas.

**Chapter 3** explores the first key aspect defined under **Goal 2**, namely PE and revision strategies of in-house linguists working at CILS that make use of MT in their workflows. It begins by examining previous survey-based studies conducted among post-editors and revisers, providing a comprehensive overview of the existing literature in this area. The chapter then proceeds to outline the structure of the questionnaire used in the current study (questionnaire Q2, sections A to D). Subsequently, it presents and analyses the results obtained, drawing comparisons with the findings from previous research.

**Chapter 4** focuses on linguists' satisfaction in performing PE and revision tasks. It begins with a discussion of the concept of satisfaction in the field of industrial-organisational psychology, reviewing relevant studies and commonly used satisfaction questionnaires in this discipline. The chapter further explores relevant work on translators' satisfaction in Translation and Interpreting studies. It then explains how the instrument used to collect the data was designed (questionnaire 2, section E) and concludes with an analysis and discussion of the results in relation to previous studies.

**Chapter 5** deals with linguists' productivity in performing PE and revision tasks. It presents a small-scale exploratory study conducted at a CILS partner, as part of a project on MT integration in their workflow. The chapter starts by presenting this collaborative project. Then, the focus shifts to previous productivity studies in professional translation workflows. Afterwards, the rationale behind our experiment is introduced, and the methodology employed is detailed. Lastly, results are presented and discussed.

**Chapter 6** delves into the modifications made to pre-translated texts during PE and revision tasks. It reports on a corpus-based study of modifications made to texts during authentic revision and PE assignments carried out at the CILS partner. After reviewing relevant studies on editing in revision and PE, we explain how the corpora were compiled and which analyses were performed. The results of the study are then presented and thoroughly discussed.

**Chapter 7** serves as the final chapter, offering a summary and a discussion of the findings in relation to our research questions and hypotheses. In the same chapter, we address the broad limitations of our research and provide perspectives for future work.

# Chapter 2

# Swiss CILS and MT use

## 2.1 Overview

Despite its active translation market and diverse linguistic landscape, Switzerland remains often underrepresented in surveys of the translation industry. Furthermore, these surveys are most often conducted with freelance translators or translation agencies, while corporate in-house translation services are rarely included. In this chapter, we address this gap by presenting a survey-based study conducted among Swiss CILS at the end of 2021. The study portrays the profile of this specific cohort of respondents, delving into their workflows, investigating whether or not they use MT, and placing particular emphasis on how PE and revision tasks are organised. By addressing these aspects, the study sets the context for our research and let us identify prospective participants for the subsequent steps of our investigation (**Goal 1**).

The chapter is structured as follows: Section 2.2 provides an overview of previous language industry surveys examining professional practices and the integration of MT into translation workflows. It also includes a review of similar studies conducted in Switzerland. Section 2.3 details the structure of our own survey, the sampling method, and the profile of the participants. Results are presented in Section 2.4 and discussed in Section 2.5. Finally, Section 2.6 provides a summary of the chapter.

## 2.2 MT in professional workflows

Regular industry surveys play a crucial role in providing insights into the state of the language industry and the operational practices of LSPs. National professional associations often conduct surveys that offer a comprehensive view of the market size within a specific

geographical area. For example, surveys conducted by AITI [1] in Italy (2018), by CBTI-BKVT [2] in Belgium (2018), SFT [3] in France (2022) and by ATC [4] in the UK (2021) provide valuable information on the market size for those specific regions. However, these surveys mainly target independent language professionals, which means that data on larger LSPs and language departments are relatively scarce.

To gain a comprehensive understanding of the industry, global surveys conducted by research and consulting firms – one notable example is Common Sense Advisory (CSA) – are often employed. In 2019, CSA released a report on the use of MT among LSPs, based on data collected from 452 respondents across several countries around the world [5] (Pielmeier and Lommel, 2019). The report provided an insightful portrait of LSPs' adoption of MT, revealing that MT was almost always integrated into CAT tools, but it was also employed in less than 10% of projects. Neural MT emerged as the preferred architecture among respondents, outperforming other alternatives. Commercially available MT solutions were identified as the most common source of MT for the surveyed LSPs.

Surveys of different stakeholders are also organized and promoted by groups of international industry associations reunited. One such example is the annual European Language Industry Survey (ELIS) [6], which focuses on

> the industry trends, expectations and concerns of a wide range of language industry stakeholders, including language service companies, independent language professionals, training institutes with translation and interpreting-related training programmes, language departments in private companies, public administrations, and language service buyers

According to the latest ELIS survey results (2023) [7], while CAT tool are approaching their full market potential, MT has still room for growth. Indeed, less than 50% of language departments reported using MT, compared to approximately 75% of independent professionals and LSPs. However, if we focus only on *corporate* language departments, approximately 70% of the respondents reported using MT in their workflows. Detailed information about how this is done – such as the type of MT architecture used, integration of MT into CAT tools, and whether post-edited output undergoes additional revision – was not provided in

1. *Associazione Italiana Traduttori e Interpreti* (Italian Association for Translators and Interpreters).
2. *Chambre belge des traducteurs et interprètes/Belgische Kamer van Vertalers en Tolken* (Belgian Chamber of Translators and Interpreters).
3. *Société Française des Traducteurs* (French Society of Translators).
4. The acronym stands for Association of Translation Companies.
5. The exact number of countries is not provided.
6. Launched in 2013 by European Union of Associations of Translation Companies (EUATC), the annual ELIS survey is a joint initiative by international language industry associations and organisations. Apart from the EUATC, those participating are the European Language Industry Association (Elia), the European Master's in Translation (EMT), the European Regional Centre of the International Federation of Translators (FIT Europe), the Globalization and Localization Association (GALA), the European Commission's Language Industry Platform (LIND) and Women in Localization (WiL) network. Source: `https://elis-survey.org/About`. Accessed 24 May 2023.
7. The 2023 survey received 1164 responses: 636 from independent language professionals, 264 from language service companies, 175 from training institutions and students, 63 from language service buyers and private and public translation departments. Although the majority of respondents were located in Europe, there were also 46 responses from other geographical areas.

the survey results. Additionally, it is worth highlighting that corporate language departments in the ELIS survey were scarcely represented, with less than 30 participants (the exact figure was not provided). Furthermore, Switzerland's participation in the survey was marginal, with only ten responses coming from this country in the 2023 edition. Among these responses, there were three companies, two individual professionals, two training institutes, and two respondents categorised as "language departments or language services buyers".

The Swiss translation market has received limited attention in research settings as well, with only a few studies dedicated to exploring its dynamics. For example, in 2015, Porro Rodríguez, Morado Vázquez, and Bouillon (2017) conducted a survey specifically targeting Swiss-based LSPs – which, however, excluded corporate translation departments. The survey findings revealed that only two out of 16 LSPs surveyed were incorporating MTPE into their workflows. Additionally, most respondents declared they were either uncertain or definitely not interested in implementing MTPE in the future. A more recent study carried out by Selinger (2020) as part of her MA thesis examined the use and perceptions of MT among Swiss translation professionals and non-professionals (170 and 115 respondents, respectively). The results showed that the vast majority of non-professionals used MT as a starting point for translations into their mother tongue. However, they did not always check the MT output against the source text. To complement the questionnaire data, interviews were conducted with five LSPs that had either integrated MT into their workflows or were in the process of doing so. These LSPs respondents expressed some concerns regarding the confidentiality of data, leading them to use or test either a customised MT system or DeepL Pro [8], which offers enhanced data privacy features. Regarding the practical details of the MT-CAT tool implementation, participants clarified that MT suggestions were fully integrated into their CAT tools or made available to internal customers as a self-service translation tool.

Given the limited number of studies conducted on the Swiss translation market, further research is warranted to gain a more comprehensive understanding of its dynamics and practices, particularly regarding technology implementation. In this regard, CILS represent an ideal context of investigation (as previously discussed in Section 1.2).

## 2.3 A survey on the use of MT in Swiss CILS

Considering the aforementioned lack of available data on Swiss CILS, we decided to conduct a survey-based study to investigate the characteristics of this specific cohort. The study aimed to achieve multiple objectives. Firstly, it sought to quantify the number of CILS operating in Switzerland, providing valuable insights into the size of this market segment. Secondly, it examined the extent to which MT has been integrated into the workflows of Swiss CILS, exploring usage patterns and integration methods employed. Thirdly, the study

8. `https://www.deepl.com/en/pro`

aimed to gather insights into the organisation of PE and revision tasks within CILS workflows. Lastly, this survey allowed us to identify prospective participants for the subsequent phases of our research.

To ensure the ethical integrity of our work, the survey-based study received the necessary approval from the Ethical Review Board of the Faculty of Translation and Interpreting at the University of Geneva (reference number 32/2021, available in Appendix J).

### 2.3.1 Sampling

Due to the specific nature of our population, namely Switzerland-based companies with an internal translation department, we recruited potential participants through a non-probabilistic approach, combining various sampling techniques [9].

Firstly, we contacted via email (Appendix A) language service directors of corporate in-house services that we directly knew (*purposive sampling* [10]); we asked them to participate in the survey and to help us recruit further participants (*snowball sampling* [11]).

Secondly, we tried to identify CILS through the profile of their employees on social media. To do so, we used the research function on the popular business-related social networking platform LinkedIn [12], looking for terms such as *translator*, *language services*, *corporate* and *project manager*, restricting the research area to Switzerland [13]. Additionally, to enhance the visibility of our questionnaire, a call for participation was also advertised through a public post on the platform (*volunteer sampling*).

Thirdly, we compiled a list of private and semi-private companies serving the Swiss public at large, including banks, insurance companies, and retail outlets, among others. We discarded from this list all the companies whose website was not translated into a different language and then contacted prospective participants using their generic email address or through a contact form on their websites. In total, we reached out to 290 companies headquartered in Switzerland, but we only received an answer by 141 of them. Thirty out of 141 companies responded that they outsource all their translation-related activities to external language service providers (agencies or freelancers). Some of them also mentioned that translations are occasionally carried out by internal employees who are not professional translators. Twelve companies declared they were not interested in taking part in the study or could not do so for internal policy. Forty-eight companies answered they do not have any in-house translation department.

9. Although less ideal than probability sampling to ensure representativeness and generalisability of results, non-probability sampling techniques are particularly useful when the population is hard to reach or when "there is not a list or a central repository of people who meet [specific criteria] and could be respondents" (Lazar, Feng, and Hochheiser, 2017, p. 113).

10. In purposive sampling, the researcher "[chooses] instances that are likely to produce valuable data to meet the purpose of the research" (Oates, 2005b, p. 98).

11. Mellinger et al. (2017, p. 13) mention that snowball sampling is closely related to *network sampling*, a technique in which "[the] researcher uses an existing social network (real or virtual) to reach people who share a common trait".

12. `https://www.linkedin.com/`

13. This method can also be classified as purposive sampling.

A link to the online survey was sent by email to the language service directors or project managers who agreed to take part in the study. In the emails, it was specified that participation was voluntary and anonymous. This information was also clearly stated on the first page of the online questionnaire, which contained a consent form (Appendix B). The questionnaire was hosted on the LimeSurvey platform [14] and was made accessible from November 11th, 2021, until February 14th, 2022. Depending on respondents' answers, the questionnaire included up to 50 questions, but not all were mandatory.

### 2.3.2 Survey structure

The questionnaire was structured in five sections:

**Section A** [*Translation service*] contained general questions about the CILS, such as the number of in-house linguists, the CAT tool(s) used and whether the CILS used or did not use MT in its production workflow.

**Section B** [*Workflow*] focused on the workflow of those language services who worked with MT, e.g. which operations were carried out in-house or outsourced. This section and the following ones were accessible only to the respondents who used MT in production (n=26). However, not all the questions were mandatory.

**Section C** [*Machine translation and post-editing*] comprised questions about the type of MT system in use, how MT was integrated in the workflow and whether different PE levels and corresponding guidelines had been implemented.

**Section D** [*Revision*] focused on revision practices, e.g. which translations were revised and whether any revision guidelines had been set up.

**Section E** [*Post-editing and revision*] included questions on the relationship between PE and revision, for instance asking whether post-edited texts were also revised.

The complete version of this questionnaire can be found in the Appendix section (Appendix C).

### 2.3.3 Participants' profile

A total of 60 responses were initially collected from the survey. Responses that were incomplete or did not meet the target criteria were excluded from the analysis, resulting in a final dataset of 52 valid respondents.

Regarding languages handled, most CILS (39 out of 52) work from and into the three main languages of Switzerland, i.e. German, French and Italian, plus English. Although English is not one of the official languages of Switzerland, it is slightly more used than Italian. Indeed 47 out of 52 respondents translate from and into English, versus 41 out of

14. LimeSurvey is an online open-source survey tool that allows university students and staff to create and manage questionnaires, as well as to store questionnaire data in a secure environment. The platform is integrated into the server infrastructure of the University of Geneva and accessible through the following link: `https://formulaire.unige.ch/outils/limesurveyfac/traduction-interpretation/index.php/admin/authentication/sa/login`.

52 respondents who translate from and into Italian. Fourteen respondents report they also translate into or from other European languages – such as Spanish – but also Romansh, Arabic, Chinese and Japanese. The most frequently reported language combination (n=31) is DE>FR, followed by DE>IT.

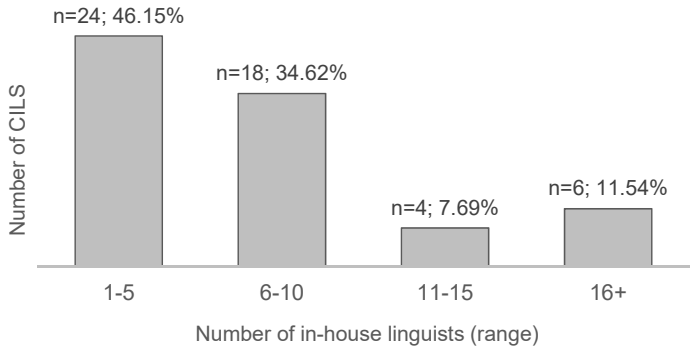



FIGURE 2.1 – Number of in-house linguists working at CILS (n=52)

Participants reported a number of in-house linguists ranging from one to 22 (Figure 2.1). Almost half of the respondents are small-sized CILS, employing one to five linguists in house. Eighteen respondents employ six to ten linguists in-house, while the remainder are large-sized CILS, with an in-house staff of more than ten linguists.

The majority of in-house linguists are French (40% of respondents) and Italian-speakers (24%), followed by German (20%) and English-speaking linguists (14%), who occasionally translate into another language. A few translators (2%) who work with other languages are also employed in-house.

Regarding the CAT tools in use, most respondents (27 out of 52) use Trados Studio. Other popular CAT tools are MemoQ and Across. Three respondents do not use any CAT tools, while three other respondents indicate they use DeepL, thus demonstrating confusion on the underlying differences between MT and TM (Figure 2.2) [15].

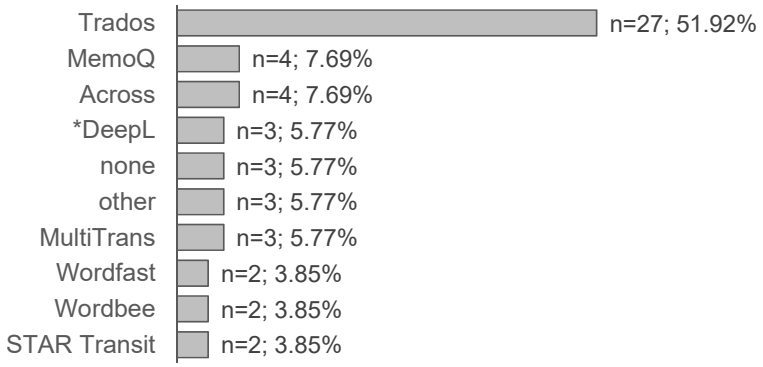



FIGURE 2.2 – CAT tool(s) currently in use (n=52)

15. Since multiple responses were allowed for this question, the sum of CAT tools in use is 53 instead of 52 (one per respondent).

## 2.4 Results

In the following sections, we will present the key findings derived from the questionnaire, beginning with an overview of the complete dataset (n=52). Subsequently, we will narrow our focus to respondents who currently incorporate MT in their production workflows (n=26), as the questionnaire was specifically designed for this group of participants. It is worth noting that not all questions were mandatory. Therefore, the number of respondents will be reported in each chart.

### 2.4.1 MT use

As shown in Table 2.1, 50% (n=26) of our respondents uses MT in production, while approximately 19% (n=10) is currently testing an MT system. Thirty-one percent (n=16) of our respondents do not use MT in their workflow.

| **Response** | ***N*** | ***%*** |
|---|---|---|
| Yes | 26 | 50.00 |
| Not yet (we are currently testing MT) | 10 | 19.23 |
| No | 16 | 30.77 |

TABLE 2.1 – Response count and percent for question A08 (Q1): *Do you use MT in your production workflows?* (n=52)

Out of those who responded *No* to the previous question, 31% (n=5) declared they are currently considering introducing MT in the next two years, while 44% (n=7) are still undecided. Four respondents do not consider introducing MT before 2024.

CILS who are currently testing MT (n=10), report they are using DeepL (n=6), a customised system provided by a specialised company, ModernMT or Microsoft Azure. Six respondents were also able to provide us with a date when they will start using MT in production: mostly in the first semester of 2022, one respondent in 2023.

### 2.4.2 Workflow

On average, respondents' main in-house operation is translation (approximately 57%), followed by revision (approximately 30%). Among other operations (13%) carried out in-house, participants mentioned post-editing, terminology management and translation memory maintenance [16].

While almost all in-house linguists revise texts, not all carry out post-editing jobs. An often-mentioned reason for this is the lack of customised systems for some language pairs. A few respondents also mentioned that PE is a recent practice and therefore the choice of using an MT system or not is currently left to the linguists.

16. This section and the following ones were accessible only to the respondents who already use MT in production (n=26).

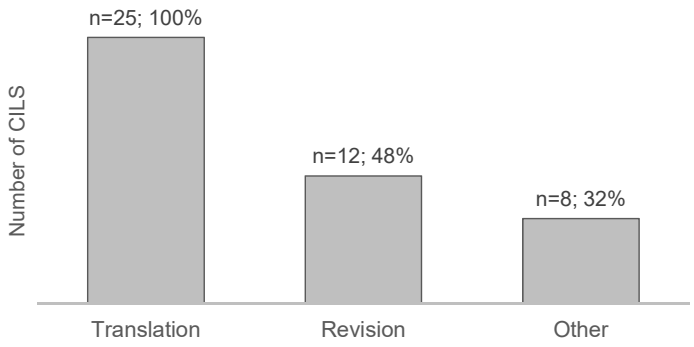



FIGURE 2.3 – Response count and percent for question B12.01 (Q1):
*Which operations do you outsource?* (n=25)

All departments except one (n=25) declare they outsource part of their operations (Figure 2.3): all of them outsource translation, approximately half of them also outsource revision jobs, while some of them (n=8) outsource other operations, including proofreading, formatting and aligning texts. In the same category, four respondents explicitly mention post-editing. Most CILS outsource to both external freelancers and agencies, with a slight preference for the latter (Figure 2.4).

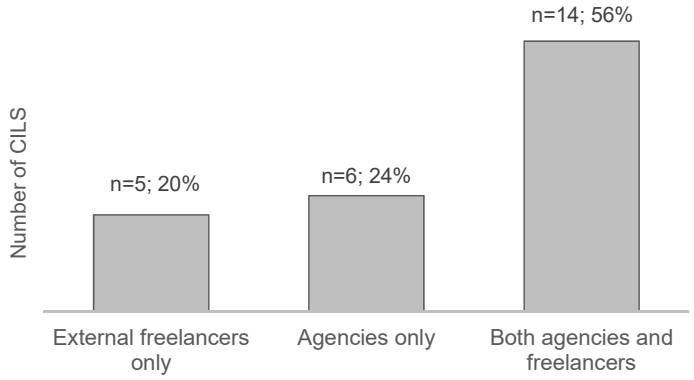



FIGURE 2.4 – Response count and percent for question B12.02 (Q1):
*To whom do you outsource your operations?* (n=25)

### 2.4.3 MT and PE

In terms of MT providers (Figure 2.5), most CILS use DeepL – two respondents explicitly mention DeepL Pro. Seven respondents are using a customised solution provided by the Swiss company TextShuttle. Three respondents use more than one system, while two respondents cannot disclose this information.

As for the type of MT system in use, 46% of respondents declare using generic engines, while 38% have opted for a customised solution. Fifteen percent of respondents use more than one type of system or a combination of the two (Figure 2.6). Expectedly, a striking majority of participants use an NMT solution, while one participant uses an SMT system. Four respondents cannot define the MT architecture they use.

Figure 2.7 shows in which year respondents started using MT in their production workflows. One respondent implemented an MT solution before 2018. Five participants started using

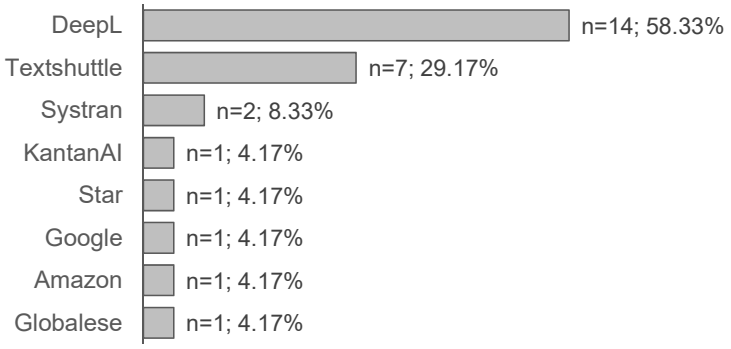



FIGURE 2.5 – Response count and percent for question C13 (Q1): *Who provides you with your current MT system?* (n=24)

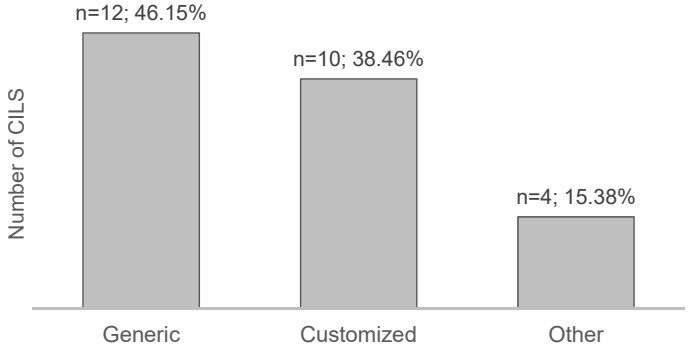



FIGURE 2.6 – Type of MT system currently in use (n=26)

MT in production in 2018. It should be noted that DeepL became available for the first time in August 2017, while the first version of DeepL Pro appeared in March 2018. This could have awakened interest in many in-house language services and could probably explain the rise in MT use in 2018. However, most of our respondents have implemented an MT solution only very recently, namely in the last two years.

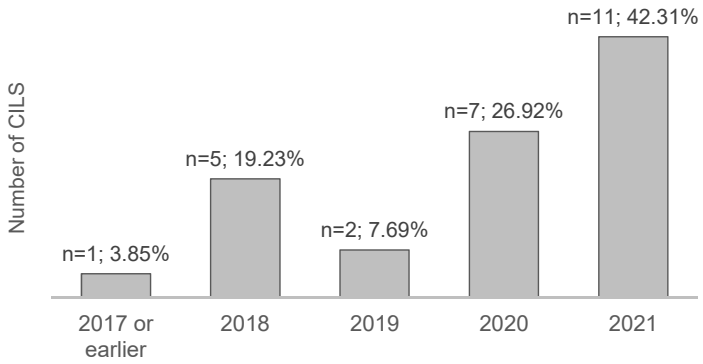



FIGURE 2.7 – Response count and percent for question C16 (Q1): *When did you start using your current MT system in the production workflow?* (n=26)

Before implementing MT in the workflow, most departments carried out at least an assessment of the quality produced by an MT engine (Figure 2.8). Several CILS have also complemented these assessments with a productivity test to verify whether using MT in production could lead to any productivity gains.

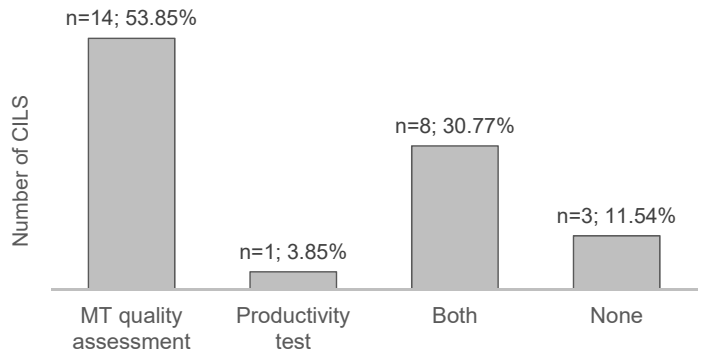



FIGURE 2.8 – Response count and percent for question C17 (Q1)*Did you assess the quality of MT or carry out a productivity test before integrating MT into the workflow?* (n=26)

The majority of respondents involved said that all in-house linguists participated in MT quality assessments or productivity tests (Table 2.2). Four respondents decided to test MT with a sample of in-house linguists. Only one respondent did not involve in-house linguists in these tests. Finally, linguists' feedback was always taken into consideration in the decision to implement MT, except for two respondents.

| **Response** | ***N*** | ***%*** |
|---|---|---|
| Yes, all of them | 17 | 77.27 |
| Yes, some of them | 4 | 18.18 |
| No | 1 | 4.55 |

TABLE 2.2 – Response count and percent for question C17-01 (Q1): *Did you involve your in-house linguists in these assessments?* (n=22)

Table 2.3 shows that 65% of respondents report that MT is only used on specific projects, while the remaining 35% apply MT on all sorts of projects. Text type is often mentioned among the criteria to apply MT or not. Tight deadlines are one of the reasons for choosing to use MT on a project, while confidentiality of text information is reported as a reason not to use MT. Some respondents also mentioned target audience, costs and productivity gains among relevant criteria.

| **Response** | ***N*** | ***%*** |
|---|---|---|
| Yes | 9 | 34.62 |
| No | 17 | 65.38 |

TABLE 2.3 – Response count and percent for question C18 (Q1): *Is MT used on each project?* (n=26)

The choice of using MT or not is often left to the linguist (44%) or made by the project manager before assigning the project (30%) (Figure 2.9). In some cases, the unavailability of a (customised) system for a specific language pair makes this choice easier. The head of language services – mentioned twice – also has a role in this decision. For one respondent, the head of the language group decides whether MT should be applied to a project or not. One respondent leaves this choice to the customers (i.e. for an in-house language department, the company's employees).

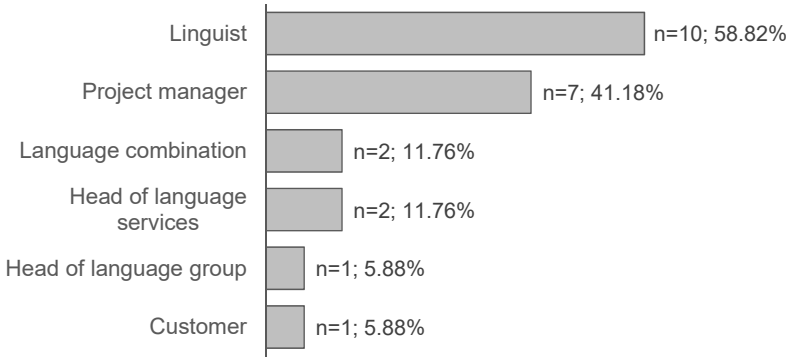



FIGURE 2.9 – Response count and percent for question C19 (Q1): *Who decides when to use MT?* (n=17)

The vast majority of respondents use the MT engine through their CAT tool, via a plugin. Those who do not integrate MT in a CAT tool use the engine via a web interface (Table 2.4).

| **Response** | ***N*** | ***%*** |
|---|---|---|
| Yes | 22 | 84.62 |
| No | 4 | 15.38 |

TABLE 2.4 – Response count and percent for question C21 (Q1): *Is MT integrated in your CAT tool?* (n=26)

Most respondents chose the option to provide linguists with an MT suggestion only if there are no other translation memory matches or when these matches are below a certain percentage of fuzziness (e.g. 85% fuzzy matches). Seven respondents stated they always provide linguists with an MT suggestion, regardless of whether the TM already offers valid matches.

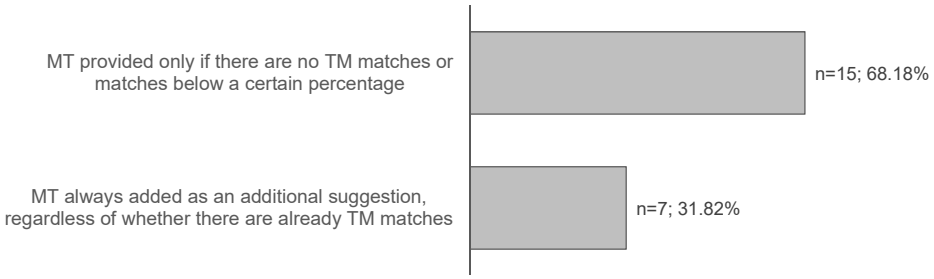



FIGURE 2.10 – Response count and percent for question C21.01 (Q1): *What is your MT/CAT tool integration scenario (regarding the display of suggestions)?* (n=22)

The majority of respondents pre-translate the entire text using MT and TM (offering only one suggestion per segment in the editing zone). Some respondents choose to present MT and TM suggestions in another window; therefore, the linguists need to select the suggestion they want to use or type it out in the editing zone. One respondent declared that their linguists could choose between both options (Figure 2.11).

In another question, we asked our respondents whether their in-house linguists could change default parameters and choose their own preferred integration scenario. This is true in

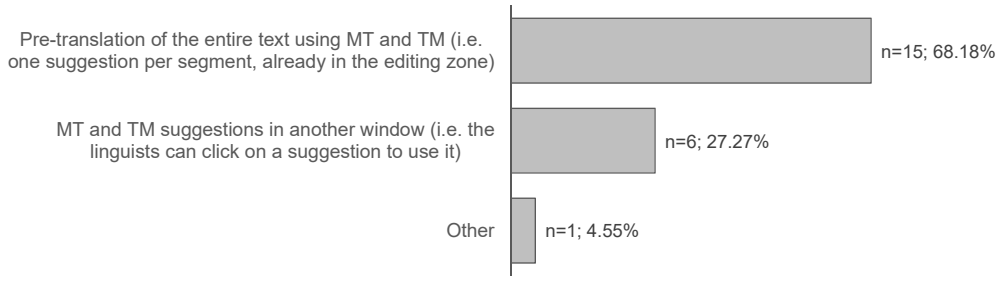



FIGURE 2.11 – Response count and percent for question C21.02 (Q1): *What is your MT/CAT tool integration scenario (regarding the use of suggestions)?* (n=22)

73% of the cases (16 out of 22 respondents), while in six CILS the default MT-CAT tool integration parameters cannot be modified by the linguists.
Most CILS (n=21) did not implement different levels of PE, while a minority of respondents (n=5) stated they apply light PE on some projects. Respondents commented that light PE is applied on internal and short-lived documents to convey the text's basic message when deadlines are very tight or when the customer requests it. As shown in Table 2.5, only 11 respondents use PE guidelines, mostly drafted in house [17].

| **Response** | ***N*** | *%* |
|---|---|---|
| Yes, drafted internally | 9 | 34.62 |
| Yes, existing guidelines (e.g. TAUS) | 2 | 7.69 |
| No | 16 | 61.54 |

TABLE 2.5 – Response count and percent for question C24 (Q1): *Do you have any PE guidelines that your in-house linguists can use for reference?* (n=26)

A slight majority of respondents provided in-house linguists with some training in PE, as shown in Table 2.6.

| **Response** | ***N*** | *%* |
|---|---|---|
| Yes | 15 | 57.69 |
| No | 11 | 42.31 |

TABLE 2.6 – Response count and percent for question C25 (Q1): *Did you provide your in-house linguists with any post-editing training?* (n=26)

### 2.4.4 Revision

We asked our respondents which terms they use to refer to internal quality assessment processes. Among these, we find *revision*, *review*, *end check*, *internal or external re-reading*, depending on who reads the final text, but also on how this is done (with or without the

17. One respondent clarified they use both existing and in-house PE guidelines

source text; the whole text or just some parts of it). Interestingly, five respondents included post-editing in this list.

A slight majority of respondents reported that not all (human and machine) translations are revised [18] (Table 2.7). This is the case for internal documents. Content-type, target audience, and customer requests are other often-mentioned discriminating factors in revising or not revising a translation. One respondent clarified that machine-translated (and post-edited) texts are never revised.

| **Response** | ***N*** | ***%*** |
|---|---|---|
| Yes | 12 | 46.15 |
| No | 14 | 53.85 |

TABLE 2.7 – Response count and percent for question D27 (Q1): *Are all translations revised?* (n=26)

In general, texts are always revised in their full length (73% of respondents). In a few cases, only some parts of the translations are revised (27% of respondents), depending on the contents' level of criticality, but also on who carried out the translation (e.g. if an external translator carried out the translation, the text must be revised in its entirety).

Revisers' corrections are usually sent back to the translator as suggestions (Figure 2.12). However, this is the case for only a slight majority of respondents (n=11), while for many other respondents' (n=9) corrections made during revision are directly applied to the final texts. Some respondents (n=6, category "Other") clarified that both scenarios apply, often depending on who carried out the translation (e.g. external or in-house translator).

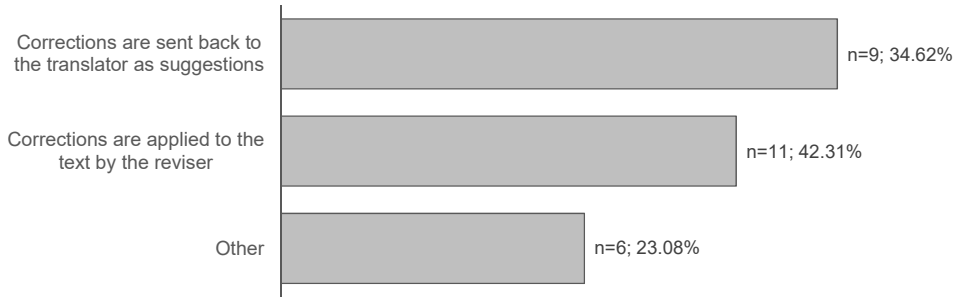



FIGURE 2.12 – Response count and percent for question D29 (Q1): *What is the status of revisers' corrections?* (n=26)

In the case of revision, the majority of respondents use revision guidelines or parameters, either existing ones (e.g. suggested by translation scholars) or drafted in-house [19] (Table 2.8). If we compare these figures with the companion ones about PE guidelines (Table 2.5), we observe that, compared to PE, revision is a more established practice.

18. In this question, we clarified that we intend revision as "the bilingual examination of source and target", carried out by "a person other than the translator", as defined in the ISO standard 17100:2015.

19. One respondent clarified they both use existing and in-house drafted revision guidelines or parameters.

| Response | *N* | *%* |
|---|---|---|
| Yes, drafted internally | 16 | 61.54 |
| Yes, existing guidelines or parameters | 2 | 7.69 |
| No | 9 | 34.62 |

TABLE 2.8 – Response count and percent for question D30 (Q1): *Do you have any revision guidelines or parameters that your in-house linguists can use for reference?* (n=26)

The fact that revision has been taught at university level for many years explains why fewer in-house departments felt the need to provide their linguists with revision training (Table 2.9) compared to PE training sessions organised by the same respondents.

| Response | *N* | *%* |
|---|---|---|
| Yes | 12 | 46.15 |
| No | 14 | 53.85 |

TABLE 2.9 – Response count and percent for question D27 (Q1): *Did you provide your in-house linguists with any revision training?* (n=26)

### 2.4.5 Revision after PE

According to our findings, post-edited texts are most often left unrevised (Table 2.10). They are sometimes revised depending on content type (e.g. sensitive documents) or target audience. In some cases, linguists themselves can ask for a revision by another colleague.

| Response | *N* | *%* |
|---|---|---|
| Yes, always | 6 | 23.08 |
| Yes, sometimes | 6 | 23.08 |
| No | 14 | 53.85 |

TABLE 2.10 – Response count and percent for question E32 (Q1): *Are post-edited texts also revised by another linguist?* (n=26)

When post-edited texts are revised, revisers usually know that the text they are asked to revise has been post-edited by a colleague (8 out of 12 respondents).

## 2.5 Discussion

Our findings indicate the growing popularity of MT and PE among Swiss-based companies with an in-house language department. DeepL is the most commonly adopted MT solution, particularly among small CILS, probably due to its user-friendly interface and quick setup process. Medium-to-large-sized CILS, on the other hand, have invested in customised MT solutions to provide better MT suggestions, especially for terminology, and to address privacy and confidentiality concerns associated with freely-available generic MT engines. This emphasis on privacy and confidentiality aligns with the concerns expressed by the respondents in the latest ELIS survey (2023).

Surveyed CILS are almost equally split between those who revise all translated content and those who do not. Similarly, for nearly half of them, the corrections made by the reviser are considered suggestions for the translators, while for a slightly smaller number of CILS, these corrections are directly applied to the text by the reviser. These practices align with those observed in interviews conducted among various Swiss in-house LSPs by Riondel (2021a, 2023), who also reports that in the former case the responsibility for the quality of the text is shared between the reviser and the translator, while in the latter case, it mainly falls on the reviser.

Our results also corroborate previous findings indicating that MT is not applied to all texts (Pielmeier and Lommel, 2019). However, it is surprising to note that when MT is used, post-edited texts are frequently left unrevised, with some exceptions depending on the content type or target audience. This observation is quite unexpected and contradicts the claims made by do Carmo and Moorkens (2021). Leaving post-edited texts unrevised enhances the risk of overlooking issues or inadvertently introducing errors, which is inherent in the PE task (Canfora and Ottmann, 2020) and could potentially be mitigated by having a second linguist check the post-edited content. It is important to acknowledge that this finding may be specific to CILS workflows and, thus, deserves further investigation.

Our findings indicate that, in the CAT tool interface, text is often displayed as fully pre-translated. A recent survey by Farrell (2022) conducted among professional translators (around 450 responses) showed that approximately 52% of respondents who employ MT are accustomed to pre-translating the entire source text. This practice is renamed *pure post-editing* by the author, in contrast to *hybrid post-editing*, that is, when MT is used in combination with TM matches within a CAT tool. In the same survey, the majority of respondents reported using MT primarily when there are no other suitable TM matches. Our findings align with this observation.

We maintain that with *pure post-editing* it becomes challenging for linguists to determine the appropriate approach to adopt – as this configuration gives the impression that the text requires revision, rather than letting linguists considering MT as *a tool*, an aid for drafting the translation. The ergonomics of MT/CAT tool integration and its impact on the translation process warrant further investigation. We also argue that this aspect could potentially influence the satisfaction of linguists (see Chapter 4 for further details).

Overall, the survey results seems to indicate that, compared to PE, revision is a more established practice. In fact, while most respondents have not yet set out any PE guidelines for in-house linguists, some revision guidelines or parameters have been defined in 17 out of 26 CILS. Almost all in-house linguists engage in revision tasks, but not all of them perform PE tasks. Furthermore, the fact that revision has been taught at university level for many years could explain why fewer in-house departments felt the need to provide their linguists with revision training. Conversely, in the case of PE, CILS management often provided some MTPE training to their in-house staff.

Respondents reported that in cases where post-edited texts undergo revision, revisers are usually aware of the text's origin (i.e., whether it has been translated or post-edited). It would be interesting to verify whether revisers modify their approach to the text based

on how the text has been produced [20].

Finally, we note that the majority of CILS organised at least one evaluation of MT quality before implementing MT into their workflows, and they actively involved in-house linguists in this assessment. We believe that this practice has the potential to foster the acceptance of MT among salaried translators, laying the foundation for a successful integration of MT into the workflow. In this regard, we share the viewpoint expressed by Silva (2014, p. 26), who stated that "LSPs efforts to push MT without taking into account feedback from translators, reviewers and other concerned language professionals are doomed to fail".

Before concluding this section, we must acknowledge two main limitations of our study. Firstly, the size of our sample is very small, which limits our ability to generalise our findings. While we made efforts to reach out to all Swiss CILS, we cannot be certain that we reached the entire population. However, we have identified an existing network of CILS that may expand as our research becomes more widely known. Secondly, our data collection method relied on self-administered questionnaires, which only allow for the collection of surface information and do not permit the researcher to clarify misunderstandings or request additional details on responses (Oates, 2005a).

Lastly, it would be useful to conduct an update of this survey to determine whether CILS that were not interested in MT at the time of data collection (end of 2021) have changed their stance in the meantime.

## 2.6 Summary of the chapter

In this chapter, we presented the findings of a survey-based study conducted among Swiss CILS to investigate their workflows, with a specific emphasis on the use of MT and the organisation of PE and revision tasks. The questionnaire was addressed to language services' directors and project managers (one answer per CILS) and we obtained 52 valid responses. At the time of data collection (November 2021–February 2022), approximately 19% of respondents were testing an MT system, while 31% did not use MT in their workflow yet. Half of our respondents (n=26) used MT in production: the remainder of the questionnaire focused on these respondents only.

DeepL was reportedly the most used MT system, followed by customised solutions. For 65% of respondents, MT was only used on specific projects, while the remaining 35% applied MT on all sort of projects. Respondents noted that the choice of using MT or not is often left to the linguist or made by the project manager before assigning the project. Only one of our respondents clarified that this choice is left to the customers. Most respondents did not implement different levels of PE (light/full PE), while five out of 26 reported they apply light PE on some internal and short-lived documents, to convey the text's basic message when deadlines are tight or when the customer requests it. Twenty-two out of 26 CILS used MT through their CAT tool, via a plugin. As for the MT-CAT tool integration, most

20. This question has been included in questionnaire Q2, detailed in the upcoming chapter.

respondents (68%) opted to provide linguists with an MT suggestion only if there are no other TM matches, or when these matches are below a certain percentage of fuzziness. Results revealed that source texts were often pre-translated using MT and TM (*de facto* providing the linguists with one suggestion at a time, in the editing zone). Nonetheless, in most of the cases, linguists are able to change the default parameters and choose their own preferred integration scenario.

Parallel questions on PE and revision practices highlighted the following results: the majority of respondents did not establish any PE guidelines for their in-house linguists. In contrast, approximately 65% (17 out of 26) of the surveyed CILS had defined revision guidelines or parameters. While almost all in-house linguists engaged in text revision, not all of them performed PE tasks. A smaller proportion of CILS found it necessary to provide revision training to their linguists compared to the number of CILS offering PE training to their in-house staff. Post-edited texts were typically left unrevised, although with some exceptions based on factors such as content type (e.g., sensitive documents) or target audience. Finally, respondents reported that in cases where post-edited texts undergo revision, the revisers are usually aware of the text's origin.

Responses from this first questionnaire let us address **Goal 1**, setting out the context for a better understanding of information gathered from a second, related questionnaire dealing with revision and PE practices from the perspective of in-house linguists. This second questionnaire will be detailed in the next chapter.

# Chapter 3

# PE and revision practices: insights from CILS linguists

## 3.1 Overview

In this chapter[1], we report on the second questionnaire (Q2, available in Appendix D) of our survey-based study. This questionnaire was addressed to CILS' linguists who translate, revise and post-edit texts, and included questions about their PE and revision practices. Specifically, we investigated how in-house linguists carry out revision and PE in terms of (i) reading strategies, e.g. whether they read the source or target text first, and (ii) overall strategies, e.g. whether linguists follow specific parameters or guidelines, whether they revise and post-edit on screen or on paper, and whether they check terminology during revision and PE tasks. Additionally, we also investigated whether linguists apply the same strategies when revising texts that have been translated or post-edited by another person. To the best of our knowledge, this is the first survey explicitly comparing revision and PE practices of professional translators.

The remainder of the chapter is structured as follows: Section 3.2 details previous survey-based studies that dealt with revision and PE practices, respectively. A special focus will be put on studies conducted in Switzerland. Section 3.3 describes the survey design, sampling methods and the profile of the respondents. Results are presented in Section 3.4 and further discussed in Section 3.5. Finally, a summary of the chapter is provided in Section 3.6.

## 3.2 Previous studies on PE and revision practices

### Revision

Several researchers have used country-specific surveys to investigate revision practices. In Belgium, Robert (2008) launched two small-scale surveys (48 and 21 responses, respectively) among translation agencies to establish which translation revision procedures and revision

1. The reader should note that a slightly modified version of this chapter has been published as Girletti (2022).

methods (revising on paper and/or on-screen) are the most used. She found that while revisers use different procedures, most compare source and target texts to make corrections and then reread the target text one last time. Results also suggest that revision is mainly carried out on screen.

This latter aspect was also included in a survey-based study conducted by Scocchera (2015, 2017) in the Italian publishing sector. The study included two questionnaires: one for translators to investigate self-revision practices (55 participants) and one for revisers to investigate other-revision practices (25 participants). Results of the latter show that revision is mainly carried out on-screen, but the choice of the medium depends on various factors and on-screen is preferred if the translation needs many corrections. Regarding revision methods, 60% of revisers do not read the whole source text before starting to revise, primarily due to "lack of time and cost-effectiveness" (2017, p. 13). Instead, participants claim they mostly compare source text and target text segment by segment.

In Denmark, Rasmussen and Schjoldager (2011) surveyed 24 translation companies about their revision policies and conducted 13 follow-up interviews with survey respondents and in-house revisers in five of these companies. Collected data suggest that not all texts are revised. This depends on different factors, including the translator who translated the text, assignment difficulty, text type/genre, intended use, and customer. The most used procedure is monolingual revision followed by a comparative revision or vice-versa. However, interviews reveal that revision is rarely fully comparative. Most companies do have revision guidelines, but not in written form.

In Austria, Schnierer (2021) surveyed translation companies to determine whether their revision practices complied with the former translation standard EN15038 (replaced by ISO 17100). She found that two out of six certified companies do not systematically revise translations, although the standard requires this. Regarding revision methods, all companies report comparing the translation with the source text. In contrast, only one uncertified company reported performing monolingual revisions of the target text (referring to the source text if needed). Five out of six certified companies use revision parameters, while this applies only to six out of thirteen uncertified companies.

Hernández Morin (2009a) conducted a survey among translation practitioners (115 respondents, primarily freelance translators) to find out about revision practices and perceptions of revision in France. Two of her questions dealt with the revision of automatically pre-translated segments, i.e., those coming from a CAT tool and machine translation, respectively. 69% of respondents state that they do not work with machine-translated texts, 23% claim they revise those texts in-depth, and 6% revise the text to ensure just its overall comprehension. In the author's thesis (Hernández Morin, 2009b), both processes are referred to as *post-editing*. Therefore, it is not clear whether respondents refer to post-editing or actual revision practices.

Lastly, a recent study by Riondel (2021a) pointed out similarities and differences between revision policies of two cohorts of salaried translators working in Switzerland. The researcher conducted 20 semi-structured interviews in a sizeable intergovernmental organisation and a medium-sized language department of the Swiss Confederation. He found that while revision

is mainly carried out on screen in the former context, at the Confederation, texts are often printed before revision. In both settings, revisers apply a complete bilingual revision, but those who work at the intergovernmental organisation also consider other types of revision (e.g. spot check for outsourced translations).

Another study by the same author (Riondel, 2022) focused specifically on reading strategies employed by the revisers interviewed (n=22). The findings indicated that approximately half of the participants start by reading the source text, while the other half begins with the target text. Those who read the source first do so to better spot mistranslations, while revisers who read the target first indicated they do so "to avoid interferences between the languages or better appreciate the readability or correctness of the target language" (*ibid.*, p.83). Furthermore, only three participants indicated that they adapt their reading strategy, while the others seem to have their preferred approach that is applied consistently.

**Post-editing**

When it comes to defining how the task is carried out, studies of revision practice outperform those on PE practice. In participant-oriented studies, PE discourse most often concerns adoption rates and attitudes toward the task (Guerberof Arenas, 2013; Gaspari, Almaghout, and Doherty, 2015; Zaretskaya, 2015; Läubli and Orrego-Carmona, 2017; Nunes Vieira, 2020a). For instance, in a survey of the state of the linguist supply chain, researchers at Common Sense Advisory (Pielmeier and O'Mara, 2020) reported that, out of 6,997 respondents, 55% use MT on most projects or when the customer requests it.

Some studies focused on salaried translators as the target population (Cadwell, O'Brien, and Teixeira, 2017; Rossi and Chevrot, 2019), but did not investigate how MT was introduced in the workflow and integrated in the CAT tool, nor how PE was carried out in terms of reading strategies, i.e., which text – source or target – is read first.

To the best of our knowledge, the only survey-based study that deals with this topic is the one by Ginovart Cid (2021), who surveyed European LSPs, university lecturers and language professionals about their MTPE practices and training protocols. Results of the questionnaire sent to PE educators – detailed in Ginovart Cid and Colominas (2021) – showed that 49% of respondents do not provide their students with any advice on whether the source or the target segment should be read first, while 33% of instructors advise reading the source text first. It must be noted that the question was asked in a close-ended, single-answer format. Therefore, other possibilities, e.g. reading the whole source or target text before starting to post-edit, are not explored. The question on reading strategies was also included in the questionnaire addressed to professional linguists, but the results are not discussed in any publication to date.

Moving on to the Swiss context, we could not find any studies on PE practices, but we found one study on MT adoption and attitudes towards MT and PE. Yuste (2002) carried out a survey among Swiss LSPs about their use and perception of translation technology. The author concluded that, at the time of writing, there was "no overall interest in MT in the Swiss translation arena". However, we are unable to further comment on these findings since, in the electronic version of the paper, the section describing collected data is missing.

As the review of existing literature pointed out, while there have been several surveys on revision procedures, PE procedures have been only scarcely investigated. Most importantly, reading strategies in revision and PE have never been studied with a contrastive approach in a context where both activities are carried out. Our research will try to fill this gap.

## 3.3 A survey on PE and REV practices of CILS linguists

### 3.3.1 Sampling

Questionnaire Q2 was distributed to in-house linguists working at CILS who use MT in production (n=26, as detailed in Chapter 2). Dissemination was mainly handled by CILS's directors or project managers who filled out the first questionnaire [2]. In most cases, these respondents included the main researcher when sending the email invitation to their employees or colleagues, enabling us to send a reminder after some time. As in the first questionnaire, both the email and the consent form (Appendix B) clearly stated that participation in the study was voluntary and anonymous. The questionnaire was hosted on the LimeSurvey platform and was made accessible from November 15th, 2021, until February 16th, 2022. Depending on respondents' answers, the questionnaire included up to 58 questions, but not all were mandatory.

### 3.3.2 Survey structure

The questionnaire was structured in five sections:

**Section A** [*Respondent's profile*] contained demographic questions, such as age and mother tongue of the respondent, years of translation experience and years of employment in the CILS. This section also included two questions about how often respondents perform revision and PE – to ensure that participants carried out these activities in their workflow.

**Sections B** [*Revision*] and **C** [*Post-editing*] contained two symmetric sets of questions related to different aspects of the two activities, such as the primary reading strategies used by respondents when revising and post-editing.

**Section D** [*Post-editing, revision and overall strategies*] comprised three questions on the relationship between revision and PE: whether participants used the same strategies when revising human-translated texts and post-editing MT content or when revising texts that had been previously translated or post-edited. The third question asked whether the introduction of MT in the workflow brought about any changes in

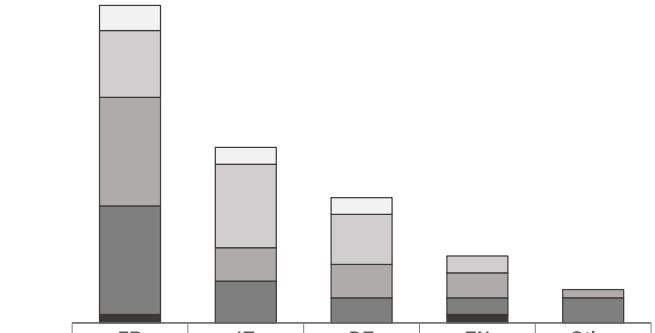
2. In some cases, the email invitations sent to CILS directors and project managers to fill out questionnaire Q1 also contained a link to questionnaire Q2. Clear instructions regarding the applicability criteria of this second questionnaire were included, along with a request to forward the link to in-house linguists. However, we did receive some responses from companies that were in a testing phase and had not yet integrated MT into their production workflows. Consequently, we made the decision to discard these responses, as they did not meet the fundamental requirement for participation in the study.

the way revision was carried out. Participants were encouraged to comment on their answers.

Lastly, **Section E** [*Satisfaction*] focused on respondents' satisfaction in performing translation, revision and post-editing. The results of this section will be presented in Chapter 4.

### 3.3.3 Participants' profile

Eighty-six CILS linguists completed the questionnaire[3]. The most represented mother tongue is French (44.19% of respondents), followed by Italian (24.42%). German ranks third (17.44%), while English is the mother tongue of 9.3% of respondents. Two respondents identified themselves as bilingual, while two others indicated different mother tongues (4.65% of respondents overall). Age is well distributed across ranges and per mother tongue, except for the most extreme categories (18-29 and 60+, including young linguists or translators approaching retirement, respectively), as shown in Figure 3.1.



| | FR | IT | DE | EN | Other |
|---|---|---|---|---|---|
| 18-29 | 3 | 2 | 2 | 0 | 0 |
| 30-39 | 8 | 10 | 6 | 2 | 0 |
| 40-49 | 13 | 4 | 4 | 3 | 1 |
| 50-59 | 13 | 5 | 3 | 2 | 3 |
| 60+ | 1 | 0 | 0 | 1 | 0 |

FIGURE 3.1 – Age range and mother tongue of participants (n=86).

Translation experience ranged from two to 36 years, with an average of 15.8 years and a median of 14.5 years. Overall, participants have been working at their respective CILS for an average of nine years and a median of seven years. All participants indicated they revise texts and use MT in their workflows, but the proportion of those who revise almost daily is slightly higher than those who post-edit texts nearly every day (79.07% versus 72.09%, respectively). Regarding participants' experience with the tasks, while 66.28% of respondents reported they already had some revision experience, most participants (80.23%) started PE at their current company. This result is expected and in line with the recent introduction of MT in many Swiss CILS. Finally, regarding training, most respondents declare having attended a PE training session (53.49%), while only 40.7% of revisers have been trained to carry out revision jobs.

3. Further information on data cleaning after data collection will be provided in the next section.

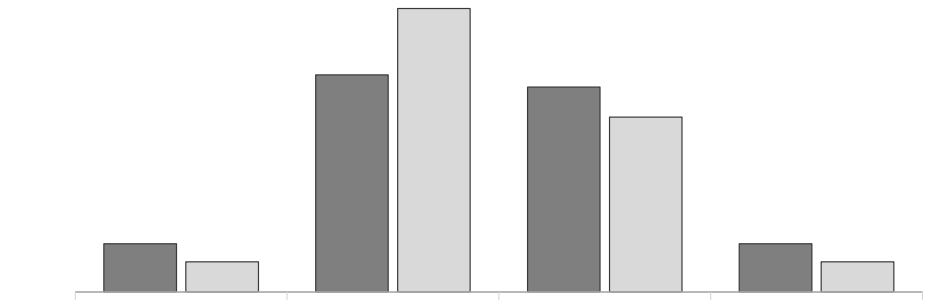
## 3.4 Results

We initially received 107 responses, but we had to discard 18 of them for various reasons. Five responses came from linguists who do not perform revision or PE in their daily jobs and were incomplete. Six responses came from linguists working at companies who did not yet use MT in their production workflow. Seven responses could not be traced back to any company that filled out the first questionnaire; this happened because we had only partial control over how the questionnaire was circulated (as explained in Section 3.3.1). Additionally, five valid responses were incomplete, but we decided to keep those who at least completed the first four sections of our survey (n=2).

In total, we retained 86 valid responses from 23 Swiss CILS. If we consider the number of in-house linguists indicated by each company in questionnaire Q1, we can calculate a response rate of 44%. However, we cannot compare this response rate with that of other surveys focusing on similar topics, mainly because we decided to address a specific group of stakeholders and focus on a geographical area that is scarcely represented. On average, 50% of linguists in each company have responded to questionnaire Q2. We did not receive any responses from linguists working in three out of 26 companies who currently use MT in production (as indicated in Chapter 2).

### 3.4.1 Reading strategies

As shown in Figure 3.2, the most used reading strategy is to proceed segment by segment, starting from the source text. This is slightly more common in PE (approximately 55% of participants) than in revision (42%). The second most used strategy is the opposite one, in which linguists start by reading the target segment (approximately 34% in PE and 40% in revision). Only a few respondents claim to read the whole target text while referring to the source in case of issues, especially when revising. Five respondents claim to use this strategy during PE.



| | I read the whole target text and refer to the source only if there are issues in target quality | I read the source segment first, then the target segment | I read the target segment first, then the source segment | Other |
|---|---|---|---|---|
| REV | 9,30% | 41,86% | 39,53% | 9,30% |
| PE | 5,81% | 54,65% | 33,72% | 5,81% |

FIGURE 3.2 – Reading strategies employed during PE and revision tasks (n=86).

A few respondents claim to use other reading strategies when revising and post-editing. From a closer inspection of their comments, we understood that revisers' strategies depend on different factors, such as the text type, the translator who carried out the translation or the customer who requested it. One respondent described their strategy, which we found to match significantly with our first-listed strategy (reading the whole target text and referring to the source in case of issues or to check numbers and tags). One reviser uses a two-step revision strategy (monolingual proofreading followed by bilingual revision), while another one reads the source and target in parallel.

Regarding PE, one respondent is unable to provide us with an answer since this person only uses MT as a further suggestion in the CAT tool. Two respondents mentioned they vary their strategies depending on the text, while two others read the source and target in parallel. Although the latter did not clarify whether they start with the source or target segment, we note that this strategy enables linguists to quickly shift attention between the source and the pre-translated text. We maintain that checking source and target text in chunks instead of reading the whole segment could benefit linguists' text comprehension, especially in case of longer sentences or complex syntactic structures.

In an additional question, we asked our participants whether they vary their strategies depending on the text or other factors. The answer was positive for approx. 63% of revisers and 37% of post-editors. Therefore, in PE, linguists tend to apply the same reading strategy more often than in revision. Criteria often cited by revisers to vary their preferred reading strategy are text type, time constraints and the translator who translated the text. In contrast, post-editors mention text type, text complexity, target audience, text length, and PE level (light or full) to be applied.

### 3.4.2 Overall strategies

| **Response** | ***N*** | ***%*** |
|---|---|---|
| Yes, I use the same strategies | 31 | 36.05 |
| No, I use different strategies | 55 | 63.95 |

TABLE 3.1 – Response count and percent for question D23 (Q2): *Do you use the same overall strategies when revising human-translated texts and post-editing machine-translated texts?*

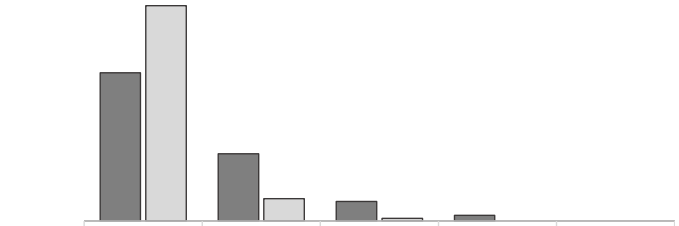
Most respondents (64%) claim to apply different strategies when working with human-translated or machine-translated texts (Table 3.1). Comments show that respondents trust MT less than human colleagues. Linguists are aware that humans and machines do not commit the same errors; therefore, they are much more careful when working with MT than when revising human-translated texts.

When analysing responses on reading strategies (Section 3.4.1), however, we found that around 65% of linguists reported using the same strategy during revision and PE. This could suggest that, although respondents claim to be aware of the differences between the two activities, in practice, they behave in the same way when revising and post-editing, at least regarding reading strategies.

### 3.4.3 Parameters and guidelines

We asked our respondents whether they use any revision parameters (Mossop, 2020) or PE guidelines (Hu and Cadwell, 2016) during revision and PE, respectively. Results show that revisers mainly follow specific parameters (72% of respondents), while only slightly more than half of respondents follow any PE guidelines (51%). These figures show that, compared to PE, revision is an established practice with a long-standing tradition.

We also asked our respondents whether and how often they verify that terms are correctly rendered in the target language (Figure 3.3). Studies on revision practice report that revisers do not always check terminology, especially if they know that the translators have already taken care of it (Allman, 2007; Riondel, 2021a). Conversely, a guideline that is often cited in full PE is to check whether terminology is correct in the target language (Hu and Cadwell, 2016).



| | Always | Often | Sometimes | Rarely | Never |
|---|---|---|---|---|---|
| REV | 61,63% | 27,91% | 8,14% | 2,33% | 0,00% |
| PE | 89,53% | 9,30% | 1,16% | 0,00% | 0,00% |

FIGURE 3.3 – Responses to question: *When revising/post-editing, do you check whether terminology is correct?*

Among our respondents, post-editors seem to be aware of this issue and systematically check whether terminology is correct in the target text (approx. 90% of respondents). A tiny percentage of post-editors check terminology "often", while only one respondent admits to only checking it "sometimes". On the other hand, when revising texts, only 62% of respondents systematically check terminology in translated texts; 28% indicated they often check terminology, 8% only sometimes, and 2% rarely do so. Some revisers commented on their answers and confirmed that they check terminology mainly depending on the translator who translated the text.

### 3.4.4 Screen vs Paper

Most respondents admit they never print texts before revising (85%) or post-editing (93%). The remaining percentage of revisers and post-editors admit printing translated texts only "sometimes", mostly when the text is deemed complex, if it has to be published, or just as a means to vary one's working method. No one revises or post-edits texts systematically on paper.

### 3.4.5 Revising post-edited or human-translated texts

In another question, we asked participants whether they use the same overall strategies when revising texts with different origins, i.e. texts that had been previously translated or post-edited by a colleague. The answer is clear-cut (Table 3.2): 78% confirm using the same strategies, thereby considering translated and post-edited text as the product of human work. Those who admit using different strategies clarify that, when revising post-edited texts, they mainly focus on textual cohesion and terminology consistency or check source and target texts very carefully to ensure that post-editors have not overlooked any MT errors.

| **Response** | ***N*** | ***%*** |
|---|---|---|
| Yes, I use the same strategies | 67 | 77.91 |
| No, I use different strategies | 19 | 22.09 |

TABLE 3.2 – Response count and percent for question D24 (Q2): *Do you use the same overall strategies when revising human-translated texts and revising texts that have been post-edited by another person?*

In the Comment section, 13 linguists reported that revision of post-edited texts is not carried out in their CILS or that they never know the origin of the text. When cross-checking these results with those from questionnaire Q1 [4], we found that 45 out of 86 respondents did not carry out revision of post-edited texts in their workflows. Nevertheless, they have answered the question based on what they *would do* if they were to revise post-edited output.

### 3.4.6 MT influence on revision procedures

We also asked our respondents whether the introduction of MT in the workflow had somehow influenced the way revision of human-translated texts is carried out. The majority of respondents (72%) consider that this is not the case. The analysis of comments from those who did notice a change (28%) revealed that this question had primarily been misunderstood. The way this question was asked has probably confused those respondents who consider PE as "the revision of MT output" (Mossop, 2020). Indeed, many participants commented again on how they tackle revision and PE, detailed their overall strategies or listed the differences between human-translated and machine-translated texts.

Only a few participants seem to have correctly understood the question and commented that, compared to what they used to do before the introduction of MT in their workflows, they now focus more on accuracy errors (typical MT errors) during revision.

4. In questionnaire Q1, we found that post-edited texts are always revised in six out of 26 CILS. Post-edited texts are sometimes revised (n=6) depending on content type or target audience. In some cases, the linguist can ask for a revision by another colleague. Otherwise, the majority of respondents (n=14) clarified that post-edited texts are never revised.

## 3.5 Discussion

The questionnaire for CILS linguists has provided interesting insights into the relationship between PE and revision in professional CILS workflows. Firstly, we observed that the majority of CILS linguists received specific training in PE. This is motivated by the fact that PE has been introduced only recently in the workflow of the surveyed CILS. Therefore, it was necessary to provide linguists with initial training to familiarise them with how MT works and on how PE should be performed. Fewer linguists received comparable training in revision. These findings are in line with those of Scocchera (2015), who found that 72% of revisers working in the Italian publishing sector had not received any revision training.

The question about reading strategies has revealed that CILS linguists proceed almost exclusively segment by segment. This findings suggests an influence of the way the source text is presented in the CAT tool, an aspect that many translators have previously lamented (LeBlanc, 2013; O'Brien et al., 2017).

Studies on the influence of different reading strategies on post-editors' and revisers' efficiency are extremely scarce. Recently, Mossop (2023) has published a contribution summarising the limited existing research and called for more studies on this topic, which is relevant for both professional practice and pedagogy. In a study of students' PE performance (Volkart et al., 2022), we found that participants who start by reading the source text during PE introduce slightly more preferential changes than those who begin by reading the target. Additionally, there was no significant influence of the strategy on the ratio of corrected errors or on the time spent on the PE task. In revision, Künzli (2009) found no statistically significant differences in the performance of participants who started by reading the target text and those who started with the source text, except for one out of three texts used in the experiment. In contrast, Ipsen and Dam (2016) found that revisers who start by reading the target text detect more errors than those who read the source text first. However, since the time to complete the task was not taken into account, it is unclear whether this procedure is faster than the opposite one.

These findings would suggest that if linguists had to choose the same reading strategy to carry out revision and PE jobs, then reading the target text first would probably be the best option. However, this contradicts the common practice of our respondents. Indeed, the slight majority of participants reported starting with the source, both during PE and revision – the latter finding aligns with those of Riondel (2023).

A few participants also reported that in PE tasks they typically read the entire target text and refer to the source text only when issues arise. It is worth noting that monolingual PE is nowadays a dangerous practice since omissions are not infrequent in neural MT, and the fluency of NMT output can be misleading (Castilho et al., 2017b).

The way many of our respondents misunderstood a question about the possible influence of MT on revision procedures makes us think that there is a sort of cognitive bias toward considering PE as the revision of MT. Such bias could affect the behaviour of some linguists who could not perceive working with MT as a means to vary their daily tasks but rather as a mere increase in the number of revision jobs to carry out. Displaying MT in a separate

window (just as with TM fuzzy matches) instead of pre-translating the entire text could perhaps help linguists consider MT as a tool supporting their translation workflow – rather than a "translation dispenser" whose output must be corrected.

While research on revision procedures has shed light – at least partially – on the impact of different revision strategies on revision quality, task duration and error detection potential (Robert, 2013; Robert and Van Waes, 2014; Ipsen and Dam, 2016), similar studies on PE strategies are extremely scarce. As a result, PE training rarely includes practical advice on how to carry out the task (Ginovart Cid and Colominas, 2021). Our questionnaire indicates that CILS linguists claim to use different approaches for PE and revision, but the majority of them reported employing the exact same reading strategy for both tasks in practice. It remains to be demonstrated whether using the same or different reading strategies in PE and revision could benefit linguists' performance or influence their attitudes toward the task.

Lastly, we acknowledge that our chosen method for investigating reading strategies (i.e., a self-administered questionnaire) may not fully capture the reality of what post-editors and revisers do. Previous eye-tracking experiments found that post-editors typically start by reading the target segment (Carl et al., 2011; Mesa-Lao, 2014). In contrast, student translators in a study conducted by Huang (2018) began by reading the source segment during a revision task. Future research employing an eye-tracker and involving the same pool of participants could verify whether our self-reported data align with actual practices, or reveal discrepancies between self-perceived practices and actual behaviours.

## 3.6 Summary of the chapter

This chapter delved into the first key aspect identified in pursuit of Goal 2, namely revision and PE practices of professional linguists. We started with a review of prior investigations of PE and revision practices within the domain of Translation Studies. This review exposed an existing imbalance, with a preponderance of studies dedicated to revision practices overshadowing those dedicated to PE practices. The focus of PE researchers had been mostly on surveying professionals regarding their attitudes towards MT and their use of MT tools. However, practical aspects such as the reading order of the texts had been somewhat overlooked.

To address this gap, we engaged in a survey of in-house linguists working at CILS that used MT in production workflows. We designed a tailored questionnaire to investigate their PE and revision practices. We were particularly interested in understanding their reading strategies when handling pre-translated texts, whether they adhered to specific parameters and guidelines, whether they performed PE and revision tasks on screen or on paper, and whether they verified terminology during these tasks. The symmetrical structure of the questionnaire facilitated our comparison of responses for both tasks.

A total of 86 valid responses were collected. The analysis of linguists' reading strategies revealed that the predominant approach involved reading the texts segment-by-segment, typically starting with the source text. This strategy was slightly more prevalent in PE than in revision. The second most used strategy was the opposite one, demonstrating the

influence of CAT tools' text segmentation on linguists' reading processes. The questionnaire also revealed that post-editors tend to adopt a more consistent approach in their reading strategies, while revisers adapt them depending on text type, time constraints, and the translator who performed the translation.

In PE tasks, an overwhelming majority of linguists systematically check terminology accuracy in the target text. In contrast, only over half the linguists do so in revision tasks. Again, the decision to check terminology often hinged on the identity of the translator. Findings showed that PE and revision tasks are mostly carried out on screen. However, some linguists occasionally opt to print texts, typically when dealing with complex documents or as a means of diversifying their working methods.

Linguists reported a tendency to employ different overall strategies when dealing with PE and revision tasks. Nonetheless, a significant proportion of respondents indicated the same reading strategy in both PE and revision tasks. This observation suggests that, although linguists are aware of the differences between these tasks and the need for applying distinct approaches, these differences do not necessarily extend to reading strategies.

Lastly, the majority of respondents stated that the integration of MT into the workflow had not significantly altered their revision procedures. Some responses, however, indicated that the phrasing of the question may have led to misunderstandings.

The findings presented in this chapter contributed to a better understanding of how CILS linguists approach the tasks of PE and revision. In the upcoming chapter, we will delve into another critical aspect of the comparison between PE and revision: the satisfaction levels of linguists engaged in these tasks.

# Chapter 4

# Satisfaction of CILS linguists in performing PE and revision tasks

## 4.1 Overview

As presented in Section 1.3, the second key aspect defined under Goal 2 is the satisfaction experienced by the linguist performing PE and revision tasks. We define a *task*, or *work activity*[1], as a set of interrelated actions carried out to complete a piece of work. Applying this definition to professional translation contexts, we consider tasks all the translation-related activities that involve text production or text correction. In particular, we will focus our attention on translation, revision and PE. These tasks share some elements and differ on others. For instance, all involve reading the source text, but revision and PE require the linguist to also read the target text – produced by a human or by a machine, respectively – before performing any other action. These three tasks constitute a fundamental part of linguists' daily work. Nonetheless, different linguists can find some tasks more or less satisfying (i.e., enjoyable) than others.

In the field of industrial-organisational psychology, *task satisfaction* has often been used as a proxy of *job satisfaction*. Companies have been interested in assessing employees' job satisfaction since it can affect performance and possibly predict turnover. Furthermore, satisfaction has a diagnostic value, since it can be used to determine which aspects of the job or the tasks performed on the job should be improved. Translators' and interpreters' professional satisfaction[2] has also been investigated by some scholars in Translation and Interpreting (T&I) studies. However, task satisfaction has been heavily under-researched in this domain. In particular, there are no studies that focus on which tasks linguists appreciate the most or the least, and on the reasons behind their attitudes towards these tasks. We tried to fill this gap in the literature by designing a number of questionnaire items related to task satisfaction and including them in our survey of CILS linguists (questionnaire Q2,

1. In the present thesis, we use *task* and *activity* as synonyms.
2. In the present thesis, *job satisfaction* and *professional satisfaction* are used interchangeably.

detailed in the previous chapter). Through the analysis of responses to these items, we will try to provide an answer to the following research question:

> **RQ2**: *In the professional workflow of Swiss CILS, how does PE compare to revision in terms of task satisfaction, for the linguist who performs these tasks?*

The present chapter is structured as follows: in Section 4.2, we will introduce the concepts of job and task satisfaction in I/O psychology and we will present studies on linguists' satisfaction conducted in the T&I domain. In Section 4.3, we will detail how we designed our questionnaire items using ten core concepts retrieved from both I/O psychology and Translation Studies. In the same section, we will also illustrate the specific hypotheses that will guide our investigation. In Section 4.4, we will provide further methodological details on our satisfaction-related questionnaire items, while in Section 4.5 we will present the findings of the study. A discussion of the findings will be provided in Section 4.6 before concluding with a concise summary of the chapter (Section 4.7).

## 4.2 An introduction to the concepts of job and task satisfaction

This section will serve as a theoretical introduction to our study of linguists' task satisfaction. We will first analyse how the concepts of job and task satisfaction are defined and measured in the field of psychology. We will also depict some of the most used questionnaires of job satisfaction, with a focus on those that deal with task satisfaction. Afterwards, we will review studies in the T&I domain that deal with translators' professional satisfaction, including satisfaction with translation technology.

### 4.2.1 Job and task satisfaction in industrial-organisational psychology

#### Job satisfaction

In the field of industrial-organisational (I/O) psychology – that is, the branch of psychology concerned with the study of human behaviour at the workplace [3] – job satisfaction is by far the most widely studied construct (Judge et al., 2017; Spector, 2022). In its almost 100 years of history, the research on job satisfaction has gone through different epochs, and the definition itself of job satisfaction has evolved accordingly [4]. Recently, Judge, Zhang, and Glerum (2021, p. 210) have defined job satisfaction as an "evaluative judgment of one's job ranging from positive to negative" – thus considering satisfaction and dissatisfaction as opposite ends of a bipolar continuum (Locke, 1969). This definition suggests that job satisfaction is an *attitude*, i.e. an evaluation of an entity which can have a positive or a negative direction and different magnitude (Eagly and Chaiken, 1993; Wagner, 2021).

3. `https://www.apa.org/ed/graduate/specialize/industrial`. Accessed 15 August 2022.
4. For a more detailed discussion of the topic, see Judge et al. (2017) and Judge, Zhang, and Glerum (2021)

As Bowling and Sessa (2021, pp. 6–8) have pointed out, the study of job attitudes[5] – job satisfaction, in particular – is paramount for several reasons. First and foremost, since work is a central part of one's life, job satisfaction is likely to contribute to individual life satisfaction (Bowling, Eschleman, and Wang, 2010). Indeed, as noted by Dalal (2013, p. 341), "[a] satisfying job can provide meaning to life and be a source of self-worth; a dissatisfying job can be intolerable and a source of sleepless nights". Second, measuring employees' job satisfaction provides diagnostic information regarding certain aspects of the work or of the organisation that need improvement. Third, job (dis)satisfaction can lead to various job-relevant behaviours, such as performance, absenteeism and turnover (Judge et al., 2001; Kim and Kao, 2014; Schaumberg and Flynn, 2017; Rubenstein et al., 2018). However, although there is a generally accepted idea that satisfied employees would also perform better, it should be noted that studies have found only inconsistent or modest correlations between job satisfaction and job performance (Judge et al., 2001; Fisher, 2003).

Scholars have investigated job satisfaction among various professional categories, such as teachers (De Simone, Cicotto, and Lampis, 2016; Troesch and Bauer, 2017; Toropova, Myrberg, and Johansson, 2021), nurses (Lu, Zhao, and While, 2012; Dilig-Ruiz et al., 2018; Waltz et al., 2020), lawyers (Hagan and Kay, 2007; Markovic and Plickert, 2018; Monahan and Swanson, 2019) and police officers (Yun, Hwang, and Lynch, 2015; Chen, 2018; Lambert et al., 2018), finding that different factors affect job satisfaction in different professions.

### Task satisfaction

One of the most frequently investigated aspects of job satisfaction is "satisfaction with the nature of the work itself", also defined as "satisfaction with intrinsic job characteristics" (Saari and Judge, 2004; Smith, Kendall, and Hulin, 1969; Weiss et al., 1967). This is often considered the most important aspect and the one that better correlates with overall job satisfaction (Jurgensen, 1978; Ironson et al., 1989; Saari and Judge, 2004; Judge and Kammeyer-Mueller, 2012). As Locke (1969, p. 330) correctly pointed out, a job is "an abstraction referring to a combination of tasks performed by an individual [for remuneration]". Therefore, when we assess satisfaction with the work itself, we are indeed assessing satisfaction with these tasks.

In Human Resource Management, particularly in the job analysis literature, a task can be defined with various levels of specificity. Indeed, it can be viewed either as "a distinct work activity carried out for a distinct purpose" (Cascio and Aguinis, 2005, p. 212) or as "a collection of activities that are directed toward the achievement of specific job objectives" (Morgeson, Brannick, and Levine, 2019, p. 36). In the first definition, an example of task is "typing an email", while the latter definition applies to tasks such as "writing a literature review", which includes various activities, such as defining the scope, reading a number of papers and summarise them. Albeit the job analysis literature defines task and job as

5. Other relevant job attitudes include *job involvement* – "the degree to which people psychologically identify with their jobs" – and *organisational commitment*, or "the extent to which employees define themselves by the same characteristics that define [their] organisation" (Robbins and Judge, 2022, p. 66)

two distinct concepts, early scholarly literature in I/O psychology has often used the terms job design and task design interchangeably (Salancik and Pfeffer, 1978; Griffin, Welsh, and Moorhead, 1981). Indeed, when investigating the motivational features of a job, I/O scholars have rather referred to the characteristics of the tasks performed on the job (Oldham and Fried, 2016).

Several researchers formulated theories on the task characteristics that would motivate employees and suggested various methods to measure these characteristics. One of the first theories of this kind is Herzberg's Two-Factor Theory (1959) [6]. This theory — also known as *Motivation-Hygiene Theory* — posits that jobs should be enriched, rather than simplified, to motivate employees to perform well. According to Herzberg, job satisfaction depends on intrinsic factors (also called *motivators* or *satisfiers*), such as the nature of the job itself, while dissatisfaction is attributable to extrinsic job factors (defined *hygiene factors* or *dissatisfiers*), such as working conditions, salary and supervision. The underlying idea is that job-related satisfaction and dissatisfaction do not pertain to the same continuum, as the opposite of satisfaction is rather no satisfaction. Therefore, factors that lead to satisfaction are different from factors that lead to dissatisfaction. Hygiene factors can prevent employees to be dissatisfied, but an optimal job context alone cannot provide satisfaction *per se*. Only motivators can lead to true satisfaction and to positive job attitudes, "because they satisfy the individual's need for self-actualization in his work" (Herzberg, Mausner, and Snyderman, 1959, p. 114). Indeed, as the authors explain, "[it] is only from the performance of a task that the individual can get the rewards that will reinforce his aspirations" (*ibid.*).

One of the most influential theories of job design to date is Hackman and Oldham's Job Characteristics Model (1976), which considers job satisfaction as one of the outcomes of intrinsically motivating characteristics of one's job. The authors argued that each job has a motivating potential that derives from five "core" characteristics (as shown in Figure 4.1), namely *Skill variety*, *Task identity*, *Task significance*, *Autonomy* and *Feedback*. A motivating job is one that (1) involves various tasks and requires a variety of skills (Skill variety), (2) includes tasks that let employees work on the whole product, rather than a small part of it (Task identity), (3) includes tasks that have an impact on the lives or work of others (Task significance), provides employees with both (4) enough freedom to schedule their own work and processes (Autonomy), and (5) a Feedback on their work.
In particular, these five core characteristics can induce three psychological states that, in turn, lead to various positive job-related outcomes, such as satisfaction, motivation and high-quality performance. Skills variety, Task identity and Task significance together induce Experienced meaningfulness of the work; Autonomy triggers Experienced responsibility for the outcomes of the work, and Feedback leads to Knowledge of the results. These psychological states induce positive personal and work outcomes. Additionally, the theory

6. This theory has inspired several investigations in the domain, but has also been widely criticized as having many flaws (Locke, 1969). Indeed, many successive studies have failed to replicate Herzberg's results with different research methods, showing that intrinsic and extrinsic factors can contribute to both satisfaction and dissatisfaction (Ewen, Smith, and Hulin, 1966; Wernimont, 1966; House and Wigdor, 1967; Dunnette, Campbell, and Hakel, 1967; Hulin and Smith, 1967).

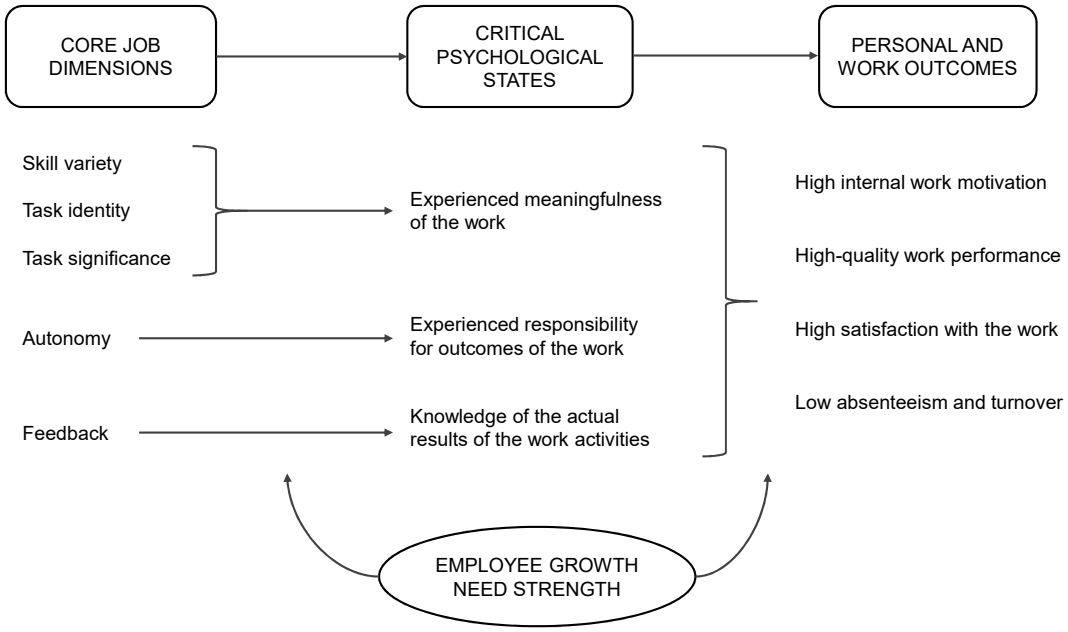



FIGURE 4.1 – Job Characteristics Model, retrieved from Hackman and Oldham, 1976.

considers that employees' need for growth and development at work (named *growth need strength*, GNS [7]) moderates [8] both the relationship between core characteristics and psychological states, and between the latter and outcome variables. In other words, people with high GNS will be more responsive to high motivating jobs than people with lower GNS. The motivating potential of each job can be calculated as a score (*motivating potential score*, or MPS), using a formula [9] that combines the five core characteristics, assigning more weight to Autonomy and Feedback. The theory has received extensive support in the literature (Loher et al., 1985; Fried and Ferris, 1987; Humphrey, Nahrgang, and Morgeson, 2007; Blanz, 2017) and inspired a great deal of research on how task characteristics contribute to employees' satisfaction. Most importantly, further studies on the topic have focused almost exclusively on the restricted number of task characteristics defined by Hackman and Oldham.

Studies on task satisfaction have mostly pursued two aims: either (i) defining the link between task satisfaction and task performance or (ii) clarifying the link between task satisfaction and overall job satisfaction. Regarding the first aspect, researchers have reported mixed results (Griffin, Welsh, and Moorhead, 1981). Some scholars have suggested that task performance induces task satisfaction (Judge et al., 2001), but other studies disconfirmed this link. For instance, Marley and Kersting (2020) have found that the level of detail

7. The idea that GNS would determine the extent to which employees would respond positively to jobs that score high in the four characteristics has been previously introduced in the work of Hackman and Lawler (1971).

8. A moderating variable affects the strength and direction of the relationship between two variables (Baron and Kenny, 1986).

9. MPS = [(Skill variety + Task identity + Task significance)/3] x Autonomy x Feedback

provided by the feedback on individuals' performance affects task satisfaction, regardless of the actual performance – which seems not to be linked with self-reported task satisfaction. Nevertheless, since in their experiment the participants were asked to perform a mundane and repetitive task, the authors note that these results are not generalisable to more creative tasks. Regarding the second aspect, although some studies have shown that task satisfaction affects overall job satisfaction only partially (Ironson et al., 1989; Wong and Campion, 1991), other scholars have used task satisfaction and job satisfaction interchangeably. For instance, Li, Wen, and Hsieh (2021, p. 4) have assumed that task satisfaction is "a reasonable simplification of [job satisfaction]".

Scholars investigated the link between task and job satisfaction focusing on a restricted number of task characteristics. For instance, Taber and Alliger (1995) found that task enjoyment and task complexity are the variables that most affect job satisfaction. Wong and Campion (1991) claimed that not only task characteristics but also task interdependence has an influence on employees' motivation. Their results show a moderate, positive relationship between the motivational value of individual tasks and the motivational value of the job overall. An inverted-U relationship has been found for task interdependence, meaning that only medium task interdependence allows for higher motivation compared to jobs with high or low task interdependence. These results show that too many unrelated tasks may decrease meaningfulness of the job, while too much interdependence among tasks would reduce variety, leading to low motivation. Furthermore, Dodd and Ganster (1996) found that perceived task variety and perceived task autonomy correlate more highly with job satisfaction than their objective counterparts.

### Measurement instruments of job and task satisfaction

Job satisfaction is usually measured through self-administered questionnaires — also called *instruments* or *measures* (Hubley and Zumbo, 2013) — as this constitutes the simplest, cheapest and fastest way to collect employees' attitudes towards their jobs [10]. It can be measured either as a global attitude toward the job or as the combination of attitudes toward different aspects – or *facets* – of the job [11] (Smith, Kendall, and Hulin, 1969; Spector, 1985; Judge et al., 2017; Judge, Zhang, and Glerum, 2021). Some of the most common job

10. Interviews are another effective method, and one that provides often more insights into the topic, but they are also more expensive and time-consuming. Nevertheless, interviewing a small sample of employees in an organisation can be a first step to select the aspects of the job that are worth investigating in a survey (Spector, 2022).

11. Several scholars who investigated job satisfaction with a faceted approach have argued that overall satisfaction corresponds to the sum of facets evaluations (Locke, 1969; Spector, 1985). This assumption has long been debated, since summing the facets means considering that such facets are all equally important to the respondent, and this is unlikely to be the case (Ironson et al., 1989; Dalal, 2013). However, recent work by Bowling and Zelazny (2022) compared several global and composite scales of satisfaction and found that they correlate well with each other in terms of construct validity. Therefore, as noted by Spector (2022, p. 16), "although the sum of facets my not be a perfect substitute for a global instrument, it can be a close approximation". In the end, the choice of focusing on global or facet measurements of job satisfaction depends on the research questions and on the goals of the study (Judge, Zhang, and Glerum, 2021).

satisfaction facets include satisfaction with pay, supervision and work tasks. Furthermore, facets of job satisfaction can be categorised into more specific sub-facets, such as satisfaction with pay levels or pay raises [12] (Heneman and Schwab, 1985).

Assessing overall job satisfaction does not necessarily mean asking participants a single question, such as *How satisfied are you with your current job?*. Both global and facet satisfaction instruments can make use of *single* or *multiple items*, i.e., statements to which respondents can agree or disagree to various extents [13]. These approaches have their own advantages and drawbacks.

A multiple-items approach allows for internal consistency testing (Ironson et al., 1989), as statements can be rephrased to verify that respondents express their attitudes toward a particular construct in a consistent manner [14]. Should the respondents misinterpret a question or select the wrong option when filling out the questionnaire, the presence of multiple statements will mitigate the effect of inconsistent responses (Spector, 2022, p. 28). However, using multiple items also increases the length of the questionnaire and, therefore, the risk of participant dropout (Hoerger, 2010). Conversely, single-item measures represent a faster way to administer questionnaires to large samples while also minimising respondent burden (Crawford, Couper, and Lamias, 2001). Their major criticisms are that single-item measures may not capture the content domain of complex constructs and that they do not allow for any calculation of internal consistency reliability (Loo, 2002; Fisher, Matthews, and Gibbons, 2016). While multiple-item measures are reported to be more accurate and more reliable than single-item measures for assessing overall job satisfaction (Oshagbemi, 1999; Judge and Klinger, 2008), the latter have been found to contain more face validity and to correlate well with their respective multiple-items measures of job satisfaction (Wanous, Reichers, and Hudy, 1997). Furthermore, single-items measures for facet satisfaction instruments have been found to correlate significantly with their corresponding facet, thus representing a more efficient and cost-effective way to measure satisfaction with specific job facets (Nagy, 2002; Lepold et al., 2018).

In the long history of job satisfaction research, various measurement instruments have been developed by scholars and practitioners alike to assess employees' satisfaction. Most

12. Questionnaires that focus on various job-related facets (facet instruments) provide a more detailed picture of individual's job satisfaction or dissatisfaction than global satisfaction measures (Rutherford et al., 2009). For instance, workers can express satisfaction on some aspects of the job and dissatisfaction on others. Thus, facet instruments constitute a preferred choice when an organisation needs to know which areas are to improve or which actions should be undertaken to increase employees' satisfaction or avoid withdrawal behaviour. At the same time, facet instruments have been often criticized as the selection of facets could omit or not address what is really important to the worker, thus leading to erroneous conclusions on one's job satisfaction (Scarpello and Campbell, 1983; Dalal, 2013).

13. For example, the Job in General Scale (Ironson et al., 1989) is a global satisfaction, multiple-items instrument that includes 18 adjectives or short sentences that respondents can select to describe their jobs. A well-known facet instrument, the short form of the Minnesota Satisfaction Questionnaire (Weiss et al., 1967), is a single-item instrument, since it employs only one statement for each of the 20 facets of job satisfaction investigated.

14. For instance, the Job Satisfaction Survey (Spector, 1985) lets respondents assess their satisfaction with pay, using both positively-worded statements, such as *I feel I am being paid a fair amount for the work I do*, and negatively-worded statements, as in *I feel unappreciated by the organisation when I think about what they pay me*.

of them include an assessment of satisfaction with the tasks performed on the job. In what follows, we will describe some instruments among the most widely used in I/O literature (Hora, Júnior, and Souza, 2018; Spector, 2022). We will present them in their chronological order of appearance and we will focus on the questions or the statements related to our subject matter, i.e., satisfaction with the work itself. These instruments have provided inspiration to design our own series of questions on linguists' satisfaction.

**Minnesota Satisfaction Questionnaire (Weiss et al., 1967)** One of the first measurement instruments of satisfaction with different job facets is the Minnesota Satisfaction Questionnaire (MSQ), which measures employees' satisfaction with 20 different aspects of their work. Respondents can answer via a five-point scale, ranging from "very dissatisfied" to "very satisfied", with "neither dissatisfied or satisfied" as the middle option. The questionnaire is available in both a long and a short version. The former includes 100 statements, while the shorter one focuses on 20 statements, one per facet. The 20 facets included in the MSQ are the following: ability utilisation, achievement, activity, advancement, authority, company policies, compensation, co-workers, creativity, independence, moral values, recognition, responsibility, security, social status, social service, supervision/human relations, supervision/technical, variety, working conditions. As noted by Spector (2022, p. 23), the nature of the work itself is expressed in various facets, including ability utilisation, creativity, achievement, activity, independence and variety.

| *"On my present job, this is how I feel about..."* | |
|---|---|
| **Item** | **Facet** |
| *The chance to do something that makes use of my abilities* | Ability utilization |
| *The chance to try my own methods of doing the job* | Creativity |
| *The feeling of accomplishment I get from the job* | Achievement |
| *Being able to keep busy all the time* | Activity |
| *The chance to work alone on the job* | Independence |
| *The chance to do different things from time to time* | Variety |

FIGURE 4.2 – Statements included in the Minnesota Satisfaction Questionnaire (Weiss et al., 1967, short form), referring to satisfaction with the work itself.

**Job Descriptive Index (Smith, Kendall, and Hulin, 1969)** The Job Descriptive Index (JDI) is a 72-items instrument that measures satisfaction with five job facets, namely pay, opportunities for promotion, coworkers, supervision and the work itself. As such, it consists of five sections, each one including a main question and a list of either 9 or 18 adjectives or short sentences. Respondents can answer with "Yes", "No" or "?", if undecided. For instance, the section investigating attitudes towards the work itself, contains adjectives such as "boring", "challenging", "creative", but also sentences such as "[the job] uses my abilities". An example of question and corresponding response options is shown in Figure 4.3. The authors of the JDI themselves do not recommend summing the five facet scores into a global satisfaction score, thus recognising that the whole is more complicated than the sum of its parts (Ironson et al., 1989).

| "Think of the work you do at present. How well does each of the following words or phrases describe your work?" | | | | | |
|---|---|---|---|---|---|
| Fascinating | Routine | Satisfying | Boring | Good | Gives sense of accomplishment |
| Respected | Exciting | Rewarding | Useful | Challenging | Simple |
| Repetitive | Creative | Dull | Uninteresting | Can see results | Uses my abilities |

FIGURE 4.3 – Example of question included in the Job Descriptive Index (Facet "Work on present job"), revised version (Bowling State University, 2009).

In 1990, the JDI has incorporated the "Job in General" scale (JIG, Ironson et al., 1989) -– a companion scale focusing on overall job satisfaction -– bringing the total number of items to 90. An abridged version of JDI (AJDI, Stanton et al., 2001), featuring only 38 items (including some items of the "Job in General Scale"), is also available. The questionnaire has been updated several times and its most recent version has been published in 2009 (JDI, 2009).

A few criticisms have been raised towards both the JDI and the MSQ, mainly due to the fact that these instruments only use non-parallel items and that selected facets could not represent what is really important for the respondent (Judge, Zhang, and Glerum, 2021; Spector, 2022). A recently-developed questionnaire, the Facet Satisfaction Scale (Beehr et al., 2006; Bowling, Wagner, and Beehr, 2018) tries to address the limitation of these two questionnaires, assessing satisfaction with the JDI facets (pay, promotion, supervision, coworkers, and the work itself) using five parallel items per subscale. In other words, the statements of satisfaction are worded in the same manner (Figure 4.4).

*Overall, I am very pleased with the types of activities that I do on my job.*
*Overall, I am very pleased with the way my manager supervises me.*
*Overall, I am very pleased to work with my coworkers.*
*Overall, I am very pleased with how much money I earn.*
*Overall, I am very pleased with my opportunities for promotion.*

FIGURE 4.4 – Example of parallel items included in the Facet Satisfaction Scale, retrieved from Bowling, Wagner, and Beehr (2018).

**Job Diagnostic Survey (Hackman and Oldham, 1975)** The Job Diagnostic Survey (JDS) was developed to investigate the characteristics of a job and their effect on employees. Its theoretical framework is rooted in the Job Characteristics Theory (Hackman and Oldham, 1976, 1980). The questionnaire is made up of seven sections (plus one section that

contains demographic questions) that employ different subscales to assess the constructs associated with the main theory.

The first two sections of the questionnaire include questions and statements aimed at describing the nature of the work that the respondents perform. The items point at the five "core" dimensions of autonomy, skill variety, task identity, task significance and feedback from the work itself, plus two additional dimensions, namely feedback from agents and dealing with others, that are deemed "helpful in understanding jobs and employee reactions to them" (Hackman and Oldham, 1974, p. 9).

Sections three to five include measures of the three critical psychological states and positive outcomes, such as general satisfaction, internal work motivation and satisfaction with various aspects of the job – namely job security, pay, co-workers, supervision and opportunities for growth. Finally, sections six and seven assess to what extent respondents would like their job to possess some characteristics, such as friendly co-workers, quick promotions or opportunities for personal growth. An example of items and relative constructs is shown in Figure 4.5. Considering the noticeable length of the original version, the authors developed a short form of the main questionnaire that can be completed in ten minutes. Although the instrument has shown good reliability and validity and has been extensively used in the literature, its main criticisms is that the model does not provide a comprehensive description of job characteristics with a motivational power (Fried and Ferris, 1987). Alternatives such as the Job Characteristics Inventory (Sims, Szilagyi, and Keller, 1976) or the Work Design Questionnaire (Morgeson and Humphrey, 2006) have also been developed to overcome the pitfalls of the JDS.

| Item | Construct |
|---|---|
| *The job provides me the chance to completely finish the pieces of work I begin.* | Task identity |
| *The job gives me considerable opportunity for independence and freedom in how I do the work.* | Autonomy |
| *The job is quite simple and repetitive* | Skill variety |
| *The job requires a lot of cooperative work with other people* | Dealing with others |
| *Supervisors often let me know how well they think I am performing the job* | Feedback from agents |
| *The job provides me the chance to completely finish the pieces of work I begin.* | Task identity |

FIGURE 4.5 – Example of statements included in the JDS (Hackman and Oldham, 1974).

**Job Satisfaction Survey (Spector, 1985)** The Job Satisfaction Survey (JSS) is a multidimensional tool that assesses global job satisfaction through nine facets, namely satisfaction with pay, promotion, supervision, coworkers, fringe benefits, contingent rewards (assigned to the employees for good performance), operating conditions (rules and procedures that the employee should follow), the nature of the work, and communication – intended as "how well management keeps employees informed" (Spector, 2022, p. 17).

Each facet contains four items to which respondents can agree or disagree (as shown in Figure 4.6). The strongest disagreement corresponds to one point, while the strongest

agreement assigns six points. The questionnaire contains a mix of positively and negatively worded statements, that is, statements that express satisfaction or dissatisfaction, respectively. Table 4.6 below shows the four items used to assess satisfaction with the nature of the work. Single scores are then summed to obtain an overall satisfaction score. The original version is in English, but a translation in more than 30 languages is available.

| Nature of the work |
| --- |
| *I sometimes feel my job is meaningless.* |
| *I like doing the things I do at work.* |
| *I feel a sense of pride in doing my job.* |
| *My job is enjoyable.* |

FIGURE 4.6 – Items pointing at satisfaction with the work itself in the JSS (Spector, 1985).

In this section, we have defined job and task satisfaction in I/O psychology and related fields, outlining the main instruments used to assess the satisfaction of different categories of workers. This analysis lays the groundwork for the development of a novel measurement instrument that we will use to assess professional linguists' task satisfaction. In the next section, we will examine how T&I studies have tackled the question of job-related satisfaction and which instruments have been used to assess translators' and interpreters' satisfaction.

### 4.2.2 Job and task satisfaction in Translation Studies

#### Job satisfaction

Though extensively studied in I/O psychology, the concept of job satisfaction has been long neglected in Translation Studies. Interpreters' satisfaction has been investigated more often than translators' satisfaction, with overall results showing that interpreters are generally highly satisfied with their job (Choi, 2007; Setton and Liangliang, 2009; Lee, 2017; Martikainen, Karkkola, and Kuittinen, 2018). Nevertheless, in the last decade, a few studies have appeared in what can be considered the fields of Translation Psychology (Jääskeläinen, 2012; Bolaños-Medina, 2016; Zhu, 2020) and Sociology of Translation (Inghilleri, 2005; Chesterman, 2006, 2017). Many of these studies have approached satisfaction as a secondary variable, focusing on principal constructs such as translators' status (Dam and Zethsen, 2016; Ruokonen and Mäkisalo, 2018), occupational stress (Courtney and Phelan, 2019; Korpal, 2021), or emotional intelligence (Hubscher-Davidson, 2016), among others.

One of the first studies to approach the concept of job satisfaction in translation, though only indirectly, is the one by Hermans and Lambert (1998) conducted among Belgian business translators to gather information on non-literary translators' status. Their interviews showed that translators were dissatisfied with their rates and workload and did not feel they made good use of their expertise. In contrast, in Katan's (2009, 2011) international survey of 890 translators, interpreters and students on perceptions of the role and status of

the profession, the researcher included a question on the overall satisfaction of the respondents. Answers to this question were very positive – a striking majority of respondents were "pretty" or "extremely" satisfied – despite respondents being aware of the many drawbacks of their profession, such as the lack of regulation and career structure. Their satisfaction was mainly attributed to the "inner joy" respondents experience when translating. This finding was echoed in Dam and Zethsen (2016), who analysed the narratives of 15 experienced agency translators — considered "the core of the profession" (*ibid.*, p.183) — to identify the sources of their job satisfaction. Translation was described as "a varied and stimulating job" (since it comprises different tasks, genres, tools), "an intellectual and creative challenge", "exciting and satisfying" and a "meaningful activity" (*ibid.*, pp. 180–182).

Translators' job satisfaction has mainly been investigated in survey-based studies conducted in specific geographical areas or among specific professional categories in the translation and localisation industry. For instance, Liu (2013) launched a survey among Chinese translators (n=193) to find out whether translators' visibility correlates with translators' job-related happiness. The researcher proposed an instrument to measure job-related happiness using Bordieu's (1986) concept of *capital* as theoretical framework of the study. The main idea is that happiness depends on the alignment between capital sought (understood as desired economic, cultural, social or symbolic goods) and the actual capital translators receive from the job. The researcher found that translators who are more visible (that is, those who interact directly with their clients and end-users) are also more satisfied with the capital obtained from their job. In the study, satisfaction is only seen as a component of translators' job-related happiness and it is not investigated as a construct *per se*. Therefore, as already noted by Dam and Zethsen (2016), the researcher does not provide much data on the actual satisfaction of the respondents, nor on its sources.

Satisfaction is also touched upon in Ruokonen and Mäkisalo (2018), who surveyed 450 Finnish translators from different cohorts. The study focused on the relationship between perceptions of status (dependent variable) and background factors, working conditions and job satisfaction (independent variables). The authors found that respondents perceived the status of the profession as middling, but their own professional status as high. Satisfaction was investigated in relation to income levels and through professional well-being indicators, namely experienced negative stress, external pressure to lower their quality standards (e.g. to meet deadlines) and intentions to quit the job. Results showed interesting differences among respondents. While satisfaction with income seemed to influence status perceptions of business and audiovisual translators, it did not seem to affect status perceptions of literary translators. Items pointing at professional well-being proved to have only limited influence on translators' perception of the status of the profession in general, while they all had a significant influence on the perception of the status of one's own work. Unfortunately, since the focus of investigation was on status, researchers did not report the percentages of translators that are indeed satisfied with their income, nor included the full set of results on professional well-being indicators. More recently, these findings have been compared with those of two other studies conducted in Finland, to gauge sources of translators' overall job satisfaction (Ruokonen, Lassus, and Virtanen, 2020). These were found to "[stem] from

translation itself, or task-related resources" (*ibid.*, p. 116; see also Virtanen, 2019). The general picture showed a cohort of satisfied or very satisfied professionals, who consider translation "a varied, challenging and meaningful job that [allows] control over one's work and [offers] opportunities for personal achievement and growth" (Ruokonen, Lassus, and Virtanen, 2020, p. 116).

In the UK and Ireland, Courtney and Phelan (2019) carried out a survey of professional translators (n=474) focusing on the relationship between occupational stress and job satisfaction. Their questionnaire contained only seven questions, two of which dealt with overall job satisfaction and sources of stress and satisfaction, respectively. Although the level of work-related stress was reported to be medium to high, and mostly due to poor remuneration and unfair treatment from agencies and clients (an issue also discussed in Moorkens and Rocchi, 2021), respondents reported to be mostly satisfied or highly satisfied with their job. Therefore, the authors concluded that "no relationship [can be] established between [occupational stress and job satisfaction]" (Courtney and Phelan, 2019, p. 110). Interestingly, it was also reported that "the growing interest in machine translation and CAT-tools is a particular source of occupational stress for many respondents" (*ibid.*). However, this aspect is not further investigated. Finally, respondents reported that their major sources of job satisfaction were flexibility and autonomy, and "the art of translation itself", i.e., the fact of creating new texts and producing high quality texts (*ibid.*, p. 109).

A similar construct was mentioned in Bednárová-Gibová (2020, p. 77) and called *linguistic hedonism*, or "the sense of pleasure [deriving] from working with words". Linguistic hedonism is one of the variables of *Happiness at work* (HAW), defined as translators' "experience of subjective well-being at work that involves contentment, positive assessment of aspects of their professional lives and preponderance of positive over negative feelings" (*ibid.*, p. 71). In two different studies, the authors investigated HAW of literary, sworn and institutional Slovak translators, and found that they often report "the ability to do quality translations" and "the increase in knowledge and skills" as their main sources of satisfaction [15] (Bednárová-Gibová and Madoš, 2019; Bednárová-Gibová, 2020). In another study carried out in the same geographical context, Bednárová-Gibová and Majherová (2021) surveyed agency (i.e., full or part-time salaried, 93 subjects) and freelance translators (i.e., those who work with direct clients, 84 subjects) about similarities and differences in perceived HAW. Their main hypothesis was that self-employed translators would have shown greater HAW than salaried translators. However, this was not confirmed by their findings, who showed statistically significant differences in respondents' answers only regarding satisfaction with income and linguistic hedonism (freelancers being more satisfied than agency translators). For both cohorts, a stimulating work environment was found to correlate positively with HAW, and once again, the ability to do quality translations was considered a major source of satisfaction.

15. Slovak institutional translators indicated salary as their primary source of satisfaction, while the "increase in knowledge and skills" and the "ability to deal with difficult translations" were the second and third most chosen categories, respectively (Bednárová-Gibová and Madoš, 2019, p. 235).

Similarly, the "chance to perform tasks which make use of the respondents' abilities" was reported as a major source of satisfaction in Piecychna (2019, p. 138), who investigated Polish sworn translators' job satisfaction, using an adapted, shorter version of the Minnesota Satisfaction Questionnaire (Weiss et al., 1967). In her results, the second most valued characteristic contributing to job satisfaction was the variety of tasks to carry out. This finding was echoed by Korpal (2021), who investigated the same cohort of Polish translators and found that the variety of topics that respondents encountered in their job was also the aspect they were most satisfied with. Job satisfaction levels among different categories of translators are also investigated by Moorkens (2020). The researcher used agreement to 13 statements to measure job satisfaction of freelance and salaried translators in Ireland. Other sections of the same questionnaire focused on respondents' attitudes towards language technology. Results showed that translators take pride in their work, but freelancers' perceptions of purpose, fairness in work, salary, colleagues and job security are more negative than those of directly-employed colleagues. Directly-employed translators are reported to be better disposed towards technology than freelance translators, who feel threatened by MT in particular. Unfortunately, the fairly small sample size – only 52 responses – and unbalanced categories (many more freelancers than salaried translators), did not let the researcher draw any significant conclusions. Furthermore, possible correlations between attitudes towards MT and levels of individual satisfaction are not discussed.

While several scholars have surveyed translators about their attitudes towards technology (García, 2006; Lagoudaki, 2008; Dillon and Fraser, 2007; Gough, 2011; Zaretskaya, 2015; Bundgaard, 2017b; Rossi and Chevrot, 2019; Salmi, 2021), the relationship between professional satisfaction and attitudes towards translation technology has received scant attention. One example is the study conducted by LeBlanc (2013, 2017), who investigated translators' attitudes towards TM, and the effects of TM implementation on translators' autonomy and professional satisfaction. Through semi-directed interviews and ethnographic observation in three Canadian LSPs, the researcher found that the use of TMs undermines professional satisfaction of in-house translators, since this technology changes the fundamental nature of their work. TMs limit translators' control over the target text; these tools force segmentation, obstacle creativity, put more pressure on individual productivity and require translators to reuse existing translations, thus making them "lazy and increasingly passive" (LeBlanc, 2013, p. 7). Most importantly, TMs are said to render the work more mechanical and to deprive translators of the texts' ownership, making skilled professionals work only with recycled segments. Finally, the interview showed that the lack of translators' involvement in crucial decisions about the workflow, like the introduction of new guidelines to work with TMs, was a major source of dissatisfaction. Therefore, the author claimed that the actual reason behind the translators' lamented disempowerment is a shift in business practice, rather than the use of the TM tool itself.

Marshman (2012, 2014) tackled the question from a different perspective and surveyed users of language technology about perceived effects of technologies on control over several job-related aspects, as an indicator of satisfaction with the technology. Among these aspects, the researcher included workload, tasks performed, quality of work, relationship with

clients/employers, remuneration and working methods (e.g. "working on paper or on screen, with or without tools, choosing the tools and settings used, or changing the approach to the text" (2014, p. 393).

Results showed that most respondents considered technology as an asset to language professionals, since it allows users to have greater control on their work. Factors reported to be the most influenced by language technology are the control over the quality of work done, over the amount of work and over working methods. Regarding the quality of their work, in particular, respondents stated they perceive both greater control, e.g. thanks to QA tools, but at the same time lesser control because they "[felt] forced to re-use inadequate or inappropriate solutions" or because of "problems relating to segmentation of texts" (Marshman, 2012, p. 6).

The study also highlighted interesting differences between subgroups; for instance, freelancers reported a greater (negative) influence of language technology over remuneration, while salaried translators were more concerned with control over the amount of work and over working methods. The author refers to a general term ("language technology") including translation memories, terminology management systems, concordancers, term extractors, localisation tools and MT. Therefore, since the survey does not focus on MT, we do not know the proportion of MT users among respondents. Overall, respondents felt they were much more in control of the quality of the texts and working methods, but in their comments, they cited TMs much more frequently than MT. At the same time, the fact that feeling forced to reuse segments contributes to perceptions of reduced control (thus leading to dissatisfaction) could be a serious issue when working with MT. Furthermore, regarding perceptions of control over the types of tasks performed, the analysis of respondents' comments showed that the increase in revision-related tasks is considered a source of dissatisfaction – an issue that could apply to both TM and MT.

#### Task satisfaction

Satisfaction related to performing a particular task has rarely been investigated in T&I studies. In a pilot study for his masters' thesis, Wallis (2008) found that two out of four student translators exhibit lower levels of satisfaction when working with fully pre-translated texts from TMs compared to when they use TMs in interactive mode — that is, consulting and choosing TM matches segment by segment. As an explanation, the author maintains that the pre-translated mode force translators to "adapt their style to that which is already contained in the [pre-translation], which means they will have less control over the creation of a holistic text" (*ibid.*, p. 626). Unfortunately, the very small number of participants does not allow to draw any definite conclusions. In fact, the other two participants expressed no difference in satisfaction when using the two working methods. Nonetheless, the hypothesis that translation working methods would affect satisfaction is worth further investigation. Based on these findings, we could assume that text pre-translation with MT, too, could affect linguists' task satisfaction.

Task satisfaction has been considered one of the facets of translators' satisfaction by Rodríguez-Castro (2011, 2015, 2016, 2019), who devoted her PhD thesis and subsequent

articles to the development and testing of the first model of satisfaction in T&I studies. The author applied Herzberg's Two-Factor Theory (detailed in Section 4.2.1) to the translation domain and identified intrinsic and extrinsic factors that could play a role in determining translators' satisfaction and dissatisfaction, respectively. Her model is made up of three facets without any hierarchical relationship (Rodríguez-Castro, 2011, p. 379), namely task satisfaction, job satisfaction and professional satisfaction. Each facet included a list of interrelated concepts and each concept, in turn, consists of interrelated properties, called factors, which are measurable using variables, i.e., questions or statements in the survey (Figure 4.7).

To test her model, the author developed a detailed questionnaire using a combination of scales, direct and indirect measurement of individual's satisfaction, as well as closed and open-ended questions. In this instrument, respondents could provide their level of satisfaction towards different items [16], using a five-point Likert scale ranging from "very satisfied" to "very dissatisfied". Some other items were evaluated in the form of agreement to various statements, using a corresponding Likert scale.


FIGURE 4.7 – Translator Satisfaction Model, adapted from Rodríguez-Castro (2011).

Among the three facets that compose the model, Job satisfaction is defined as "an individual affective reaction that reflects how the individual feels about the combination of routine tasks done [alone or in a team]" (*ibid.*, p. 125). Therefore, it encompasses extrinsic sources of satisfaction, i.e., those related to the work environment, such as workload

16. For each page of the survey listing various survey items, respondents were asked the question "Are you generally satisfied with the following?"

and team collaboration. In Rodriguez-Castro's thesis (2011), this facet comprises the concepts of individual-job fit, individual-organisation fit, individual upper-management fit and individual-client relationship and includes, among others, factors such as interpersonal relationships, client involvement and nature of the job, which is measured through the following survey items:

— *My job is stressful*

— *Flexible schedule*

— *Working overtime or weekends*

— *Policies for working remotely*

It is worth noting that this facet has undergone multiple restructuring in subsequent articles; in the more recent one (Rodríguez-Castro, 2016), the concepts of individual-upper management fit and individual-client relationship have been included as sub-concepts of individual-organisation fit. Professional satisfaction consists of both intrinsic and extrinsic job factors and "groups those concepts that are associated with the development of a sense of identity as a translator", that is, "a sense of belonging to a community" (Rodríguez-Castro, 2011, p. 162). This facet includes the concepts of professional profile, job security, occupational flexibility, self-concept and career development. Examples of statements associated with these factors are:

— *My role meets my professional expectations*

— *Opportunities for career development*

— *I am flexible to adjust to continuous changes*

The third facet in Rodríguez-Castro's model is Task satisfaction [17], defined as "the positive attitude experienced by a worker during, or upon completion, of a task" (Fisher, 1980). The author defines a task as "[a set] of routine activities conducted at a particular point in time during the enactment of a role" (Rodríguez-Castro, 2016, p. 203).

Figure 4.8 details the facet of task satisfaction and its related concepts of self-efficacy, nature of the task, self-fulfilment and job-fit, which are considered "strong intrinsic motivators" and compose together the *Task Satisfaction Index* (Rodríguez-Castro, 2015, p. 34). Self-efficacy is intended as "self-assessment of capabilities and of the power to influence outcomes and goals associated with the tasks being performed" (Rodríguez-Castro, 2011, p. 105). This concept is associated with the "feeling of knowing" the task – that is, task familiarity – and comprises the factors of task scope and task description.

The concept of nature of the task (also called *work itself*) — a core-concept that most literature on I/O psychology has associated with job-related satisfaction — is measured through four factors, namely task complexity, novelty of the task, variety of tasks performed, and more in general collecting data on tasks performed by survey respondents.

17. In the original model, task satisfaction is the first facet presented. However, we decided not to follow this order of presentation.

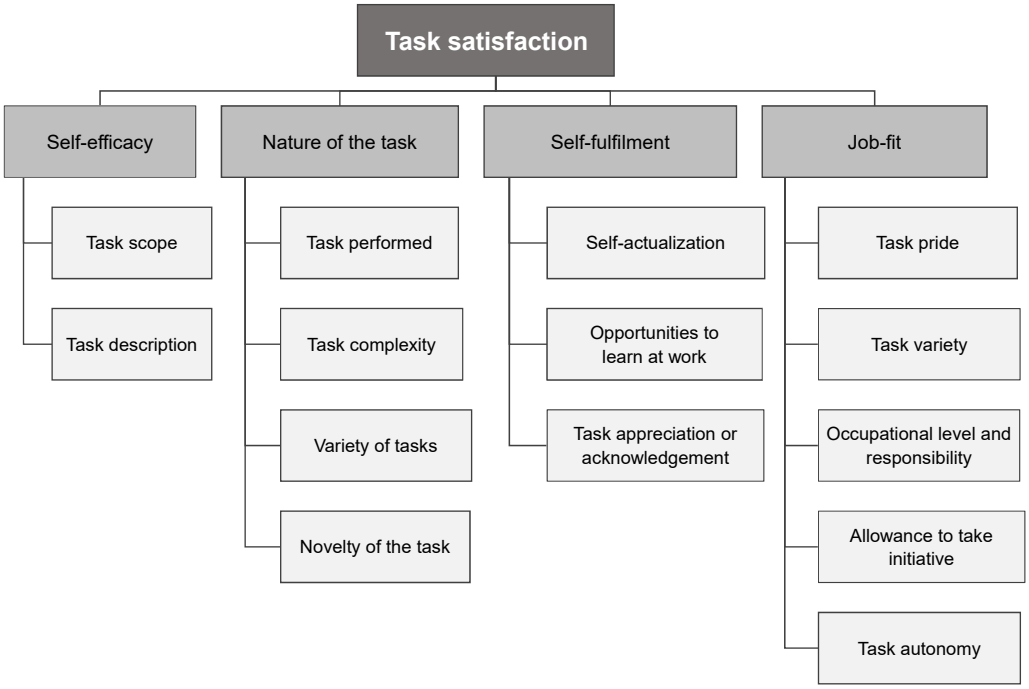



FIGURE 4.8 – Facet of task satisfaction, adapted from Rodríguez-Castro (2011).

Self-fulfilment represents satisfaction of individual needs and is determined by self-actualization (desire to grow and motivation for achievement), opportunities to learn at work (through feedback) and task appreciation (performance appraisal).

Finally, job fit is intended as "compatibility between an individual and the work setting" (Rodríguez-Castro, 2016, p. 203). It encompasses various characteristics, such as task autonomy (that is, the level of autonomy granted to the individual to make decisions concerning the task), occupational level and responsibility associated with experience and technical expertise, the opportunity to choose new tasks or to perform multiple tasks (task variety), allowance to take initiative to learn new processes and tasks, and the feeling of ownership and pride that derives from producing a new text (task pride).

Rodríguez-Castro tested her questionnaire with 250 translators from various countries that were reached through social networking platforms, industry contacts and international translators' associations. Results showed that task satisfaction is mainly attributable to the ability to perform a wide variety of tasks, intrinsic task pride and successful completion of projects (Rodríguez-Castro, 2016, p. 223), while no significant sources of dissatisfaction have been reported. The author argues that this latter finding "might be attributed to the fact that translation professionals are highly qualified individuals who enjoy their tasks [...]" (*ibid.*, p. 224). However, the question of whether translators enjoy the tasks they perform has not explicitly been included in the questionnaire, nor investigated further. A more recent paper from the same author (Rodríguez-Castro, 2019) described the sources of satisfaction and dissatisfaction between expert and novice translators, using the data collected in the main study. Results highlighted significant differences between the two groups, with experienced translators exhibiting higher levels of professional satisfaction than

novice translators.

While the relevance of this investigation does not need to be pointed out, the study presents several limitations. First, in Rodriguez-Castro's instrument, there are no questions of task satisfaction referring to specific tasks (e.g. revision or PE). Considering the variety of tasks and roles that language professionals can assume nowadays (e.g. a professional translator can also serve as reviser or post-editor in the same working day), we can expect different roles to induce different satisfaction levels. Second, the selection of survey items included in her Task Satisfaction Index is sometimes questionable. Indeed, some concepts seem to partially overlap, such as task autonomy and allowance to take initiative. Some statements have only a vague relationship with the concepts they are expected to represent – such as job turnover, measured with the statement "Overall satisfaction toward the current job", or self-efficacy, captured through "[I am satisfied with my] understanding of the scope of a task before undertaking the task". Additionally, some statements, such as "Clarity of task descriptions" or "Level of autonomy given to make decision", are rather ambiguous and could be misunderstood if no further information is provided – as it is the case in the questionnaire. Similar criticisms toward this work have been expressed by Mohammadi Dehcheshmeh (2017) who suggested some adjustments to Rodriguez-Castro's task satisfaction model. Third, the fact that some labels of concepts and factors associated with facets – and the number of facets itself – are inconsistent throughout the author's thesis and subsequent articles makes it difficult to describe her model. Finally, the length of the questionnaire is recognised by the researcher herself as a main drawback (Rodríguez-Castro, 2011, p. 376).

As we have seen from the review of the literature, while some research has been carried out on translators' satisfaction, far too little attention has been paid to translators' task satisfaction. Since no measurement instruments were adapted to research task satisfaction in professional translation contexts, we have set up a questionnaire for this purpose. In the next section, we will detail how this questionnaire was drafted. Afterwards, we will present the research hypotheses tested using our questionnaire.

Our work responds to Hubscher-Davidson's call (2018, p. 3) for the inclusion of more studies in the field of psychology of translation, which should "encompass [translators'] attitudes, personalities and dispositions". By focusing on professional translators' task satisfaction, our intent is to extend existing knowledge on translators' work-related attitudes and provide new research perspectives. If linguists enjoy what they do, introducing new tasks should add variety and new skills to their work, thus enriching rather than impoverishing their daily workflows. Studies have found that this is not always the case with MT and PE. Analysing how this task compares with other tasks — revision, in particular — in terms of satisfaction can help us reshape PE to address sources of (dis-)satisfaction and suggest new ways to introduce this activity in educational contexts, as well as in professional workflows.

To the best of our knowledge, the present work is the first survey of professional translators' attitudes towards specific translation-related tasks, and also the first conducted on a particular category of professional translators, namely professional salaried translators working at in-house corporate language services in Switzerland.

## 4.3 Designing a questionnaire on professional linguists' task satisfaction

Since there were no existing measurement instruments of task satisfaction that would fit our particular context, we developed for this purpose a number of questionnaire items relating to various aspects of linguists' task satisfaction. To do so, we have first reviewed the literature on professional satisfaction in I/O psychology and in T&I studies (Section 4.2) and selected the aspects (renamed *core concepts*) that would be relevant for our study. Statements referring to each core concept have been retrieved or adapted from existing measurement instruments of job satisfaction (Section 4.2.1) or formulated based on further investigation of the core concept itself. Additionally, we included a statement measuring overall task satisfaction. Table 4.1 schematises the core concepts used to develop the *Satisfaction section*.

| **Core concept** | **Retrieved or adapted from** | **Item(s) (question code)** |
|---|---|---|
| **Task-specific self-efficacy** | Rodríguez-Castro (2011); Konttinen (2021) | E27.01<br>E27.02<br>E27.03 |
| **Creativity of the task** | MSQ (Weiss et al., 1967);<br>JDI (Smith, Kendall, and Hulin, 1969);<br>Moorkens (2020) | E30.01<br>E32.01<br>E34.01 |
| **Task complexity** | JDI (Smith, Kendall, and Hulin, 1969);<br>Rodríguez-Castro (2011); JDS (Hackman and Oldham, 1974) | E30.03<br>E32.03<br>E34.03 |
| **Stimulating nature of the task** | Dam and Zethsen (2016); JDS (Hackman and Oldham, 1974) | E30.02<br>E32.02<br>E34.02 |
| **Control (over the text's final quality)** | Marshman (2012, 2014);<br>Rossi and Chevrot (2019) | E31.01<br>E33.01<br>E35.01 |
| **Ability utilisation** | MSQ (Weiss et al., 1967) | E31.02<br>E33.02<br>E35.02 |
| **Task identity** | JDS (Hackman and Oldham, 1974); Rodríguez-Castro (2011) | E31.03<br>E33.03<br>E35.03 |
| **Work-scheduling autonomy** | Morgeson and Humphrey (2006) | E29 |
| **Task variety** | Rodríguez-Castro (2011);<br>Morgeson and Humphrey (2006) | E28 |
| **Enjoyment (overall task satisfaction)** | JSS (Spector, 1985) | E26.01<br>E26.02<br>E26.03 |

Table 4.1 – Design of the Satisfaction section in questionnaire Q2.

Each core concept is expressed in the survey with one statement and applied to the three tasks of translation, revision and PE. Although our research focuses on the comparison between revision and PE, we also surveyed our respondents about their attitudes towards

translation. The latter, considered the core task of our profession (the very reason why one chooses to start a career as translator), serves as a baseline, and has been included to gather a more comprehensive picture of respondents' task satisfaction. In the questionnaire, we also specified that we meant translation as *translating without using MT*, to clarify the differences with PE. Post-editing has been defined as *working with MT*, to account for cases where MT is used as an additional suggestion during the translation process. In order to clarify what we meant by *revising*, we provided the ISO definition that describes revision as the "bilingual examination of source and target" carried out by a reviser (International Organization for Standardization, 2015, p. 2).

In the following sections, we will discuss the core concepts selected to design the Satisfaction section and present the related statements.

### 4.3.1 Core concepts

In this section, we present the main characteristics of the core concepts used to investigate linguists' task satisfaction (Table 4.1). Each core concept is examined from both the point of view of I/O psychology and Translation Studies. Coverage of these concepts varies inasmuch that some are well-researched in both fields while others have received less attention. The statements associated with each core concept are presented at the end of each subsection.

#### Task-specific self-efficacy

Self-efficacy refers to "beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" (Bandura, 1997, p. 3), or, simply put, individuals' self-perceptions of their own abilities (Wohn and Sarkar, 2018). Given its focus on perceived abilities rather than on actual abilities, self-efficacy differs from expertise, which represents "knowledge and actual skills", but comprises perceived expertise and perceived confidence in one's own abilities (*ibid.*, p. 2337). It also differs from competence, understood as the proven ability to use knowledge and skills (EMT Expert Group, 2017). Indeed, while competence assumes an external assessment, self-efficacy refers to self-assessed competence (Konttinen, 2022). Finally, self-efficacy improves with successful completion of tasks and at the same time is threatened by failures, particularly if these occur before one reaches a certain level of self-efficacy [18] (Bandura, 1997).

Scholars have conceptualized three levels of self-efficacy, depending on the context of applicability, namely general self-efficacy, domain-specific and task-specific self-efficacy (Grether, Sowislo, and Wiese, 2018). General self-efficacy (GSE) refers to "individual's perception of their ability to perform across a variety of situations" (Judge, Erez, and Bono, 1998, p. 170). Scholars have proposed various scales for the measurement of GSE that include statements such as "I can always manage to solve difficult problems if I try hard enough" and "It is easy

18. According to Bandura (1997), enactive mastery experiences – i.e., perceived success or failure of past experiences – are "the most influential source of efficacy information" (*ibid.*, p. 80). Other sources of self-efficacy are vicarious experiences, verbal persuasion, and physiological and affective states.

for me to stick to my aims and accomplish my goals" (Sherer et al., 1982; Schwarzer and Jerusalem, 1995; Chen, Gully, and Eden, 2001). While GSE has been found to correlate with task-specific self-efficacy (Sherer et al., 1982; Miyoshi, 2012), the construct has also been questioned, mainly for its lack of clarity regarding the object of assessment that could lead to different interpretations by different respondents (Bandura, 1997).

Domain-specific self-efficacy refers to "confidence in one's coping ability within a specific setting" (Grether, Sowislo, and Wiese, 2018, p. 132) and it is associated with job satisfaction and job performance [19] (Caprara et al., 2003, 2006; Klassen and Chiu, 2010; Federici and Skaalvik, 2012). An example of self-efficacy scale for a specific domain is the occupational self-efficacy scale proposed by Schyns and Collani (2002), which includes statements such as "I feel insecure about my professional abilities" and "I feel prepared to meet most of the demands in my job".

The third type of self-efficacy discussed in the literature refers to individuals' self-assessment of their abilities in performing specific tasks, such as solving anagrams or mathematics problems (Smith et al., 2006; Bonne and Johnston, 2016; Siefer, Leuders, and Obersteiner, 2021).

In the T&I domain, the concept of self-efficacy is relatively young. It has been investigated mainly for interpreters (Bontempo and Napier, 2011; Macnamara, 2012; Mashhady, Fatollahi, and Pourgalavi, 2015; Liu, 2021) and for translation students (Atkinson, 2014; Bolaños-Medina, 2014; Yang, Guo, and Yu, 2016; Haro-Soler, 2017; Haro-Soler, 2019; Konttinen, 2022), while professional translators' self-efficacy has attracted less attention.

Self-efficacy is a component of the freelance translators' psychological skills model developed by Atkinson (2012), who investigated the relationship between translators and interpreters' psychological skills and professional success. The author found that occupational self-efficacy was statistically significantly related to job satisfaction, but also to other measures of success, such as income and quantity of work.

Self-efficacy has also been associated with task satisfaction in Rodríguez-Castro's Translator satisfaction model (2011, 2019). Here, self-efficacy is associated with (translators' attitudes towards) task scope and task descriptions. The underlying idea is that self-efficacy is a characteristic of expert translators [20], who have been exposed to various tasks and therefore exhibit a better understanding of task scope and task descriptions. However, these constructs are expressed with statements such as "Task description shows accurate nature of work" and "Knowledge of CAT tools is essential to meet my deadlines" which bear apparently no direct relationship with self-efficacy. Therefore, rather that investigating self-efficacy as a variable affecting task satisfaction, it seems that Rodriguez-Castro simply focused on translators' satisfaction with their own perceived self-efficacy.

19. Studies have shown GSE and domain-specific self-efficacy are correlated (Sherer et al., 1982; Chen, Gully, and Kilcullen, 2000; Chen, Gully, and Eden, 2001; Schyns and Collani, 2002). Furthermore, GSE, too, has been found to correlate positively with job satisfaction and job performance (Judge and Bono, 2001).

20. Researchers have highlighted the links between self-efficacy and translation expertise (Ho, 2010; Muñoz Martín, 2014).

Translation self-efficacy has been defined as "individuals' beliefs in their competence to successfully perform the courses of action needed to produce an acceptable translation for a given translation brief" (Bolaños-Medina and Núñez, 2018, p. 55). The authors (*ibid.*) have also proposed the first translation-specific self-efficacy scale [21], that includes items such as "Recognising translation mistakes as a whole" and "Evaluating different alternative solutions for translation problems". While this scale is made up of translation-relevant items and shows good psychometric properties, it does not consider other translation-related tasks. More recently, Konttinen (2021) has developed a task-specific scale for measuring translation students' self-efficacy in translation company simulation courses. This scale touches upon various strategic, project management and support activities, as well as on translation-related tasks, including items referring to specific core-production tasks, such as "I am able to work as a translator in a translation project" and "I am able to work as a post-editor in a translation project".

As we have seen, self-efficacy plays a role in determining job satisfaction. Given that our focus of investigation is on translators' satisfaction in performing specific translation-related tasks, our survey includes task-specific items. In the present thesis, task-specific self-efficacy is defined as the extent to which individuals feel they own the necessary skills and knowledge to carry out specific translation-related tasks. This concept is expressed by the following statements:

E27.01 – *I have the necessary skills and knowledge to carry out translation jobs (without using MT).*

E27.02 – *I have the necessary skills and knowledge to carry out revision jobs.*

E27.03 – *I have the necessary skills and knowledge to carry out post-editing jobs.*

### Creativity of the task

Creativity can be defined as "the production of a novel and appropriate response, product, or solution to an open-ended task" (Amabile, 2011, p. 1). Research in social and I/O psychology has shown that engaging in creative tasks can enhance positive emotions (Amabile et al., 2005; Bujacz et al., 2016) and also that creativity influences the relationship between job satisfaction and job performance (Kato-Nitta and Maeda, 2013). Indeed, some of the most famous measurement instruments of job satisfaction in I/O psychology include this construct among their variables. For instance, creativity is one of twenty facets of job satisfaction defined in the MSQ (Weiss et al., 1967) and it is understood as "[t]he chance to try my own methods of doing the job" — although this definition considers creativity as a process and bears more similarities with the concept of autonomy. "Creative" is also one of the adjectives listed in the JDI (Smith, Kendall, and Hulin, 1969) that respondents can select to describe their own jobs. Furthermore, the JDS (Hackman and Oldham, 1974) mentions creativity in two different sections, as a job characteristic that a worker would need or wish to have in an

21. Interpreting-specific self-efficacy scales had been previously developed by Jiménez Ivars, Pinazo Catalayudb, and Ruiz i Forés (2014) and Lee (2014).

ideal job [22]. These statements are then used to determine the individual's need for growth and personal development.

In T&I studies, scholars and practitioners alike have put forward the value of creativity in translation (Sorvali, 1998; Kuznik and Verd, 2017; Malmkjær, 2020) and praised creativity as one of the main characteristics contributing to professional satisfaction (Dam and Zethsen, 2016; Piecychna, 2019; Moorkens, 2020). Nonetheless, many professional translators find that the use of translation technology in general (Ehrensberger-Dow and Massey, 2014; Bundgaard, Christensen, and Schjoldager, 2016; LeBlanc, 2017; Salmi, 2021) and of MT in particular (Moorkens and O'Brien, 2015; Cadwell, O'Brien, and Teixeira, 2017; Daems et al., 2017b; Sakamoto, 2019; Álvarez-Vidal, Oliver, and Badia, 2020; Guerberof Arenas and Toral, 2022) constitutes a hindrance to their creativity and freedom of expression. In other translation-related activities, such as revision, creativity has received far less attention. O'Brien (2012, p. 13) claimed that "[f]or many, editing is seen as a less creative task than translation". This assumption is also echoed in Mossop (2020, p. 226), who argued that "[r]evising the work of others is, for many people, not enjoyable because there is no opportunity for creativity". Creativity is also mentioned among the features that make up the psycho-physiological component of the reviser in the translation revision competence model proposed by Robert et al. (2017) and Robert (2018), but this construct is not investigated further.

Drawing on previous literature, we decided to verify to what extent CILS linguists in Switzerland consider translation, revision and PE as creative tasks and whether this would correlate with their level of satisfaction in performing these tasks. Therefore, we included the following statements in our questionnaire:

E30.01 – *Translation (without MT) is a creative task.*

E32.01 – *Revising human translations is a creative task.*

E34.01 – *Post-editing MT output is a creative task.*

### Task complexity

There is little consensus on the definition of task complexity in different disciplines or even within the same field. In their thorough analysis of the construct in industrial ergonomics, Liu and Li (2012) identify three points of view that researchers have adopted to define task complexity. First, if we consider the structure of a task, task complexity refers to the number of interconnected elements that compose the task. Second, we can define task complexity as resulting from the interaction between the task and the person who performs it; in this sense, we deal with perceived or subjective task complexity, as opposed to objective task complexity. Third, from the point of view of the resources required, a task is considered complex if it demands cognitive and physical abilities — in this sense, it is "synonymous with task load or task demand" (*ibid.*, p. 555).

22. These statements are: "Opportunities to be creative and imaginative in my work" (*ibid.*, p. 8) and "A job where there is considerable opportunity to be creative and innovative" (*ibid.*, p. 10).

Several studies have found that task complexity, defined as in the latter viewpoint, correlates positively with job satisfaction (Taber and Alliger, 1995; Dodd and Ganster, 1996; Schyns and Croon, 2006; Humphrey, Nahrgang, and Morgeson, 2007), particularly for workers that exhibit higher needs for personal growth and development (Hackman and Oldham, 1976; Loher et al., 1985). Furthermore, job complexity is considered a moderator in the relationship between job satisfaction and job performance (Judge et al., 2001). Task complexity is included in various measurement instruments of job satisfaction, often under the name "skill variety", such as in the JDS (Hackman and Oldham, 1974) where it is expressed with the item "The job requires me to use a number of complex or high level skills".

In her Translator Satisfaction Model, Rodríguez-Castro (2011) includes task complexity among factors that define the nature of the work itself. The author focused on a specific aspect of complexity in translation tasks, namely terminological complexity [23]. In her words, complex tasks in the translation domain are those that require "advanced knowledge and skills in order to deal with terminological complexity and domain specific usage" (*ibid.*, p. 111). She also includes a more generic statement to measure respondents' satisfaction with the opportunity to work with complex tasks and found that responses to this statement reveal a strong, positive correlation with overall professional satisfaction.

As noted by Bolaños-Medina (2014, p. 197), "it is a well-established fact that translation is a complex activity, involving not only various skills but also affective and attitudinal factors". The same can be said for revision, an activity that is usually reserved to more experienced professionals (Mossop, 2020; Riondel, 2021a) and that involves specific competences, many of which are shared with post-editing (Nitzke, Hansen-Schirra, and Canfora, 2019; Konttinen, Salmi, and Koponen, 2021; Robert, Ureel, and Schrijver, 2022). Nevertheless, the perception of post-editing as a complex activity has gathered different opinions. As reported by Konttinen, Salmi, and Koponen (2021, pp. 194–195):

> It is interesting to note that in previous studies, revision is often seen as a demanding task more suited to later stages of translator education and for more experienced translators [...], whereas post-editing is sometimes envisioned as a task for beginning translators or even non-translators familiar with the subject matter. [...] The different perceptions of reviser versus post-editor backgrounds may reflect cases where MT post-editing is associated with 'fit-for-purpose' or 'good enough' quality, emphasising the correctness of information content over fluent language.

Using NMT in everyday professional translation workflows, with different texts and quality levels (e.g. full PE), has unveiled the complexity of PE processes compared to revision ones (do Carmo, 2020). It remains to be determined, however, whether professional translators working with MT perceive PE as a complex task and whether this complexity plays a role in determining their task satisfaction. Following Liu and Li's categorization (2012), we adopt a "resource requirement viewpoint" for our own conceptualization of task complexity. We define a task as complex if it requires various skills and knowledge to be completed. Thus,

23. In a more recent paper (Rodríguez-Castro, 2019), items originally referring to task complexity appear associated with self-efficacy.

in designing our questionnaire on linguists' satisfaction in performing various translation-related tasks, we include the following statements referring to task complexity:

E30.03 – *Translation (without MT) is complex (i.e., it requires skills and knowledge).*

E32.03 – *Revising human translations is complex (i.e., it requires skills and knowledge).*

E34.03 – *Post-editing MT output is complex (i.e., it requires skills and knowledge).*

**Stimulating nature of the task**

We define an intellectually stimulating task as one that is intrinsically rewarding, challenging and engaging. This construct is intimately related to task complexity. Indeed, in the JDS (Hackman and Oldham, 1974), the adjectives "stimulating" and "challenging" are used in conjunction. This construct has even been used as a synonym for job satisfaction (Forsgren, Forsman, and Carlström, 2009) and investigated in I/O psychology as a factor that could potentially influence the relationship between satisfaction and performance (Baird, 1976; Kim, 1980; Wong and Campion, 1991).

Several studies in the T&I domain report that translation is perceived as a stimulating activity (Katan, 2011; Ehrensberger-Dow and O'Brien, 2015; Dam and Zethsen, 2016), but little is known on whether this adjective applies to revision and PE tasks, too. Therefore, we have included the following statements in our questionnaire:

E30.02 – *Translation (without MT) is stimulating.*

E32.02 – *Revising human translations is stimulating.*

E34.02 – *Post-editing MT output is stimulating.*

**Control (over the text's final quality)**

One of the factors that contributes to making a task meaningful and satisfying is the performer's perception of exerting some control on it. In translation-related tasks, this control extends over the text's final quality. Interviewees in Virtanen's study (2019) mentioned that translation is a meaningful job that allows control over the quality of one's work. However, Marshman (2012, 2014) found that the use of translation technology (particularly translation memories and QA features in CAT tools) affects users' perception of control over the quality of their work. While most participants reported that technology allows them to better control the quality of the final product, others claimed that text segmentation interferes negatively in this respect. This point is also echoed in Doherty (2016) and in LeBlanc (2013), where in-house translators mentioned that being forced to reuse matches from the TM limits their control over the target text, leading to "a sense of disempowerment" (*ibid.*, p.10).

Similar considerations have been made for PE. Rossi and Chevrot (2019) investigated how MT influenced perceptions of control among salaried translators working at the European Commission's Directorate-General for Translation (DGT). The majority of respondents feel that when using MT they have less control over their working methods, while their control over the quality of the target text seems to be affected only in part.

As for revision, to the best of our knowledge, this question has not been investigated. While revisers are often the last ones to check the quality of the text, they do not always have the last say on the changes to be made (Riondel, 2021a). Most importantly, revisers are instructed to keep their interventions to a minimum (Mossop, 2020) and this could be perceived as a diminished control over the quality of the final product.

To investigate linguists' perceptions of control over the final product in translation, revision and PE, we included the following statements:

E31.01 – *When translating without using MT, I feel I have control over the text's final quality.*

E33.01 – *When revising human translations, I feel I have control over the text's final quality.*

E35.01 – *When post-editing MT output, I feel I have control over the text's final quality.*

**Ability utilisation**

The construct of ability utilisation – also known as skill utilisation – refers to the extent to which an individual perceives their skills and knowledge being used effectively in their job role. Research within I/O psychology has found that this construct is one of the strongest predictors of job satisfaction (O'Brien, 1982a; O'Brien, 1982b; Humphrys and O'Brien, 1986; Burr and Cordery, 2001; Morrison et al., 2005; Tian, Wang, and Chia, 2018; Wang et al., 2020). This construct has been included in established instruments like the JDI (Smith, Kendall, and Hulin, 1969) with the item "[the] job uses my abilities" and the MSQ (Weiss et al., 1967) where it appears as one of the facets of job satisfaction.

Using the MSQ, Piecychna (2019) identified *ability utilisation* as one of the aspects of their job with which Polish sworn translators reported the highest satisfaction. Similarly, Bednárová-Gibová and Madoš (2019) found that the ability to handle demanding legal translations, together with the associated growth in knowledge and skills, were among the main sources of satisfaction for institutional Slovak translators.

To the best of our knowledge, the construct of ability utilisation has not been investigated in relation to specific tasks. Therefore, we have included the following statements in our questionnaire:

E31.02 – *When translating without using MT, I make good use of my skills and knowledge.*

E33.02 – *When revising human translations, I make good use of my skills and knowledge.*

E35.02 – *When post-editing MT output, I make good use of my skills and knowledge.*

**Task identity**

Dealing with an identifiable piece of work can enhance feelings of job satisfaction. This concept, named "task identity", has been included by Hackman and Oldham (1974) in the JDS and addressed the cases where employees could see the completeness of their work, e.g. compared to those who work on the assembly line. Indeed, a question in the JDS asks to what extent employees perform "a whole and identifiable piece of work [that] it is not a small

part of the overall piece of work, which is finished by other people or by automatic machines". As Morgeson and Humphrey (2006, p. 1323), citing Hackman and Oldham (1980) explain:

> Jobs that involve an intact task, such as providing a complete unit of service or putting together an entire product, are invariably more interesting to perform than jobs that involve only small parts of the task.

In Hackman and Oldham's model, task identity contributes to the experienced meaningfulness of the work, together with skill variety and task significance. Rodríguez-Castro (2011, p. 117) mentioned task identity as synonym of task pride. The latter concept is part of her Translator Satisfaction Model, under the facet *Task satisfaction*. It is operationalised as "taking pride in [one's own] work; improving overall project quality; feeling of great contribution to the team" (Rodríguez-Castro, 2016, p. 204).

We define task identity in translation-related tasks as the perception that the task lets the linguist create a new text. The following statements in the questionnaire point to the construct of task identity:

E31.03 – *When translating without using MT, I feel I am creating a new text.*

E33.03 – *When revising human translations, I feel I am creating a new text.*

E35.03 – *When post-editing MT output, I feel I am creating a new text.*

**Work-scheduling autonomy**

One of the most studied job characteristics is autonomy, defined as "the degree to which one has control and discretion over how to conduct one's job" (Judge and Klinger, 2008, p. 399). Several studies in I/O psychology have found that autonomy correlates strongly with job satisfaction (Loher et al., 1985; Finn, 2001; Skaalvik and Skaalvik, 2014) and with job performance (Morgeson, Delaney-Klinger, and Hemingway, 2005; Saragih, 2011). Autonomy is one of the job characteristics included in the JDS and has particular importance in calculating the motivating potential of the job (see Section 4.2.1). This concept has been further elaborated by Morgeson and Humphrey (2006, p. 1323) for their Work Design Questionnaire – a self-administered questionnaire for the assessment of job characteristics – which include various statements referring to three different aspects of autonomy, namely "freedom in work scheduling, decision making and work methods".

In her Translator Satisfaction Model, Rodríguez-Castro (2016, p. 204) defines *task autonomy* as the "level of autonomy given to make decisions", therefore using a statement that is quite generic and open to different interpretations. Surveys in our industry have shown that one of the aspects that freelancer translators enjoy the most is the chance to decide autonomously on their work schedule, including taking short breaks and choosing the tasks they wish to perform (Ehrensberger-Dow et al., 2016; Courtney and Phelan, 2019; Pielmeier and O'Mara, 2020). In this sense, salaried linguists working at language service providers have less autonomy, since it is usually the project manager that assigns them the tasks to perform, depending on the customers' needs.

Being forced to perform tasks that linguists do not appreciate can be frustrating. Even more so if the task involves working with a technology that constrains our agency. For

instance, Cadwell, O'Brien, and Teixeira (2017) reported that institutional translators were more open to using MT compared to translators working in an LSP, because the former could choose when to use this technology.

In our questionnaire, we included a statement to determine the level of work-scheduling autonomy of our respondents, which will provide us with additional information to interpret the results. The statement reads:

> E29 – *I can choose the tasks I want to perform on a daily basis (e.g. ask to perform more translation, rather than revision).*

**Task variety**

Task variety is defined as "the extent to which an individual performs different tasks at his or her job" (Humphrey, Nahrgang, and Morgeson, 2007, p. 1335). The concept has often-times been associated to — and confused with — *skill variety*, i.e., the extent to which a task or a job requires various skills to be performed. As such, it has been included in various measurement instruments of job satisfaction, e.g. in the JDS and in the MSQ.

Working in a language service usually involves performing different tasks depending on daily requirements. In this sense, the introduction of PE can be perceived as an element of disruption, which can be positive or negative depending on the individual personality and attitudes towards MT. To determine the level of task variety of our respondents, we introduced the following statement:

> E28 – *The introduction of machine translation in my workflow has helped to make my daily work more varied and stimulating.*

**Task enjoyment**

Finally, we included in our questionnaire a statement to gauge respondents' overall satisfaction in performing each of the three tasks investigated. Task enjoyment has been used as a proxy of task satisfaction in I/O psychology (Judge, Bono, and Locke, 2000; Li, Wen, and Hsieh, 2021), after several studies found that this was the variable that most influences satisfaction with the work itself (Taber and Alliger, 1995; Skaalvik and Skaalvik, 2014; Chen, 2018). Furthermore, the percentage of time spent on enjoyable tasks has been found to correlate positively and significantly with global satisfaction (Taber and Alliger, 1995). To the best of our knowledge, apart from the few studies on task satisfaction detailed in Section 4.2.2, the concept of task enjoyment has been neglected in our field.

The statements included in the questionnaire to measure respondents' overall task satisfaction are the following:

> E26.01 – *I enjoy translating (without using MT).*
>
> E26.02 – *I enjoy revising (human translations).*
>
> E26.03 – *I enjoy post-editing MT output.*

### 4.3.2 Research question and hypotheses

Following Rodríguez-Castro (2015), in this thesis, task satisfaction is understood as "the positive attitude experienced by a worker during, or upon completion, of a task". To answer our research question (**RQ2**: *In the professional workflow of Swiss CILS, how does PE compare to revision in terms of task satisfaction, for the linguist who performs these tasks*[24]*?*), we formulated several hypotheses ($H$) related to the core concepts defined in the previous section. To ease text readability, we will only indicate alternative hypotheses ($H_1$). The reader shall note that null hypotheses ($H_0$) will assume no effect or no relationship between the variables considered (Mellinger et al., 2017).

#### Task-specific self-efficacy

**H01.01**: *Trained post-editors will report higher levels of task-specific self-efficacy than post-editors who did not attend any PE training.*

**H01.02**: *Trained revisers will report higher levels of task-specific self-efficacy than revisers who did not attend any revision training.*

#### Creativity of the task

**H02.01**: *Post-editors who use MT as a suggestion in a separate window will find PE more creative compared to those who pre-translate the entire text, as this setting could constrain their creativity.*

**H02.02**: *Post-editors who use MT as an additional suggestion will find PE more creative compared to those who display only one suggestion at a time, as more suggestions would enhance their creativity.*

**H02.03**: *Post-editor whose CILS implemented light post-editing will find PE a lesser creative activity compared to those who do not deal with light PE, as this practice restrains their creativity.*

#### Task complexity

**H03.01**: *Since revision is usually reserved to more experienced professionals, it will be perceived as more complex than PE.*

24. As previously mentioned (Section 4.1), in this thesis, we understand as *tasks* the main activities carried out by professional translators in their daily workflows, namely translation, bilingual revision of human translation and post-editing of MT output.

**Stimulating nature of the task**

**H04.01**: *There is a positive correlation between perception of a task as complex and perception of a task as stimulating.*

**Control (over the text's final quality)**

**H05.01**: *PE working mode (use of MT suggestions) affects perceptions of control, as users of pre-translated texts will feel less in control during PE.*

**Ability utilisation**

**H06.01**: *There is a positive correlation between perceptions of a task as complex and respondents' perceptions of making good use of their abilities.*

**Task identity**

**H07.01**: *PE allows for a higher level of task identity than revision.*

**Task variety**

**H08.01**: *There is a positive correlation between perceptions of PE as a stimulating activity and perceptions of task variety on the job.*

**Task enjoyment (overall task satisfaction)**

**H09.01**: *PE working mode (use of MT suggestions) has an impact on the level of PE enjoyment.*

**H09.02**: *Younger translators (<30 years old) will enjoy PE more than mature translators.*

**H09.03**: *Linguists who value revision tasks as much as translation tasks will also enjoy PE, since they enjoy working with pre-translated texts.*

**H09.04**: *There is a positive correlation between MT system's perceived quality and PE enjoyment.*

**H09.05**: *Linguists who received PE training will enjoy PE more than those who did not receive any training in this task.*

**H09.06**: *Linguists who received revision training will enjoy revision more than those who did not receive any training in this task.*

(continues)

**H09.07**: *Linguists who declare having more autonomy in choosing their daily tasks will also enjoy PE more than those who do not have the same autonomy.*

**H09.08**: *There is a positive correlation between perception of a task as complex and level of enjoyment for that task.*

**H09.09**: *Post-editors who use a customised system will enjoy PE tasks more than those who use a generic system.*

**H09.10**: *Previous PE experience positively influences the level of PE enjoyment.*

**H09.11**: *Previous revision experience positively influences the level of revision enjoyment.*

## 4.4 Methods

In what follows, we will provide further information on how Section E – the portion of questionnaire Q2 that deals with linguists' task satisfaction – was designed and tested, and which methods were used for data analysis. Eighty-five participants answered all the mandatory questions of the Satisfaction section [25].

### 4.4.1 The Satisfaction section

Section E contained 26 statements to which respondents could answer using a five-point, bipolar Likert scale ranging from *Strongly disagree* to *Strongly agree*, with *Neither disagree nor agree* as a middle point. In addition, a *Cannot choose* option was included to allow for non-applicability of the statement. Four open-ended, optional questions were also included to elicit comments on aspects that linguists did or did not appreciate about working with machine translation and revising human translations, respectively.

All the statements were positively-worded and statements associated with the same construct had a parallel structure across the three tasks – that is, they used the same wording. While some scholars recommend using a mix of positively and negatively-worded statements to verify whether participants answered consistently (see for instance Spector, 2022), others advise against this practice. In fact, studies have found that negatively-worded items in surveys can also lead to a misunderstanding of the questions and to inconsistent results (Barnette, 2000; Chyung, Barkin, and Shamsy, 2018). Furthermore, our choice to only use parallel statements was made to maintain a parallel structure between survey items, thus avoiding a possible influence of different wordings in investigating the same construct across different tasks (Bowling, Wagner, and Beehr, 2018). Additionally, we decided not to add further (negatively-worded) statements in an effort to reduce the length of

25. For information on sampling and dissemination methods, as well as on the profile of the participants, the reader can refer to the previous chapter (Chapter 3).

the instrument and avoiding participant dropout (Hoerger, 2010). Indeed, task satisfaction was discussed in only one of five sections of the instrument, which was already rather long. Other methods were employed to shorten completion time, such as including filter questions that allow participants to skip statements that did not apply to their case.

### 4.4.2 Pre-test and reliability

Both questionnaire Q1 (Chapter 2) and Q2 have been pre-tested internally with several colleagues from the Faculty of Translation and Interpreting at the University of Geneva, many of whom are active freelance translators, or have some experience as in-house staff in an internal translation service. Additionally, two external professional translators – a project manager working in an Italian-based translation agency and a freelance translator – filled out questionnaire Q1 and Q2, respectively. Feedback from colleagues and external professionals lead to some amendments that improved the clarity of the original questionnaires.

Internal consistency reliability of the satisfaction scale and of the subscales referring to each task was assessed by means of Cronbach alpha coefficients, which were all above the minimum acceptable level of 0.75.

### 4.4.3 Data analysis

Responses to survey statements were coded in the SPSS software and analysed using descriptive and inferential statistics. Responses to Likert-scale questions were assigned a numerical value ranging from 1 (*Strongly disagree*) to 5 (*Strongly agree*); the higher the score, the higher the agreement toward the statement investigated. A score of 0 was assigned whenever a participant selected the *Cannot choose* option; these observations were excluded from further analysis.

Results were described with measures of central tendency, such as mean and standard deviation. Data were also analysed by means of hypothesis testing. Unless stated otherwise, all hypotheses are directional. T-tests were performed to verify whether the difference between two groups were statistically significant ($p<.05$). Analysis of variance (ANOVA) was employed to reveal a significant effect of an independent variable on a dependent variable when the former had more than two levels. Correlations among variables were performed using Spearman's rho, a preferred choice when dealing with ordinal data, such as Likert scale items (Saldanha and O'Brien, 2013). Lastly, a qualitative analysis was carried out on responses to open-ended questions, which were coded in a spreadsheet tool and categorised by theme.

## 4.5 Results

Responses to parallel questions will be presented per core concept and reported as percentages. Tables will also include the number of observations ($n$), mean values ($M$) and standard deviations ($SD$). For each core concept, we will test the hypotheses formulated in Section 4.3.2.

### 4.5.1 Task-specific self-efficacy

The questionnaire included a statement aimed at measuring linguists' perceived self-efficacy level in each of the three tasks. In other words, we wanted to verify to what extent linguists felt prepared to carry out translation, revision and PE. The statement read "I have the necessary skills and knowledge to carry out [this task]". Since several studies have reported that training plays a role in improving self-efficacy (Torkzadeh and Van Dyke, 2002; Nørgaard et al., 2012; Black et al., 2013; de Sousa Mata et al., 2021), we also considered the interaction between these two variables while formulating hypotheses.

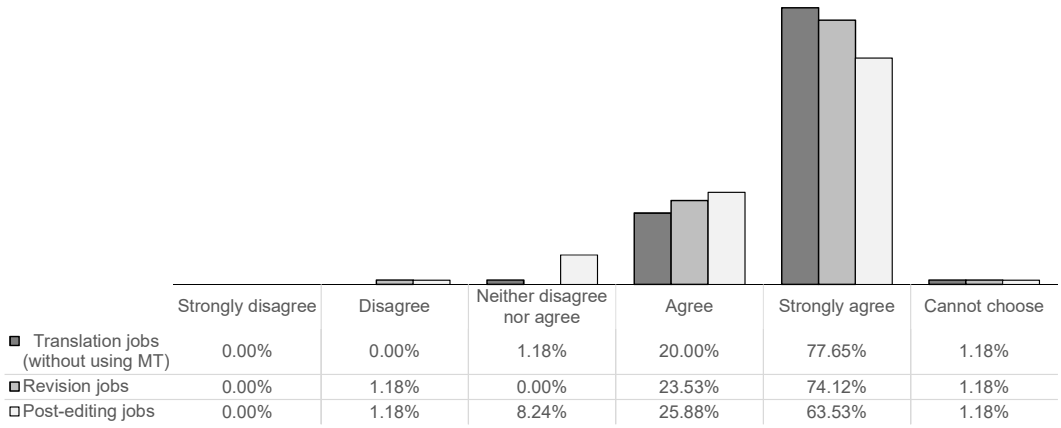



FIGURE 4.9 – Responses to statement *I have the necessary skills and knowledge to carry out...*

Figure 4.9 shows that respondents assume they all have the necessary skills and knowledge to carry out the three tasks. However, while 98% (n=83) of linguists indicate that they feel well prepared to carry out both translation and revision, some participants are slightly less confident regarding PE. Indeed, only 89% (n=76) of post-editors agree or strongly agree with the statement, while 8% (n=7) remain neutral. No one feels unfit to carry out translation jobs, but one respondent expressed disagreement in the case of revision and another one regarding PE jobs.

*H01.01: Trained post-editors will report higher levels of task-specific self-efficacy than post-editors who did not attend any PE training.*

| | *Self-efficacy in PE* | | |
|---|---|---|---|
| **PE training** | *n* | *M* | *SD* |
| Yes | 45 | 4.60 | .720 |
| No | 39 | 4.46 | .682 |
| Total | 84 | 4.54 | .702 |

TABLE 4.2 – PE self-efficacy scores, broken down per PE training.

We hypothesised that attending a PE training would enhance linguists' self-efficacy during PE. Therefore, we compared the means of the two subgroups (Table 4.2) and found that

post-editors who were trained to carry out PE jobs reported a slightly higher level of self-efficacy (n=45, M=4.60; $SD$=.720) compared to those who did not receive any training (n=39, M=4.46, $SD$=.682). We ran a t-test for independent samples to assess whether this difference was significant, but we failed to reject the null hypothesis ($t(82) = .901, p > .05$).

*H01.02: Trained revisers will report higher levels of task-specific self-efficacy than revisers who did not attend any revision training.*

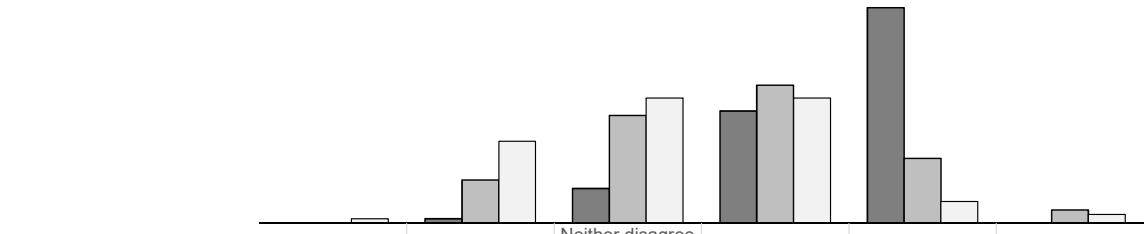
| | *Self-efficacy in revision* | | |
|---|---|---|---|
| **REV training** | *n* | *M* | *SD* |
| Yes | 34 | 4.88 | .327 |
| No | 50 | 4.62 | .602 |
| Total | 84 | 4.73 | .523 |

TABLE 4.3 – Revision self-efficacy scores, broken down per revision training.

We applied our hypothesis about the effectiveness of training in improving self-efficacy to revision tasks, too. Mean comparison (Table 4.3) showed a small difference between the two groups, with trained revisers reporting a higher level of self-efficacy compared to those who did not receive any training. Results of a t-test for independent means confirmed that this difference is statistically significant ($t(78) = 2.572, p = .006$).

### 4.5.2 Creativity of the task

We were interested in finding out to what extent participants considered post-editing and revision as creative activities. Mean values show that translation is considered the most creative among the three tasks investigated, with a mean of 4.47 out of five (n=85; $SD$=0.717), while revision and PE obtain a lower, very similar score. Revision scores on average 3.63 (n=82; $SD$=0.923), while PE has a slightly lower mean of 3.22 (n=83; $SD$=0.911). Figure 4.10 details these scores.



| | Strongly disagree | Disagree | Neither disagree nor agree | Agree | Strongly agree | Cannot choose |
|---|---|---|---|---|---|---|
| Translating (without MT) | 0.00% | 1.18% | 9.41% | 30.59% | 58.82% | 0.00% |
| Revising human translations | 0.00% | 11.76% | 29.41% | 37.65% | 17.65% | 3.53% |
| Post-editing MT output | 1.18% | 22.35% | 34.12% | 34.12% | 5.88% | 2.35% |

FIGURE 4.10 – Responses to statement *[This task] is a creative task.*

Most in-house linguists (89%, n=76) agree more or less strongly that translation can be considered a creative task when MT is not involved. Few translators do not position themselves (9.4%, n=8) and only one respondent disagrees. Revision, too, is considered a creative task, but to a lesser extent; in this case, 55% (n=47) agree or strongly agree,

29% of linguists (n=25) neither agree nor disagree, while 12% (n=10) simply disagree with the statement. Finally, only 34 out of 85 respondents (40%) agree more or less strongly that post-editing MT output requires creativity, but 34% of linguists (n=29) do not position themselves, and 22% (n=19) disagree. One respondent strongly disagrees with the statement, indicating that creativity is not a characteristic of PE.

For this core-concept, we formulated the following hypotheses:

*H02.01: Post-editors who use MT as a suggestion in a separate window will find PE more creative compared to those who pre-translate the entire text, as this setting could constrain their creativity.*

We hypothesised that dealing with a pre-translated text would constrain the creativity of respondents during PE, while having the opportunity to use MT as an additional suggestion in another window would rather enhance linguists' creativity.

| | *PE is a creative task* | | |
|---|---|---|---|
| **MT/CAT tool integration scenario (use of suggestions)** | *n* | *M* | *SD* |
| Pre-translation of the entire text (i.e., suggestion in the editing zone) | 46 | 3.35 | .900 |
| Suggestions in a separate window (i.e., the linguist clicks on a suggestion to use it) | 26 | 3.00 | .938 |

Table 4.4 – PE creativity score, broken down per MT/CAT integration scenario (use of suggestions).

Although mean values show a small difference between the two sub-groups (Table 4.4), a t-test for independent samples has shown no significant results ($t(70) = 1.552, p > 0.05$), meaning that we cannot extend our assumption to the entire population. It is worth noting, however, that most linguists use MT in pre-translation mode, while only 26 out of 85 linguists deal with MT suggestions in another window inside the CAT tool.

*H02.02: Post-editors who use MT as an additional suggestion will find PE more creative compared to those who display only one suggestion at a time, as more suggestions would enhance their creativity.*

We hypothesised that dealing with multiple suggestions would enhance linguists' creativity, rather than constraining it. In our sample, most linguists access MT only if no usable TM matches are available, while only 19 out of 85 respondents always use MT in addition to TM matches. When comparing the two sub-groups, we notice that the first one reports a slightly higher mean value than the second sub-group (Table 4.5), thus contradicting our hypothesis. A t-test for independent means shows that this difference is non-significant ($t(72) = .978, p > .05$).

| MT/CAT tool integration scenario (display of suggestions) | *PE is a creative task* | | |
|---|---|---|---|
| | *n* | *M* | *SD* |
| MT provided only if there are no TM matches or matches below a certain percentage | 55 | 3.29 | .936 |
| MT always added as an additional suggestion, regardless of whether there are already TM matches | 19 | 3.05 | .848 |

TABLE 4.5 – PE creativity score, broken down per MT/CAT integration scenario (display of suggestions).

*Hyp2.03: Post-editor whose CILS implemented light post-editing will find PE a lesser creative activity compared to those who do not deal with light PE, as this practice restrains their creativity.*

Finally, we formulated the hypothesis that light PE would contribute to picture PE as a task that requires low creativity. Mean comparison of our sub-samples confirm this hypothesis; post-editors who carry out light PE reported that PE is slightly less creative compared to those who only deal with full PE (Table 4.6).

| Light PE implemented | *PE is a creative task* | | |
|---|---|---|---|
| | *n* | *M* | *SD* |
| Yes | 19 | 3.16 | .765 |
| No | 64 | 3.23 | .955 |

TABLE 4.6 – PE creativity scores, broken down per light PE implementation.

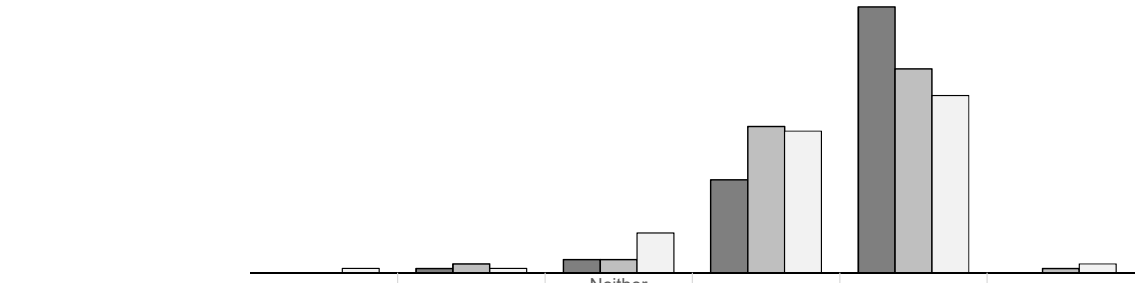
We tested this difference for its statistical significance using a t-test for independent means, but we could not reject the null hypothesis ($t(81) = .319, p > 0.05$). However, it is worth noting that in our sample, only 3 out of 23 CILS who participated in the survey implemented light PE. Therefore, the size of the two sub-groups is unbalanced, with only 19 respondents who carry out light PE and 64 respondents who do not.

### 4.5.3 Task complexity

As discussed in the literature review, a task is deemed complex when it requires specific skills and knowledge. In our questionnaire, we asked to what extent linguists perceive translation, revision and PE as complex tasks. We formulated the following hypothesis:

*H03.01: Since revision is usually reserved to more experienced professionals, it will be perceived as more complex than PE.*

Respondents' opinions converge on the complexity of the tasks; the three are considered complex, in the sense that they require specific skills and knowledge. Translation is somehow



| | Strongly disagree | Disagree | Neither disagree nor agree | Agree | Strongly agree | Cannot choose |
|---|---|---|---|---|---|---|
| Translating (without MT) | 0.00% | 1.18% | 3.53% | 24.71% | 70.59% | 0.00% |
| Revising human translations | 0.00% | 2.35% | 3.53% | 38.82% | 54.12% | 1.18% |
| Post-editing MT output | 1.18% | 1.18% | 10.59% | 37.65% | 47.06% | 2.35% |

FIGURE 4.11 – Responses to statement *[This task] is complex (i.e., it requires skills and knowledge).*

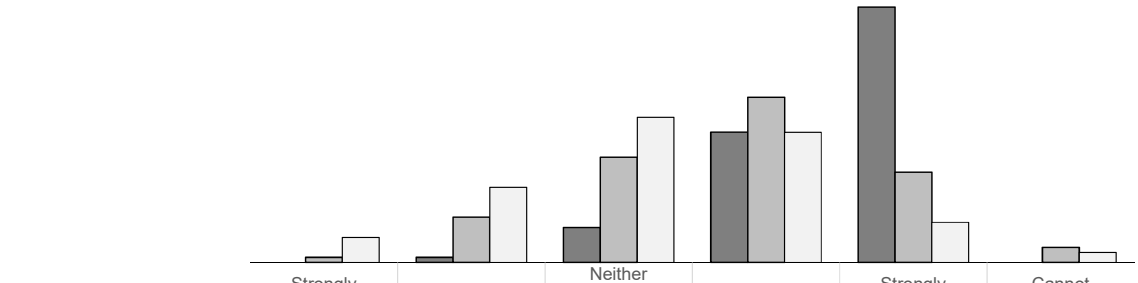
considered more complex than both revision and PE ($M$=4.65, $SD$=.612, n=85). Revision ($M$=4.46, $SD$=.685, n=84) is deemed slightly more demanding than PE ($M$=4.31, $SD$=.810, n=83).

In Figure 4.11, we observe that around 95% (n= 81) of translators and 93% (n=79) of revisers agree that their tasks require skills and knowledge. For PE, only 85% (n=72) of respondents agree on this statement. The percentage of respondents who do not position themselves is higher for PE (11%, n=9) than for both translation and revision (3.5% for both activities, n=3). Very few respondents disagree or strongly disagree with the statement, primarily for revision and PE.

We employed a paired samples t-test to verify the statistical significance of the difference between PE and revision scores on the statements. This difference has been found marginally significant ($t(81) = 1.754, p = .042$).

### 4.5.4 Stimulating nature of the task

In our questionnaire, we asked to what extent translation, revision and PE are considered stimulating tasks.



| | Strongly disagree | Disagree | Neither disagree nor agree | Agree | Strongly agree | Cannot choose |
|---|---|---|---|---|---|---|
| Translating (without MT) | 0.00% | 1.18% | 8.24% | 30.59% | 60.00% | 0.00% |
| Revising human translations | 1.18% | 10.59% | 24.71% | 38.82% | 21.18% | 3.53% |
| Post-editing MT output | 5.88% | 17.65% | 34.12% | 30.59% | 9.41% | 2.35% |

FIGURE 4.12 – Responses to statement *[This task] is stimulating.*

Mean values for the three tasks show that PE is deemed less stimulating than revision. While translation scores 4.49 out of five ($SD$=.701, n=85), revision and PE are assigned lower scores. PE ($M$= 3.20, $SD$=1.045, n=83) is deemed slightly less stimulating than revision ($M$=3.71, $SD$=.975, n=82). Figure 4.12 represents the distribution of scores. Translation is viewed as a stimulating task by a striking majority of participants (91%, n=77), and revision ranks second (60%, n=51), with a higher percentage of linguists who do not position themselves (25%, n=21) and a few respondents who disagree (12%, n=10). Thirty-four respondents (40%) find that PE is a stimulating task, while 29 linguists (34%) neither agree nor disagree, and 15 (18%) simply disagree. Five participants expressed strong disagreement with this statement.

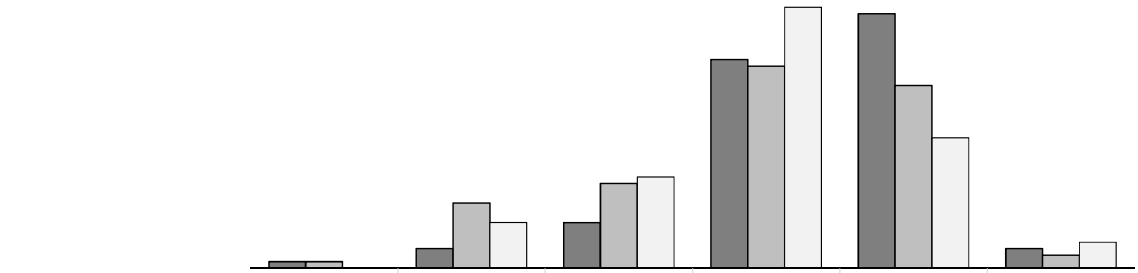
*H04.01: There is a positive correlation between perception of a task as complex and perception of a task as stimulating.*

Spearman's rank-order correlations were computed to assess the relationship between complexity and stimulating nature of the task, for each of the three tasks investigated. For translation, we found a moderate, positive and significant correlation between complexity and stimulating nature of the task, $r$=.550, n=85, $p$<.001. Positive and significant correlations, although weaker, were also found for revision ($r$=.269, n=82, $p$=.014) and PE ($r$=.346, n=83, $p$=.001). The null hypothesis stating that there is no correlation between perceptions of task complexity and perception of that task as stimulating is therefore rejected.

### 4.5.5 Control (over the text's final quality)

We assumed that perceptions of control depend on the role a linguist plays in the translation workflow. Indeed, while translators work on almost the whole process (going from a blank page to a complete draft), revisers are instructed to avoid changing the whole pre-translation, unless it contains major issues, such as inconsistencies or mistranslations (Mossop, 2020). It is not clear, however, to what extent post-editors perceive they are in control of their final texts. On one hand, post-editors are translators who can rely on additional suggestions — that is, MT suggestions — to translate the source text and are therefore responsible for the final quality. On the other hand, for the sake of efficiency and productivity, post-editors are required to reuse acceptable MT suggestions, even if these do not match their own translation style. PE working mode, regarding the use of suggestions, could exacerbate this situation. In particular, post-editors who work on a pre-translated text (just like revisers do) could perceive they are not entitled to change the whole text. Figure 4.13 shows the results for this statement.

The majority of respondents (83.5%, n=71) found that, when translating, they can control the text's final quality ($M$=4.28, $SD$=.865, n=82). Compared to translation, both revision and PE seem to allow for less control. Indeed, the two tasks obtain the same mean value (3.90), but a higher standard deviation in revision ($SD$=1.043, n=83; compared to $SD$=0.875, n=81 in PE) accounts for slightly more variation among the respondents. Sixty-nine percent of revisers (n=59) agree with the statement, while 15% (n=13) do not position



| | Strongly disagree | Disagree | Neither disagree nor agree | Agree | Strongly agree | Cannot choose |
|---|---|---|---|---|---|---|
| Translating (without MT) | 1.18% | 3.53% | 8.24% | 37.65% | 45.88% | 3.53% |
| Revising human translations | 1.18% | 11.76% | 15.29% | 36.47% | 32.94% | 2.35% |
| Post-editing MT output | 0.00% | 8.24% | 16.47% | 47.06% | 23.53% | 4.71% |

FIGURE 4.13 – Responses to statement *When [performing this task], I feel I have control over the text's final quality.*

themselves and a few linguists express some disagreement (13%, n=11). The majority of post-editors agree with the statement (71%, n=60), 16.5% (n=14) select the middle option *Neither disagree nor agree*, and 8% simply disagree with the statement. We formulated the following hypothesis:

*H05.01: PE working mode (use of MT suggestions), affects perceptions of control, as users of pre-translated texts will feel less in control during PE.*

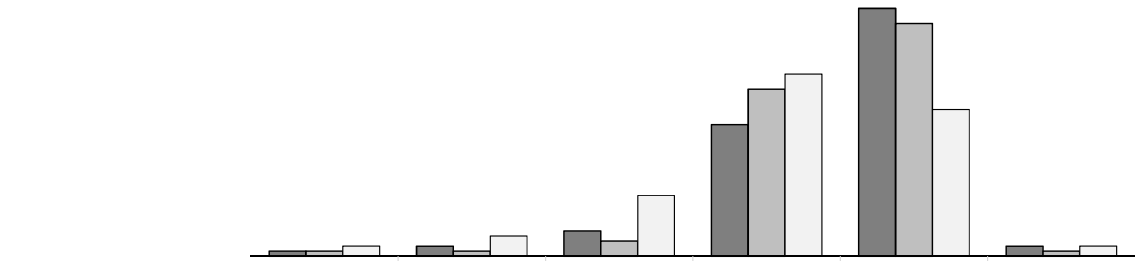
| **MT/CAT tool integration scenario (use of suggestions)** | *Control* | | |
|---|---|---|---|
| | *n* | *M* | *SD* |
| Pre-translation of the entire text (i.e., suggestion in the editing zone) | 46 | 4.09 | .784 |
| Suggestions in a separate window (i.e., the linguist clicks on a suggestion to use it) | 24 | 3.63 | .970 |

TABLE 4.7 – Perceptions of control over the text's final quality during PE, broken down per MT/CAT integration scenario (use of suggestions).

As Wallis (2008) found that pre-translation would affect satisfaction, we also hypothesised that post-editors working with MT in pre-translation mode would report a lower level of control over the text's final quality compared to post-editors who use MT as a suggestion in a different window. Our findings did not provide support to this hypothesis, as mean comparison showed that post-editors who work on pre-translated texts feel also more in control than those who check MT in another window (Table 4.7). Results of a t-test for independent means showed that this difference is statistically significant ($t(39) = 2.015, p = .017$).

### 4.5.6 Ability utilisation

We asked our respondents whether they found that their abilities were put in good use during translation, revision and PE. Furthermore, since we defined complex tasks as those that require skills and knowledge, we also expected to see a positive correlation among complexity of the task and ability utilisation when performing that task.



| | Strongly disagree | Disagree | Neither disagree nor agree | Agree | Strongly agree | Cannot choose |
|---|---|---|---|---|---|---|
| Translating (without MT) | 1.18% | 2.35% | 5.88% | 30.59% | 57.65% | 2.35% |
| Revising human translations | 1.18% | 1.18% | 3.53% | 38.82% | 54.12% | 1.18% |
| Post-editing MT output | 2.35% | 4.71% | 14.12% | 42.35% | 34.12% | 2.35% |

FIGURE 4.14 – Responses to statement *When [performing this task], I make good use of my skills and knowledge.*

Figure 4.14 shows that most in-house linguists who filled out our survey feel they make good use of their skills and knowledge when translating ($M$=4.45, $SD$=.815, n=83), when revising ($M$=4.45, $SD$=.735, n=84), and, to a lesser extent, when post-editing, too ($M$=4.04, $SD$=.956, n=83). The majority of translators express strong (58%, n=49) or simple (31%, n=26) agreement, while a few respondents remain neutral (6%, n=5) or disagree to various extents (4%, n=3). The same applies to revisers, with 54% (n=46) who strongly agree and 39% (n=33) who simply agree with the statement. Three respondents (3.5%) remain neutral and very few revisers express simple (n=1) or strong (n=1) disagreement. The majority of post-editors agree (42%, n=36) or strongly agree (34%, n=29) with the statement, while 14% of respondents (n=12) neither agree nor disagree. A tiny percentage of post-editors disagree or strongly disagree with this statement (7%, n=4 and n=2, respectively).

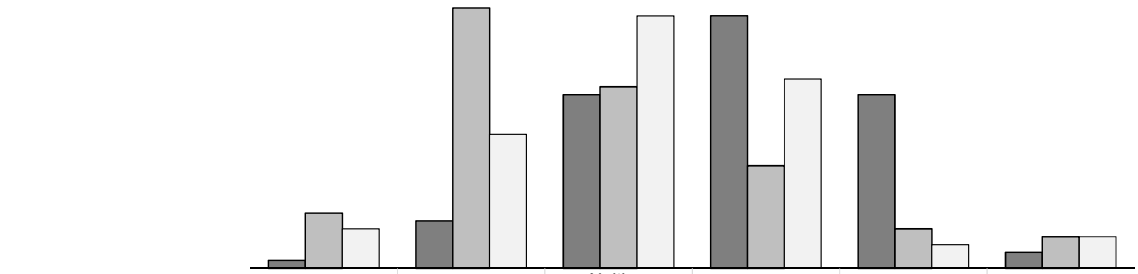
*H06.01: There is a positive correlation between perceptions of a task as complex and respondents' perceptions of making good use of their abilities.*

Spearman's correlations were computed to assess the existence of a positive correlation between perceptions of a task as complex and ability utilisation when performing that task. Results indicated positive, moderate and significant correlations ($p$<.001) for the three tasks investigated, with $r$=.472 for translation (n=83), $r$=.562 for revision (n=84) and $r$=.425 for PE (n=83).

### 4.5.7 Task identity

In our questionnaire, we included a statement about the level of task identity associated with translation, revision and PE, i.e., to what extent these tasks provide linguists with the feeling of creating a new text. We assumed that some tasks induce more task identity than others. As previously discussed, since revisers work on (finished) human-translated texts and should not retranslate the whole content, we hypothesised that PE would allow for a higher level of task identity than revision.

*H07.01: PE allows for a higher level of task identity than revision.*



| | Strongly disagree | Disagree | Neither disagree nor agree | Agree | Strongly agree | Cannot choose |
|---|---|---|---|---|---|---|
| Translating (without MT) | 1.18% | 7.06% | 25.88% | 37.65% | 25.88% | 2.35% |
| Revising human translations | 8.24% | 38.82% | 27.06% | 15.29% | 5.88% | 4.71% |
| Post-editing MT output | 5.88% | 20.00% | 37.65% | 28.24% | 3.53% | 4.71% |

FIGURE 4.15 – Responses to statement *When [performing this task], I feel I am creating a new text.*

Results in Figure 4.15 show that, while a majority of translators agree with the statement ($M$=3.82, $SD$=.952, n=83), most revisers tend to disagree ($M$=2.70, $SD$=1.042, n=81). Indeed, 63.5% of translators (n=54) agree or strongly agree with the statement, compared to only 21% (n=18) of revisers. Forty revisers out of 85 (47%) disagree or strongly disagree with the statement and 23 respondents (27%) remain neutral. Compared to revision, PE allows for a higher level of task identity ($M$=3.04, $SD$=.955, n=81). Although many post-editors cannot position themselves (38%, n=32), the number of respondents who find that working with MT means creating a new text is slightly higher than the number of those who disagree (32%, n=27, and 26%, n=22, respectively). This difference has been tested with a one-tailed t-test for paired samples and found to be statistically significant ($t(78) = 2.257, p = .013$).

### 4.5.8 Work-scheduling autonomy

We were interested in finding out whether linguists had some freedom in deciding the tasks to carry out daily – a factor which contributes to overall job satisfaction (Loher et al., 1985; Finn, 2001; Skaalvik and Skaalvik, 2014). It also represents an informative independent variable for interpreting *overall task satisfaction* (see Section 4.5.10).
Responses to this statement (Figure 4.16) show that only a minority of in-house linguists (16.5%, n=14) can negotiate with project managers the tasks they carry out daily. For most respondents (56.5%, n=48), daily tasks are imposed, and this lack of freedom could

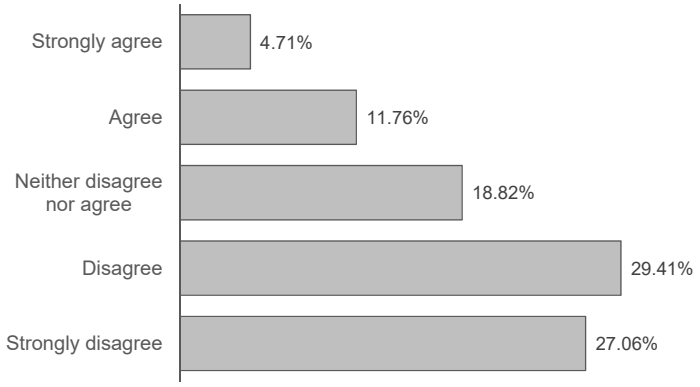



FIGURE 4.16 – Responses to statement *I can choose the tasks I want to perform daily (e.g. ask to perform more translation, rather than revision).*

exacerbate negative attitudes towards certain tasks that are already less appreciated than others. Seven respondents chose not to answer this question. In the next sections, we will use this data to test our hypotheses.

### 4.5.9 Task variety

A statement was included for the respondents to find out whether the introduction of MT was perceived to vary their daily routine. Responses to this statement were expected to correlate positively with perceptions of PE as a complex and stimulating task. The rationale behind this lies in the assumption that viewing PE as complex and stimulating might lead linguists to consider MT as a valuable addition to their workflows. Results are available in Figure 4.17.

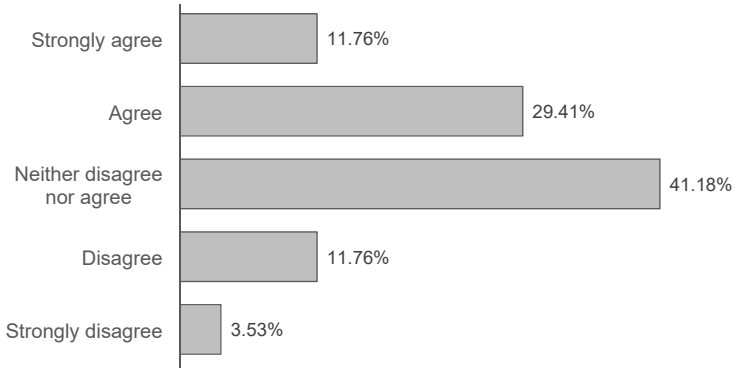



FIGURE 4.17 – Responses to statement *The introduction of MT in my workflow has helped to make my daily work more varied and stimulating.*

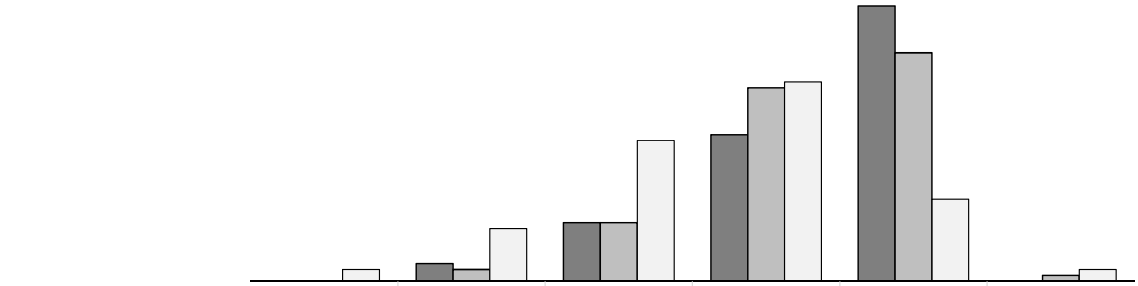
With a mean value of 3.35 out of five ($SD$=.968, n=83), responses show that linguists partially agree with the statement. For 41% of respondents (n=35), the introduction of MT in the workflow has helped to make their daily work more varied and stimulating (sum of options *Agree* and *Strongly agree*). However, there is another 41% (n=35) that neither agree nor disagree with this statement. Only 15% of respondents (n=13) disagree more or less strongly with the statement, while two respondents could not choose an answer.

*H08.01: There is a positive correlation between perceptions of PE as a stimulating activity and perceptions of task variety on the job.*

Spearman's rank-order correlation was computed to assess the relationship between stimulating nature of PE and perceptions of task variety on the job. We found a moderate, positive and significant correlation between the two variables ($r$=.565, $p$<.001, n=81). The null hypothesis is therefore rejected, and we can extend to the whole population the observation made on our sample.

### 4.5.10 Task enjoyment (overall task satisfaction)

Finally, we aimed at eliciting respondents' overall task satisfaction while performing translation, revision and PE, through the statement *I enjoy [performing this task]*. Mean values show that post-editing is less appreciated than the two other tasks. Indeed, while translation obtains a mean of 4.36 (n=85; $SD$=0.829) and revision a mean of 4.30 (n=84; $SD$=0.773), PE scores only 3.59 (n=83; $SD$=0.976) out of five. Results are further detailed in Figure 4.18.



| | Strongly disagree | Disagree | Neither disagree nor agree | Agree | Strongly agree | Cannot choose |
|---|---|---|---|---|---|---|
| Translating (without MT) | 0.00% | 3.53% | 11.76% | 29.41% | 55.29% | 0.00% |
| Revising human translations | 0.00% | 2.35% | 11.76% | 38.82% | 45.88% | 1.18% |
| Post-editing MT output | 2.35% | 10.59% | 28.24% | 40.00% | 16.47% | 2.35% |

FIGURE 4.18 – Responses to statement *I enjoy [performing this task]*.

Translation and revision are considered gratifying activities by approx. 85% of participants (n=72), and the majority of them strongly agree with the statement. Translation is slightly more enjoyable than revision, that is, more respondents strongly agree with the statement (55%, n=47, compared to 46%, n=39 for revision). Only three and two respondents (for translation and revision, respectively) express disagreement, while ten linguists (12%) place themselves in the middle of the scale. Regarding PE, only 56.5% of participants (n=48) seem to enjoy this activity, but not as much as translation or revision, as only 16.5% (n=14) of respondents agree strongly with the statement. In contrast, 28% (n=24) of respondents neither agree nor disagree, while 13% (n=11) disagree more or less strongly with the statement.

*H09.01: PE working mode (use of MT suggestions) has an impact on the level of PE enjoyment.*

Table 4.8 details answers to the statement "I enjoy PE MT output", broken down per PE working mode, i.e., whether the text is entirely pre-translated or whether suggestions are presented in a separate window, inside the CAT tool [26].

| **MT/CAT tool integration scenario (use of suggestions)** | *I enjoy PE* | | |
|---|---|---|---|
| | *n* | *M* | *SD* |
| Pre-translation of the entire text (i.e., suggestion in the editing zone) | 46 | 3.70 | 1.030 |
| Suggestions in a separate window (i.e., the linguist clicks on a suggestion to use it) | 26 | 3.58 | .857 |

TABLE 4.8 – PE enjoyment score, broken down per PE working mode (use of suggestions).

It seems that linguists who work with pre-translated texts enjoy PE slightly more than those who select MT suggestions from a separate window [27]. Results of a t-test for independent samples show that this difference is not statistically significant ($t(70) = .498, p > 0.05$). This result seems to contradict the assumption — retrieved from the work of Wallis (2008) — that pre-translation would affect linguists' satisfaction in a negative way. However, as already discussed in previous sections, it is worth noting that the two sub-groups are unbalanced.

*H09.02: Younger translators (<30) will enjoy PE more than mature translators.*

We hypothesised that linguists' age would affect the level of PE enjoyment, as younger translators would be more open toward using MT as an aid in their workflow. Table 4.9 confirms our assumption and shows that there is a difference in means between sub-groups. However, sub-groups are particularly unbalanced in terms of number of observations, with only eight linguists in the category of younger translators (18-29 years). A one-way ANOVA showed that the effect of age on PE enjoyment is not statistically significant, $F$=(4, 78), $p$>0.05.

*H09.03: Linguists who value revision tasks as much as translation tasks will also enjoy PE, since they enjoy working with pre-translated texts.*

We hypothesised that respondents who enjoy revision at least as much as they enjoy translation are also likely to enjoy PE more than those who prefer translating above all other activities. Our assumption is confirmed by mean comparison of the two sub-groups (Table 4.10). Results of a one-tailed t-test for independent samples indicate that the difference is

26. As per data collected in questionnaire Q1.
27. While the majority of respondents use MT through their CAT tool, nine out of 85 linguists use MT via a web interface, while two other linguists use a combination of methods.

| Age range | | *I enjoy...* Translating (without using MT) | Revising human translations | Post-editing MT output |
|---|---|---|---|---|
| | *n* | 7 | 7 | 7 |
| **18-29** | *M* | 4.57 | 4.29 | 4.43 |
| | *SD* | .787 | .951 | .787 |
| | *n* | 26 | 26 | 25 |
| **30-39** | *M* | 4.69 | 4.31 | 3.60 |
| | *SD* | .618 | .679 | .913 |
| | *n* | 24 | 24 | 23 |
| **40-49** | *M* | 4.21 | 4.50 | 3.43 |
| | *SD* | .884 | .511 | .843 |
| | *n* | 26 | 25 | 26 |
| **50-59** | *M* | 4.19 | 4.08 | 3.42 |
| | *SD* | .801 | .997 | 1.102 |
| | *n* | 2 | 2 | 2 |
| **60+** | *M* | 3.50 | 4.50 | 4.50 |
| | *SD* | 2.121 | .707 | .707 |

TABLE 4.9 – Mean comparison: *I enjoy [this task]*, broken down per respondents' age range.

statistically significant ($t(81) = 3.023, p = .002$).

| *I enjoy [this task]* | *I enjoy PE* *n* | *M* | *SD* |
|---|---|---|---|
| Revision >= Translation | 58 | 3.79 | .894 |
| Revision < Translation | 25 | 3.12 | 1.013 |

TABLE 4.10 – Mean comparison: *I enjoy PE*, broken down per revision/translation preference.

*H09.04: There is a positive correlation between MT system's perceived quality and PE enjoyment.*

In questionnaire Q2, respondents were asked to rate the quality of the MT system used at the CILS (Appendix D, question C22). We assumed that there would be a positive correlation between perceived quality of the MT system and linguists' PE enjoyment. Spearman's rank-order correlation confirmed that the two variables were moderately correlated, $r$=.561, $p$<.001 (Table 4.11).

*H09.05: Linguists who received PE training will enjoy PE more than those who did not receive any training in this task.*

We assumed that understanding how MT works and how one can interact with MT in PE tasks would enhance PE enjoyment. The two sub-groups, namely trained and untrained post-editors, reported different mean values for this statement, with the first sub-group reporting a higher level of PE enjoyment compared to the one who has not received former

| | *I enjoy PE* | | |
|---|---|---|---|
| **MT system's perceived quality** | *n* | *M* | *SD* |
| Very bad | 2 | 1.50 | .707 |
| Bad | 5 | 2.20 | .837 |
| Neither good nor bad | 27 | 3.30 | .869 |
| Good | 43 | 3.88 | .731 |
| Very good | 6 | 4.67 | .516 |

TABLE 4.11 – Mean comparison: *I enjoy PE*, broken down per perceived quality of MT system.

training (Table 4.12). Results of a t-test for independent samples showed that this difference is statistically significant ($t(81) = 2.319, p = .011$).

| | *I enjoy PE* | | |
|---|---|---|---|
| **PE training** | *n* | *M* | *SD* |
| Yes | 44 | 3.82 | .971 |
| No | 39 | 3.33 | .927 |

TABLE 4.12 – Mean comparison: *I enjoy PE*, broken down per PE training.

*H09.06 Linguists who received revision training will enjoy revision more than those who did not receive any training in this task.*

We also assumed that trained revisers would enjoy revising texts more than untrained revisers. Compared to PE training, revision training seems less effective in enhancing revision enjoyment, as the two sub-groups report very similar mean values for the statement *I enjoy revising (human) translations* (Table 4.13). The difference is also non-significant ($t(82) = 1.027, p > .05$).

| | *I enjoy revision* | | |
|---|---|---|---|
| **REV training** | *n* | *M* | *SD* |
| Yes | 35 | 4.40 | .736 |
| No | 49 | 4.22 | .798 |

TABLE 4.13 – Mean comparison: *I enjoy revision*, broken down per revision training.

*H09.07: Linguists who declare having more autonomy in choosing their daily tasks will also enjoy PE more than those who do not have the same autonomy.*

We hypothesised that interacting with MT as a deliberate choice, rather than an imposed task, could contribute to enhancing PE enjoyment. Therefore, we tested whether post-editors who reported a higher level of autonomy in choosing the daily tasks to perform (autonomy>3) would also report higher PE enjoyment levels than post-editors who have

lower autonomy (autonomy<3) [28]. We excluded from this comparison respondents who decided not to position themselves toward the statement (autonomy=3 or option *Cannot choose*). Results of mean comparison are reported in Table 4.14.

| | *I enjoy PE* | | |
|---|---|---|---|
| **Work scheduling autonomy** | *n* | *M* | *SD* |
| >3 | 14 | 3.71 | .914 |
| <3 | 46 | 3.41 | .979 |

TABLE 4.14 – Mean comparison: *I enjoy PE*, broken down per level of autonomy in choosing daily work tasks.

We notice a small difference between the two subgroups, with respondents with high autonomy reporting slightly higher levels of PE enjoyment. We recognise however that the two groups are particularly unbalanced in terms of number of observations. The difference in PE enjoyment between post-editors who have high autonomy and those who have lower autonomy in choosing their daily tasks is non-significant ($t(58) = 1.023, p > .05$).

*H09.08: There is a positive correlation between perception of a task as complex and level of enjoyment for that task.*

The literature on satisfaction in psychology has shown that challenging jobs enhance job satisfaction, particularly for employees with positive core-self evaluations (Judge and Klinger, 2008). In order to test this correlation on our sample, we computed Spearman's rho for each task investigated. We found weak but significant correlations between *task complexity* and *task enjoyment*, for translation ($r$=.287, $p$=.008) and revision ($r$=.249, $p$=.023). PE complexity seems not to correlate with PE enjoyment ($r$=.108, $p$>.05).

*H09.09: Post-editors who use a customised system will enjoy PE tasks more than those who use a generic system.*

Customised MT systems, i.e., those that are trained with in-domain data, mainly CILS's translation memories and glossaries, are reported to provide post-editors with higher-quality output compared to generic MT systems. Therefore, we assumed that users of customised MT would also report higher levels of PE enjoyment. As shown in Table 4.15 below, mean comparison of the three sub-groups [29] shows an improvement in PE enjoyment that corresponds with the specificity of the MT system in use. Indeed, while users of a generic engine assign to this statement a value of 3.40 on average, post-editors who use a combination of generic and custom MT report a slightly higher value (3.55). Finally, post-editors who work

28. Agreement to the statement *I can choose the tasks I want to perform daily (e.g. ask to perform more translation, rather than revision)*.

29. Data on MT systems used at the companies were gathered in questionnaire Q1. While the majority of respondents work with custom MT, generic MT engines are used by 25 respondents. Eleven respondents use a combination of customised and generic systems.

with custom MT output report the highest mean value, 3.70 out of 5. However, the differences among sub-groups were not statistically significant, as per the results of a one-way ANOVA ($F(2, 80) = .792, p > .05$).

| | *I enjoy PE* | | |
|---|---|---|---|
| **MT type** | *n* | *M* | *SD* |
| Customised | 47 | 3.70 | .954 |
| Generic | 25 | 3.40 | 1.000 |
| Other | 11 | 3.55 | 1.036 |

TABLE 4.15 – Mean comparison: *I enjoy PE*, broken down per MT type used at the CILS.

*H09.10: Previous PE experience positively influences the level of PE enjoyment.*

We hypothesised that linguists with previous PE experience would enjoy PE more than those who started working with MT only recently. In questionnaire Q1, which was addressed at CILS' directors or project managers, we asked in which year the linguists had started using MT in production. We found that most of our CILS had implemented an MT solution between 2020 and 2021, with an additional peak in 2018 (see Figure 2.7 in Chapter 2).

In questionnaire Q2, we asked respondents whether they already had some previous, professional PE experience before starting to work at their respective CILS. However, we did not add a follow-up question to find out in which year respondents had started working with MT. Although we cannot calculate the actual PE experience, we assume that having worked with MT in different contexts could benefit linguists and positively affect their "relationship" with MT. Our assumption is confirmed by the results of mean comparison (Table 4.16).

| | *I enjoy PE* | | |
|---|---|---|---|
| **Previous PE experience** | *n* | *M* | *SD* |
| Yes | 17 | 4.06 | .899 |
| No | 66 | 3.47 | .964 |

TABLE 4.16 – Mean comparison: *I enjoy PE*, broken down per previous PE experience.

We observe a difference between linguists who had previous PE experience and those who did not, with the first sub-group reporting a higher mean value of PE enjoyment. While this difference is statistically significant, as per the results of a t-test for independent means ($t(81) = 2.276, p = .013$), we note that two sub-groups are unbalanced, as most linguists had no previous PE experience (understood as having worked with MT in a different workplace).

*H09.11: Previous revision experience positively influences the level of revision enjoyment.*

Similarly, we hypothesised that previous professional revision experience could positively affect revision enjoyment levels. In this case, most participants had already some revision experience in a different working context. The two sub-groups exhibit slightly different mean

| | *I enjoy revision* | | |
|---|---|---|---|
| **Previous REV experience** | *n* | *M* | *SD* |
| Yes | 56 | 4.39 | .779 |
| No | 28 | 4.11 | .737 |

TABLE 4.17 – Mean comparison: *I enjoy revising*, broken down per previous revision experience.

values for revision enjoyment, with experienced revisers reporting slightly higher values than those who revised always in the same working context (Table 4.17). However, the results of a t-test for independent means showed that this difference was not statistically significant ($t(82) = 1.613, p = .055$).

### 4.5.11 Most and least appreciated aspects in PE and revision

The second part of Section E included four non-mandatory, open-ended questions on aspects that linguists appreciate or do not appreciate about working with MT or revising human translations.

Seventy-five linguists commented on the positive aspects of working with MT. Most recurrent factors are speed and productivity gains, the fact that MT provides suggestions and a starting point to initiate the translation process. MT quality was praised by several linguists who answered this question. Help with unengaging and repetitive texts, efficiency, and the chance to have a customised MT tool were mentioned by a few respondents. Two respondents also cited bad MT quality as a positive aspect since it is a sign that human intervention is still needed to produce a good final result. Furthermore, some MT errors are arguably amusing, which can make working with MT somehow entertaining.

Respondents listed a slightly higher number of positive than negative aspects of working with MT, but the latter are more diverse. Seventy-two linguists commented on this question, most often lamenting a loss of human creativity: while, on the one hand, having an MT suggestion can help initiate the translation process, on the other hand, it hinders creativity, and it is less intellectually stimulating. Another often-mentioned, less appreciated aspect is that MT is unable to provide correct and consistent terminology throughout the text. This aspect, together with typical NMT "hidden" errors (especially omissions and fluent mistranslations), leads to a higher cognitive effort than revision. Bad MT quality on some text types, errors that no human would ever make, lack of consistency, lack of cohesion between sentences, and lack of context, contribute to making post-editing a time-consuming task. Some respondents also lamented that the quality of post-edited texts is downgraded compared to human-translated texts, since MT tends to produce very similar texts and an impoverished language.

Seventy-two linguists commented on the most appreciated aspects of revising human translations. This task is often considered "enriching", both for the reviser and the translator, who discover new wordings and translation styles and have thus the chance to broaden their horizons and improve their skills. Another greatly appreciated aspect is the human interaction between the reviser and the translator. Several respondents clarify that they

like revising human-translated texts because translation is the product of a human brain. As such, translations are of good quality and require fewer modifications than in MT. Furthermore, usually, translators carry out research work on the text and check terminology; therefore, human-translated texts are more reliable, and revision is faster than PE.

Only 60 linguists listed some unappreciated aspects of revising human translations. The most recurrent comment concerns bad quality translations, whose revision is time-consuming and sometimes frustrating, or working with unskilled, unreliable translators who do not carry out the required terminology verification. Linguists also voiced their dilemma regarding respecting the principle of avoiding preferential modifications when revising. Human interaction is mentioned among negative aspects, too, when discussions between revisers and translators "do not contribute to a better translation, but rather fuel a battle of egos", as one respondent explained [30].

## 4.6 Discussion

Table 4.18 presents, in a concise format, the outcomes of hypothesis testing.
Data on linguists' attitudes towards revision and PE indicate that the latter is generally viewed as less gratifying and slightly less demanding in terms of skills and knowledge than revision. These findings align with previous studies that have reported negative attitudes towards PE tasks and the use of MT in professional contexts (Läubli and Orrego-Carmona, 2017; Nunes Vieira and Alonso, 2018; Álvarez-Vidal, Oliver, and Badia, 2020).

However, our respondents also reported that PE, more than revision, enables them to create new content and exert greater control over the text's final quality. This contrasts with previous studies in which professionals lamented a perceived lack of control over the target text when using translation technology (LeBlanc, 2013; Marshman, 2014; LeBlanc, 2017; Rossi and Chevrot, 2019). It is worth noting that these studies did not focus on specific translation-related tasks and therefore could not provide nuanced insights in this regard. Additionally, some of these studies did not focus specifically on MT, but rather on TMs or translation technology in general. Previous research on salaried translators' perceptions of control over text quality in PE (Rossi and Chevrot, 2019) suggested that MT limits perceptions of control only in part. Our data offer additional insights on this point. It seems that the common practice of instructing revisers not to make extensive modifications to translated texts restrains their perception of control over the final text's quality than the use of MT tools alone.

From our questionnaire responses, we observed that linguists who had received PE training generally expressed higher levels of satisfaction with the PE task compared to those who had not received such training. In contrast, revision training did not seem to significantly influence linguists' satisfaction with revision tasks. It is worth noting that a slight majority of respondents had received training for PE tasks, while fewer had undergone revision

30. *"Les échanges entre réviseur et traducteur qui ne servent pas une meilleure traduction mais alimentent plutôt une guerre d'égo".*

| Hypothesis | Independent variable | Dependent variable | Null hypothesis (no effect/difference) |
|---|---|---|---|
| H01.01 | PE training | Self-efficacy (PE) | failed to reject |
| H01.02 | REV training | Self-efficacy (REV) | **rejected** |
| H02.01 | Use of MT suggestions in CAT tool | Creativity of the task (PE) | failed to reject |
| H02.02 | Display of MT suggestions in CAT tool | Creativity of the task (PE) | failed to reject |
| H02.03 | Light PE | Creativity of the task (PE) | failed to reject |
| H03.01 | Task | Perception of task complexity | **rejected** |
| H05.01 | Use of MT suggestions in CAT tool | Perception of control | *failed to reject |
| H07.01 | Task | Task identity | **rejected** |
| H09.01** | Use of MT suggestions in CAT tool | Task enjoyment (PE) | failed to reject |
| H09.02 | Age | Task enjoyment (PE) | failed to reject |
| H09.03 | Task enjoyment (REV) (>= Translation enjoyment) | Task enjoyment (PE) | **rejected** |
| H09.05 | PE training | Task enjoyment (PE) | **rejected** |
| H09.06 | REV training | Task enjoyment (REV) | failed to reject |
| H09.07 | Work-scheduling autonomy | Task enjoyment (PE) | failed to reject |
| H09.09 | MT system type | Task enjoyment (PE) | failed to reject |
| H09.10 | PE experience | Task enjoyment (PE) | **rejected** |
| H09.11 | REV experience | Task enjoyment (REV) | failed to reject |

*significant difference found in the opposite direction than expected in alternative hypothesis

**two-tailed

| Hypothesis | Variable 1 | Variable 2 | Correlation |
|---|---|---|---|
| H04.01 | Perception of task complexity | Perception of task as stimulating | **positive, moderate** |
| H06.01 | Perception of task complexity | Ability utilisation | **positive, moderate** |
| H08.01 | Perception of PE as stimulating | Perception of task variety | **positive, moderate** |
| H09.04 | MT perceived quality | Task enjoyment (PE) | **positive, moderate** |
| H09.08 | Perception of task complexity | Task enjoyment | none |

TABLE 4.18 – Hypothesis testing: overview of the outcomes.

training. Compared to PE training, however, our data suggests that revision training has a more pronounced effect in enhancing linguists' task-specific self-efficacy levels.

Trained revisers were more likely to report having the necessary skills and knowledge to perform revision tasks, whereas trained post-editors did not share the same sentiment. This divergence could be attributed to the fact that revision is a more established practice in the field, and consequently, it may have more robust pedagogical strategies in place. Revision has been integrated into university curricula long before PE and may be better integrated in the main curriculum than training in translation technology use. Furthermore, it is important to note that PE training for professional translators often consists of one or two full-day training sessions. While these sessions provide professionals with a basic understanding of the implications of working with an MT engine, it may not be sufficient for linguists to acquire a high level of self-efficacy in PE tasks. Additionally, our survey data (cf. Section 3.4.3 in Chapter 3) indicated that respondents make use of revision guidelines more often than PE guidelines. The majority of CILS have not yet defined comprehensive PE guidelines, potentially leaving linguists with many unsolved doubts about the PE process.

Our analysis revealed that experienced post-editors expressed higher satisfaction with PE tasks than novice post-editors. Conversely, experience in revision did not seem to affect satisfaction with revision tasks. It is important to note, however, that we did not collect specific data on respondents' experience levels with these tasks. Instead, we gathered this information by cross-referencing data from questionnaire Q1 (Chapter 2). As such, we recognise that the relationship between experience with the task and task satisfaction warrants further investigation.

Our data did not indicate an influence of the mode of using or displaying MT suggestions in the CAT tool over the levels of overall satisfaction with the PE task, or on the perception of the task as more or less creative. Contrary to our expectations, post-editors who work on fully-pre-translated texts reported a higher level of control over the text's final quality than those who display MT suggestions in a separate window. We noticed that the former mode of working is also the most common among our respondents, suggesting that they have become accustomed to post-editing fully pre-translated texts and have found other ways to exert their control over the text's final quality. Further investigation is needed to better understand this relationship.

A positive, moderate and significant correlation has been found between the perceived quality of the MT system and the level of enjoyment experienced during PE tasks. This correlation, however, does not imply a direct causal relationship. It is also possible that linguists who enjoy working with MT tend to rate MT systems more favourably. It is important to remember that correlation does not equal causation; an unaccounted-for third variable could have influenced both factors. Additionally, we found a positive correlation between the perception of a task as complex and its perception as stimulating. Yet, complexity does not guarantee enjoyment, underscoring the role of individual disposition in determining what is considered enjoyable.

Overall, our findings revealed an interesting hierarchy among the primary tasks performed by linguists on their job. There is a general tendency to value translation tasks

more highly than revision tasks and, in turn, revision tasks are valued more highly than PE tasks. This hierarchy may be attributed to the relative novelty of the PE task for our respondents, which is accompanied by a lack of clarity regarding the expectations and approaches involved in working with MT. Previous authors (cf. O'Brien, 2012; Mossop, 2020 reported that many translators perceive editing and revising as less enjoyable than translating. However, in our sample, the majority of linguists valued revision tasks as much as translation tasks. Interestingly, these linguists tended to be more satisfied with PE tasks than those who did not enjoy working with pre-translated texts. The underlying reasons for this phenomenon, whether they stem from individual predispositions or other factors, deserve further investigation.

In the open-ended questions, linguists voiced their dilemma about respecting the principle of avoiding preferential modifications when revising. While minimising interventions is a central concept in revision (Riondel, 2021b), we argue that it is crucial to emphasise, both in academic and professional training settings, that this principle does not apply in the same way to PE tasks when MT is used only as a further suggestion to streamline the translation process. This distinction is essential to elevate the status of the PE activity, which is often erroneously perceived as simply working on a fully pre-translated text.

In conclusion, focusing on various characteristics associated with task satisfaction in performing revision and PE, our study offers valuable insights into the intricate relationship between these two activities in the NMT era. Furthermore, since task satisfaction is the facet that better correlates with job satisfaction, our findings contribute to a deeper understanding of what professional, salaried translators appreciate or dislike about the very nature of their jobs.

We acknowledge, however, that our study has a number of limitations. While questionnaires are a common method to gather respondents' opinions and attitudes, they have inherent limitations (Saldanha and O'Brien, 2013, pp. 152–153). Firstly, the use of closed-ended questions forces respondents to selecting among a pre-determined set of options. To mitigate this limitation, we included open-ended questions and comments sections in several closed-ended questions, allowing respondents to provide additional explanations for their answers. Secondly, respondents may alter their actual behaviour and select responses that align with what they believe the researcher expects. Although we do not believe our questions would significantly attract social desirability bias, we could not control for this potential bias. Lastly, regarding our sample size, it is essential to consider that we focused on a very specific cohort of respondents, namely Swiss CILS who already use MT in production. We found that only 26 out of 52 CILS use MT in their workflows, and not all linguists within these CILS have access to MT (cf. Chapter 2). Therefore, our analysis exclusively targeted those CILS linguists engaged in revision and post-editing. When viewed from this perspective, our sample size becomes more reasonable and representative of the targeted population.

## 4.7 Summary of the chapter

In this chapter, we delved into an often-overlooked aspect of the comparison between PE and revision tasks in Translation Studies: the satisfaction of the linguists performing these tasks. To investigate this aspect, we incorporated a dedicated section into our questionnaire on PE and revision practices for CILS linguists, as outlined in the preceding chapter.

In developing our questionnaire items related to satisfaction, we drew inspiration from research in the field of industrial and organisational psychology, as well as from existing literature on translators' satisfaction. These sources enabled us to identify key concepts to explore various facets of task satisfaction for post-editors and revisers. These core concepts include task-specific self-efficacy; creativity; complexity; stimulating nature of the task; control; ability utilisation, task identity, work-scheduling autonomy, task variety, and, ultimately, task enjoyment (pointing to overall satisfaction with the task). In addition to PE and revision, we included translation tasks in our comparison to gain a comprehensive understanding of CILS linguists' satisfaction with their main job tasks.

We formulated 26 statements to which respondents could answer using a five-point Likert scale ranging from *Strongly agree* to *Strongly disagree*. We also included four open-ended questions to elicit comments on aspects that linguist did or did not appreciate about working with MT and revising human translations. For each core concept, we developed hypotheses to guide our data analysis. We examined how responses to various satisfaction items correlated with one another and with data from questionnaire Q1.

Our findings revealed that PE tasks are generally considered less fulfilling and slightly less demanding in terms of required skills and knowledge compared to revision tasks. However, respondents noted that PE, more than revision, allowed them to create new content and exert control over the text's final quality. We also observed that linguists who had received PE training reported higher levels of satisfaction with the PE task than those who had not receive such training. In contrast, while revision training influenced linguists' perception of having the necessary skills and knowledge required to perform revision tasks, it did not appear to influence their satisfaction with these tasks. The same trend was observed regarding experience with the task: experienced post-editors expressed higher satisfaction with PE tasks than novice post-editors. However, experience in revision did not seem to affect satisfaction with revision tasks. Lastly, we found that linguists who valued revision tasks as much as translation tasks tended to be more satisfied with PE tasks than those who did not enjoy working with pre-translated texts.

Having explored linguists' PE and revision practices and their attitudes towards these tasks, the next chapter will narrow its focus to a specific CILS context to investigate linguists' productivity in performing PE and revision tasks.

# Chapter 5

# Comparing PE and revision at the workplace: a small-scale productivity study with statistical and neural MT

## 5.1 Overview

Together with delivering high-quality translations, the ability to provide these translations in a timely manner is a key aspect in determining the success of every language service provider. To manage their internal workload and still uphold the double-step translation plus revision process [1], CILS have resorted to outsourcing part of their assignments. Indeed, as evidenced by our survey in Chapter 2, translation tasks are outsourced by all the respondents, while the QA step is often handled in house, which ensures consistency in style and terminology use for the final text. However, there are instances when requests with very tight deadlines make it challenging to follow the double-step process. In such cases, MT can be employed as a tool to boost linguists' productivity.

In this chapter, we investigate the third key aspect defined under Goal 2, namely linguists' productivity in performing PE and revision tasks. We report on a productivity test carried out in 2018 as part of a project focused on integrating an MT engine into a professional translation workflow. Unlike previous productivity studies that compared the PE task to translation from scratch, our study compares PE to revision. The study was conducted at a CILS that entrusts the majority of its translation tasks to external partners but handles the QA process internally. For the CILS, it was crucial to assess the potential productivity gain or loss associated with handling revision and PE assignments in house.

In our study, productivity is understood as encompassing both linguists' temporal effort and adequacy of the final text. The main objective of the test was therefore to determine whether the traditional workflow involving a translator and a reviser could be replaced by PE performed by a single linguist without compromising on quality. To this end, we measured the time spent in processing human and machine-pre-translated segments, and

1. As requested by the ISO 17100:2015 standard (2015).

performed an additional quality check on the final texts. The productivity test was initially carried out using a customised SMT engine trained with the CILS' data. After five months, the test was repeated, this time employing a customised NMT engine trained on the same data.

The chapter is structured as follows. Section 5.2 provides the background of our study, introducing the CILS partner and the industry-academia collaboration established for this research. Section 5.3 delves into relevant literature on productivity studies, with a particular focus on research conducted in industrial settings. The rationale behind conducting our productivity study is explained in Section 5.4. This section also includes our research questions and related hypotheses. Section 5.5 outlines the experimental design and data analysis methods adopted in our study. The findings are presented in Section 5.6 and further discussed in Section 5.7. Finally, Section 5.8 provides a concise summary of the chapter.

## 5.2 An industry-academia collaborative project

In this section, we provide relevant background information on the CILS partner and the collaborative project established between the CILS and the Department of Translation Technology at the Faculty of Translation and Interpreting (FTI, University of Geneva).

### 5.2.1 The CILS partner

Our CILS partner is the Language Services department at Swiss Post [2], a dynamic company engaged not only in the logistics market but also in various other sectors, including the retail financial market and the passenger transport market [3]. Headquartered in Bern, Swiss Post owns one of the largest CILS in Switzerland, both in terms of the number of linguists employed and the volume of pages processed annually.

The CILS operates under the company's Communication division and is responsible for providing translation services both internally and externally across the various linguistic regions of Switzerland. Its clientele consists of Swiss Post employees [4], who can request translation and interpreting services, as well as other specialized services like video subtitling. Despite being an integral part of the company, the CILS functions as a separate LSP and invoices various divisions within Swiss Post for the assignments it handles. At the time of the project, the CILS processed approximately 80,000 pages per year, employed in house 9 project managers (responsible for tasks preparation and customer service) and 30

2. For clarity throughout the thesis, we will use the terms "the CILS" or "the CILS partner" to refer specifically to the Language Services department at Swiss Post. When referring to the company in its entirety, we will explicitly mention Swiss Post.

3. `https://www.post.ch/en/about-us/profile/swiss-post-group-structure`. Accessed 16 June 2023.

4. In 2022, the company employed approximately 46,500 people. Source: `https://geschaeftsbericht.post.ch/22/ar/en/category/mitarbeitende_en/`. Accessed 16 June 2023.

in-house linguists, each specialised in one of the target languages of the CILS, namely German (de-CH), Italian (it-CH), French (fr-CH) and English (en-UK). In addition to its pool of in-house linguists, the CILS collaborated with two external LSPs (agencies), and maintained a direct collaboration relationship with some freelancers. The translation workload primarily focused on translating from German into French and Italian.

As mentioned earlier, the standard workflow of the CILS primarily consisted of outsourced translation and subsequent in-house QA. In this process, in-house linguists apply the corrections directly to the texts and are thus responsible for the texts' final quality. The QA step is usually performed with the source text, although in certain cases it may be carried out on the target text only or as a spot check. All the steps in the workflow are carried out in a CAT tool. The CAT tool in use at the CILS is XTM [5], a cloud-based translation solution. By leveraging the TM and terminology databases, this tool plays a pivotal role in streamlining and enhancing the translation process.

### 5.2.2 Project overview

In 2017, the management team of Swiss Post's CILS reached out to the department of translation technology at the FTI with a request for assistance in evaluating the viability and potential benefits of integrating MT into the workflow. The main objective of the collaboration was to receive recommendations on the most suitable MT architecture, text type and language pairs that would have yielded the best quality for an MT system trained on Swiss Post's data. The desired MT engine needed to be state-of-the-art and customisable, seamlessly integrating with the CAT tools used internally. Data confidentiality was a critical concern, and the selected engine had to ensure the utmost confidentiality. In the framework of this project, in-house linguists' productivity was to be examined and first experiences with the customers compiled. To achieve these objectives, three FTI students were assigned to the project, with their internships lasting from four months to one year. The collaboration extended beyond the internship period and continued until the end of 2020, resulting in several academic papers and conference presentations detailing the experiences and findings from the project. (Bouillon et al., 2018; Girletti et al., 2019; Mutal et al., 2019; Volkart, Bouillon, and Girletti, 2018).

During the initial phases of the collaboration, various MT integration scenarios were considered, including gisting, MT as an additional suggestion in the CAT tool, and full PE for projects with very short turnaround times. Each scenario was thoroughly evaluated, considering potential opportunities and risks associated with its implementation. The idea of using MT for gisting purposes was to integrate an MT engine into the company's intranet, making it accessible to all staff members to enable quick understanding in multilingual teams (for instance, facilitating the translation of emails and comments on the intranet). It would have also eliminated the need to use external MT interfaces like Google Translate, thereby minimising any risks associated with sensitive information being typed into an unsecured

5. `https://xtm.cloud/`. Accessed 16 June 2023.

platform. However, the quality level of raw MT output could not meet the expectation of customers for all text types, and there was a high risk of providing false information due to MT mistranslations. Instead, the focus shifted to the other two scenarios, which showed promising potential for enhancing linguists' productivity. Introducing MT into the CAT tool as additional suggestion or as the default choice when no suggestions from the TM were available, would have allowed for greater control over the quality of the final output and addressed the concerns related to false information. The same applied to the second scenario, i.e. using full PE for projects with very tight deadlines, where the traditional double process (translation plus revision) might not be feasible due to time constraints.

The pilot project involved a test team consisting of 14 in-house linguists and one project manager: four linguists each for French and Italian, and three linguists each for English and German. Due to the relatively smaller volume of translation tasks into German, the decision was made to prioritise the other three target languages for the pilot project. Nonetheless, human evaluations of the NMT engine were conducted for all target languages [6]. Participation in the pilot project was entirely voluntary, and linguists were provided with clear information about the test steps and the overall project objectives during one of the routine staff meetings.

At the outset of the project, we organised a one-day PE training session for the pilot test team. The primary objective of this training session was twofold: firstly, to equip linguists with the necessary knowledge and skills to work effectively with MT by sensitising them to common errors made by MT systems and explaining them how to evaluate MT output. Secondly, we aimed to foster acceptance of the new technology. Our approach focused on empowering the users, providing them with specific training before involving them in the evaluation process. In this regard, we agree with Silva (2014, p. 26), who stated that "LSPs efforts to push MT without taking into account feedback from translators, reviewers and other concerned language professionals are doomed to fail". We believe that engaging linguists in this way is crucial to avoid biased evaluations stemming from reluctance or skepticism toward MT usage. By actively involving them in the process, we sought to instill a sense of ownership and confidence in their ability to adapt to the change. The training sessions covered various aspects of MT, including its historical development, different MT architectures (RBMT, SMT, NMT), and typical MT errors. We also emphasized the distinctions between MT and TM technology and introduced the concepts of pre-editing and controlled languages. Participants were presented with different definitions of PE, PE levels, and associated guidelines, accompanied by insights from academic research on PE. To put theoretical notions into practice, we organised brief PE exercises, enabling participants to gain hands-on experience with the task.

As for the text material used in MT engine training and subsequent testing with linguists, we considered four different subject areas (henceforth 'domains'): *annual reporting documents*, *vocational training*, *financial services*, and *process manuals for post offices*. These

6. Further details about this evaluation will be provided in the next paragraph.

domains were chosen because they correspond to the largest workload for the CILS. We informed Swiss Post employees (i.e. CILS' customers) responsible for these subject areas about the project's goals and sought their agreement to participate. This step was crucial as it provided us with access to relevant TM and glossary data for the selected domains. For the *de-fr* and *de-it* language pairs, the TMs for these domains were nearly parallel, indicating a substantial overlap of at least 65% of source sentences. This allowed us to make meaningful comparisons between the results obtained for these language pairs. However, the volume of translated material into English was significantly lower. Consequently, for *de-en*, we decided to focus solely on the *annual reporting documents* domain.

The main steps of the project are outlined in Figure 5.1 and briefly presented in the following paragraphs.

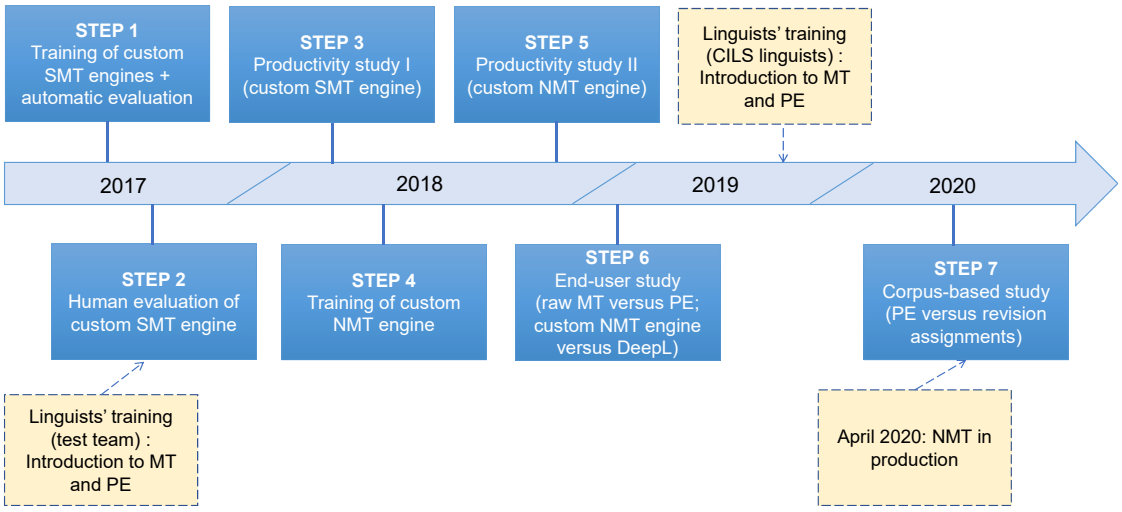



FIGURE 5.1 – Overview of the main steps of the collaborative project between the FTI and the CILS partner.

— **STEP 1 – Training and automatic evaluation of customised SMT engines**. In 2017, NMT was still in its infancy and mainly used in research contexts. Therefore, we started by training SMT engines, as they were state-of-the-art at that time. We considered both a commercial platform, namely Microsoft Translator Hub (MTH) [7] and an open-source platform such as Moses (Koehn et al., 2007). The decision to

7. The platform has been discontinued in 2019, but some information can still be found at `https://www.microsoft.com/en-us/research/project/microsoft-translator-hub/`. Accessed 16 June 2023.

compare these two solutions was practical: the CILS partner did not have the necessary infrastructure and knowledge to maintain an open-source solution, making a commercial solution more feasible for the in-house team. However, due to company privacy requirements, personal data such as names, addresses, and phone numbers had to be anonymised before training an MT engine in a commercial platform.

To evaluate the quality of raw SMT, we used two standard automatic metrics on four different test sets, one for each subject area: BLEU (Papineni et al., 2002) and Word Error Rate. Additional information on the training and testing material can be found in Bouillon et al. (2018). The results of this evaluation showed that Moses outperformed MTH in all subject areas, with *process manuals for post offices* achieving the highest scores for the language pairs *de-fr* and *de-it*. Consequently, we decided to proceed to the next step of the project with the customised Moses engine, hereafter denoted as "the customised SMT engine".

— **STEP 2 – Human evaluation of Moses-based SMT engine**. The second step of the project involved a human evaluation to assess the potential suitability of MT for PE in various language pairs and subject areas, from the perspective of Swiss Post's linguists. Prior to involving them in a real post-editing task, we first allowed linguists to assess the quality of the segments to give them an idea of the expected quality. Eight linguists participated in the evaluation: three for *de-fr* and *de-it*, and two for *de-en*. As mentioned earlier, to prepare the test team for the evaluation task, we organised a one-day training course on MT and PE.

Instead of relying on traditional notions widely used in the MT evaluation domain, such as accuracy and fluency (Moorkens et al., 2018), we opted for a customised evaluation setup. For each source sentence in the test sets (1,000 segments in total), linguists were presented with raw MT output and were asked to answer the following question: "In a post-editing task, would you reuse this translation?". They could choose from three possible answers: "Yes, I would leave it as it is"; "Yes, I would use it with some changes" and "No, I would translate from scratch". As the evaluators were already familiar with the material being evaluated, we did not include any reference translation in our test. However, the translators were aware of the origin (i.e., the subject area) of each segment, enabling them to evaluate if the terminology used was appropriate. Although some might argue that the category "Yes, I would use it with some changes" is too broad, as it encompasses both segments requiring minor changes and those needing intensive PE, our main focus in this evaluation was to determine whether the translators were willing to post-edit the raw MT. While it is true that this type of evaluation can be subjective, we were aware of this risk. However, since the same evaluators were also the linguists who subsequently had to work with the output, the subjectivity aspect was not considered a significant concern.

Overall, results for this evaluation step indicated that the percentage of usable raw MT sentences ranged from 84% to 96% for the language pairs *de-fr* and *de-it*. However, for the *de-en* language pair, only 63% of sentences were deemed usable on average. It is important to note that this lower percentage is not necessarily related to the

quality of the raw SMT output suggestion itself, but also partly attributable to the fact that we calculated a majority judgment on each segment. Since we had only two evaluators for the *de-en* language pair, we considered only those sentences on which the English-speaking evaluators agreed. For further details on the human evaluation and its results, please refer to Bouillon et al. (2018).

— **STEP 3 – Productivity study I (SMT engine)**. The third step of the project involved a productivity study using the SMT engine. Detailed information about this study will be provided in Section 5.4 and subsequent sections.

— **STEP 4 – Training of a customised NMT engine**. In 2018, we began experimenting with NMT. An NMT engine was trained with the same material as the SMT engine, using the open-source toolkit OpenNMT (Klein et al., 2017). Further details on the training of the NMT engine can be found in Mutal et al. (2019) and in Section 5.5.1.

— **STEP 5 – Productivity study II (NMT engine)**. Five months after the first productivity study, a second productivity study was conducted to test the performance of linguists using the customised NMT engine. The details of this second study will also be presented in Section 5.4 and subsequent sections. Following this step, the decision was made to adopt the customised NMT engine for the remainder of the project.

— **STEP 6 – End-user study**. To assess the impact of the newly-introduced NMT customised engine on end-users (i.e. company's employees, which are also customers of the language services), we conducted two comparative evaluations of raw and post-edited versions of machine-translated texts extracted from Swiss Post's manuals. The study aimed to achieve three main objectives: first, to assess whether end-users would rate post-edited MT more favorably than raw MT. Second, to verify whether end-users would find that the CILS' customised NMT system produced better results than a general-purpose, off-the-shelf NMT engine such as DeepL. Third, we also wanted to evaluate whether, when aware of translation production metadata, the end-users/customers would be willing to pay for post-edited texts. This aspect was particularly relevant in determining whether customers value human intervention and would be willing to invest in higher-quality translations or if they prioritise cost-saving even if it means accepting lower quality and potential risks. Additional information and findings from this study can be found in Girletti et al. (2019).

— **STEP 7 – Corpus-based study**. After providing one-day training sessions on MT and PE for all in-house linguists of the CILS, the CILS integrated NMT into its production workflow in April 2020. Four months later, we collected several PE and revision assignments to compare the modifications made to the texts during these two tasks. The findings and analysis of this corpus-based study will be presented in Chapter 6.

The remainder of this chapter will focus on the productivity study (**Step 3** and **Step 5** of the timeline in Figure 5.1). In the next section, we will present and discuss previous productivity studies in the translation domain.

## 5.3 Measuring productivity in professional translation workflows

The measurement of the time spent on a text holds significant relevance for LSPs, as it is an aspect "that can show immediate financial savings" (O'Brien and Duarte, 2015, p. 97). Linguists' productivity is commonly defined in terms of the number of source words processed per hour or per day, making it a widely used metric in both industrial and academic contexts (Silva, 2014). The productivity of post-editors has been a prolific subject of study, while comparable investigations into the revision task have been scarce. These studies will be examined in the following sections. In each section, we will first review productivity studies conducted in academic settings, then we will present those conducted in real-life, professional settings. Finally, we will examine productivity studies that focused on a two-step workflow involving PE and subsequent revision.

### 5.3.1 Studies on productivity in post-editing tasks

The effectiveness of PE as a time-saving strategy has been evidenced in several studies. Since its inception, temporal effort (Krings, 2001) has stood out as one of the most studied PE effort indicators, as it is straightforward to measure and "reflects [both] the technical effort needed to perform the editing [and] the cognitive effort required to detect errors and plan the necessary corrections" (Koponen et al., 2012, p. 12). Researchers have focused on the factors that could impact PE time, such as source sentence structure or length (Tatsumi, 2010; Temnikova, 2010; Koponen et al., 2012; Moorkens and Sasamoto, 2017), the type of edits performed (Popović et al., 2014; Koponen et al., 2012), and the professional experience of the post-editor (de Almeida, 2013; Guerberof Arenas, 2014b; Moorkens and O'Brien, 2015; Yamada, 2011), obtaining mixed results. Some of these studies have been conducted with translation students or non-professional translators (Garcia, 2011; de Sousa, Aziz, and Specia, 2011; Läubli et al., 2013; Aranberri et al., 2014; Koehn and Germann, 2014; Jia, Carl, and Wang, 2019a), as these cohorts are usually easier to reach, but the extent to which their findings are applicable to professional contexts is questionable (Saldanha and O'Brien, 2013; Mellinger et al., 2017).

A study conducted by Green, Heer, and Manning (2013) with 16 professional translators across three language pairs revealed that post-editing raw SMT can yield time savings in comparison to translation from scratch, all while providing similar or even improved quality in the final texts. In another test involving 12 professional translators, Federico, Cattelan, and Trombetti (2012) found that the inclusion of SMT suggestions within a commercial CAT tool improved processing speed for all participants. Similarly, Guerberof Arenas (2014a) conducted a productivity test with 24 participants, focusing on the language pair *en-es*.

The task involved translating from scratch, editing high-quality TM matches (85-94% fuzzy match range) and post-editing some segments coming from a customised SMT engine. The test was conducted using an *ad hoc* PE interface which presented individual segments in random order, and measured the time spent on the segments, while hiding their origin. The researcher also conducted an analysis of the final text quality. The findings indicated an increase in processing speed (words per minute) during PE, but no statistically significant differences were found between PE speed and the speed to process TM matches. Regarding final text quality, the study showed that more errors remained in segments translated from scratch, than in segments coming from TM and MT suggestions[8]. Teixeira (2014a) undertook a similar experiment with ten professional translators, and found that participants spent less time, on average, editing medium and high-quality TM matches than post-editing MT matches. This finding is echoed in Sánchez-Gijón, Moorkens, and Way (2019) who conducted an experiment with NMT output. However, the difference in the time required to post-edit NMT segments and to edit high-quality fuzzy matches was not statistically significant.

Other studies conducted within (or mimicking) authentic translation settings have shown nuanced perspectives on the effectiveness of PE for linguists. For instance, in a study involving the English to Japanese language pair, Moorkens and Sasamoto (2017) reported that PE can be slower than translating from scratch. Similar observations were made by Skadiņš et al. (2014) when considering a text with formatting tags, rather than plain text.

The specific type of MT engine used also plays a role in determining productivity gains. A study by Sánchez-Torron, Koehn, and Phillipp (2016) employing a range of SMT systems has found a linear relationship between raw MT quality and PE speed: the better the raw MT, the higher the PE speed. In a study on the use of PE in literary translation, Toral, Wieling, and Way (2018) found that NMT yields twice the productivity gain (approximately 36%) compared to SMT (+18%). However, Castilho et al. (2017a) reported only marginal improvements on temporal effort when using an NMT engine rather than an SMT one. A more recent study by Pereira (2019) compared the productivity of professional translators when using SMT and NMT from two commercial providers. Although NMT was perceived as requiring fewer corrections and faster to post-edit, results of the productivity test showed that PE time was actually longer for NMT than SMT – though the difference was not statistically significant.

Another noteworthy study conducted by Macken, Prou, and Tezcan (2020) involved 20 in-house translators from the DGT and two language directions (each using a different MT architecture). The study aimed to compare the time spent on PE and translation from scratch. Product and process data from authentic translation assignments were collected over one month, in which linguists were asked to activate MT suggestions for half of their

8. It is worth noting that in a preliminary study with 8 professional translators, TM segments contained more errors compared to both MT segments and segments translated from scratch. The author attributed this outcome to an excess of trust in the TM, which let errors propagate throughout the text (Guerberof Arenas, 2012).

texts. Time measurements were retrieved from SDLXLIFF files using time stamps. Results showed that average processing speed was generally reduced in PE, particularly in NMT, although substantial individual variation was evidenced. Indeed, in some cases, post-editing SMT required more time than translation from scratch. Overall, the average speed gains in PE compared to translation from scratch were 14% for NMT and 12% for SMT.

Research conducted within professional settings has a strong focus on productivity. As aptly noted by Guerberof Arenas (2014a, p. 168), studies conducted in a commercial setting denote "a slight change in how post-editing and post-editing 'experiments' are being viewed and carried out [...], since an increase in productivity can mean a reduction in costs". Indeed, the time savings achieved through PE, as compared to translation from scratch, are often used to compute discounts for the final customer (Scansani and Mhedhbi, 2020; Parra Escartín and Arcedillo, 2015b) – a practice originating from TM tools (García, 2006).

One of the most cited productivity studies within the localisation industry is the one by Plitt and Masselot (2010) conducted at Autodesk, involving 12 participants with no PE experience and three language combinations. The test aimed to compare PE and translation from scratch for IT documentation. The MT engine was a customised SMT engine, and a custom PE tool was used to record processing speed (words per second) and edit distance. Final quality was assessed in a blind test by the in-house QA team. Results showed that PE reduces time by 74% on average and also decreases error rate. All translators were faster with PE, albeit with varying degrees of productivity improvement, ranging from 20% to 131%.

Interestingly, slower translators benefitted more from MT than faster translators. The authors posited that fast translators "have a smaller margin of progression because they have already optimised their way of working" (*ibid.*, p. 11). Echoing the results of Guerberof Arenas (2014a), more errors were found in final texts that had been translated from scratch, compared to those that had been post-edited, across all the language pairs. In a follow-up study with 36 participants and nine language directions (all from English), Zhechev (2014) found that PE enhanced productivity across all language pairs, but to varying degrees. In a blind assessment of final quality, reviewers were unable to distinguish between segments that underwent PE and those translated from scratch.

It is noteworthy that in these studies, the recording of the performance activity only takes into account first-pass editing time – namely from the initial selection of a segment, until the participant moves to the next segment. However, Moran, Lewis, and Saam (2014) showed the importance of capturing performance activity within a CAT tool, accounting for multiple passes on the same segment.

Other industry-based productivity studies are presented within the contexts of larger projects, shedding light on the challenges and strategies for integrating MT in company's production workflows (Schäfer, 2003; Silva, 2014; Flournoy and Duran, 2009; O'Brien and Duarte, 2015; Groves and Schmidtke, 2009; Nunziatini, 2019). For example, Silva (2014) reports productivity figures of a small-sized LSP over a ten-years span. Among the various experiments carried out during this time frame, the author reports on a productivity test involving monolingual PE plus CAT translation. Results showed a minimum of 14% up to

100% productivity increase for all in-house translators. Furthermore, the author reports a steady increase in productivity figures (measured in words processed per project per day) following the integration of MT into the workflow – although other changes in the process, such as better TM management and document pre-processing, are likely to have contributed to this trend. Most importantly, the significant reduction of the time to complete the assignments over the years did not compromise quality, which rather improved.

In a workplace study conducted within a large Danish LSP, Bundgaard (2017a) organised an experiment with 8 in-house linguists in the language pair English to Danish. The aim was to find out how much time in-house translators invest in editing TM and SMT matches on two texts from two different domains. For the purpose of the experiment, source segments were pre-translated with TM (fuzzy matches above 70%) and MT suggestions from a customised SMT engine. Nonetheless, MT matches remained active in a separate window of the CAT tool, enabling linguists to replace TM matches with MT suggestion by simply selecting them. The origin of the segment was clearly marked. Participants worked in their usual CAT tool. To determine the time spent on each segment, the researcher employed a combination of screen recording, keystroke logging and process observation. Results showed that participants invest less time in editing TM matches (starting from 75%, with lower matches demonstrating reduced efficiency) than MT matches. Additionally, the study revealed high individual variation in the time spent on the tasks. However, the statistical significance of this disparity was not evaluated.

While numerous studies revolve around language pairs with English, Läubli et al. (2019) investigated language pairs that have received less attention in PE research, namely *de-fr* and *de-it*. This study is also relevant to our own study because it is conducted in a Swiss CILS. The researchers tested the impact of customised NMT on translation speed and quality in an experiment with high ecological validity, involving four in-house translators (two for each language pair). Participants carried out the tasks in their usual CAT tool, on four source texts in the financial domain, under two conditions: either performing translation with domain-specific TM and termbase, or integrating MT matches when high-quality fuzzy matches (at least 80%) were not available. Final text quality underwent evaluation by two external experts through a blind assessment process. The findings showed that, on average, NMT enables translators to work faster (approximately 60% in *de-fr* and 9% in *de-it*) and does not have a negative impact on overall quality, although it does seem to slightly impact textual cohesion.

### 5.3.2 Studies on productivity in revision tasks

Productivity is not often cited as in issue in revision, where the focus is on achieving maximum quality. However, within professional translation settings, the revision task constitutes part of the workflow and is therefore subject to time constraints. As indicated by Künzli (2007, p. 124), "revisers are very well aware of the fact that time is money" and they even "hesitate to bill the client for the actual time they have spent [on the revision task]". This is an example of what Koponen et al. (2021b, p. 16) define as "the conflict between professional and business concerns" (p.16):

> [Revision] seeks to create adequate quality but it takes time, and therefore — unless the time for the drafting phase can be reduced — it increases costs. This situation gives rise to an ethical question: to what and to whom will the reviser be loyal? More generally, little is known about the usefulness of revision: how many problems in translations — especially serious problems — are being corrected (or not corrected) per hour of revision effort in translation services?

Although the topic has profound practical relevance, academic research on productivity in revision tasks is scant. In an empirical study with ten participants revising different text types, Künzli (2009) investigated the relationship between time and quality of the final text, finding a strong correlation between the variables.

Robert (2012) found a statistical significant effect of the revision procedure on the time spent revising. In an experiment comparing four procedures, she found no statistically significant difference between monolingual and bilingual revision when these are carried out in only one row. Statistical analyses confirmed that a bilingual revision carried out in one row does not take less time than a double-step revision, i.e. a bilingual revision, followed or preceded by a monolingual reading of the target text. She also found that longer revision tasks in terms of duration are those that allow for the higher revision quality, confirming, as Künzli (2007, p. 124) suggested, that "quality takes time". Interestingly, no statistically significant correlations were found between translation or revision experience and the time to complete the revision task.

Other researchers have investigated the time to revise as a further step following a PE task. For instance, as part of her experiment on the productivity of translators in MT-assisted TM translation, Bundgaard (2017a) planned a subsequent revision phase in which each participant had to revise the production of the other participants. However, differently from the first part, the time was not gathered through the analysis of screen recordings but self-reported by revisers. The results showed substantial individual differences in the amount of time spent on the task.

A study by Temizöz (2016, 2017) merits further attention, as it is unique in its kind. Indeed, in an experiment with 20 participants (10 translators and 10 engineers) the researcher compared two workflows in technical translation, involving post-editing by professional translators or subject-matter experts (engineers) followed by a revision step carried out by the opposite category. The author investigated which workflow allowed for faster turnaround times and produced the highest quality in the final product. All participants worked on the same English source text (587 words) pre-translated into Turkish with Google translate, which used a phrase-based MT architecture (PBMT). They had first to post-edit the text, then to revise the same text after two months, in order to avoid a learning effect. Speed was measured as number of words processed per minute and also as total processing time, while quality of final texts was assessed by comparing the texts with a reference translation. T-tests did not show any significant difference in revision speed between the two groups, nor in the total task time for the two workflows, while subject-matter experts' PE results show higher quality than translators' PE for the terminology category. Furthermore, when counting recurring errors, it turned out that revision made by subject-matter experts on translators' PE increased the quality of final texts (almost +20%) more than

revision by professional translators did on engineers' PE (around +4%) and this difference was statistically significant ($p<0.05$). Translators made more terminology errors that have been corrected by engineers during the revision step. In light of these results, the author concluded that the revision step increases text quality overall, and also that the quality of post-editing affects the quality of revision. Finally, the author notes that revisers' subject-matter knowledge is key in ensuring final text quality.

It should be noted that the study focuses on technical translation, where expert knowledge plays a major role. Different results could be obtained if we carried out a similar experiment in less technical or less critical domains. In addition, the study results were statistically significant only when counting repeated errors. Therefore, Temizöz (2017, p. 20) suggested the translation and localisation industry should pay more attention to this point when assessing the quality of final texts. Lastly, we argue that the two groups of participants are somehow unbalanced. Translators involved in the study were not specialised in technical translation nor engineering, while engineers were proficient in English but had no training in translation. While both cohorts reported using Google Translate sometimes, the latter "deal with translation as a component of their daily work" (*ibid.*, p. 5). Consequently, engineers were more familiar with post-editing practice, while translators did not post-edit in their professional practice (Temizöz, 2016, p. 654).

## 5.4 A small-scale productivity study comparing PE and revision

In this section, we explain the reasons for undertaking the study and we introduce the specific research questions addressed.

### 5.4.1 Rationale

In the previous sections, we have seen that several productivity studies have been published throughout the years, most of them showing that MT is beneficial to linguists' productivity, although to varying degrees. However, in many of these studies ecological validity is compromised, since they are conducted in *ad hoc* testbenches and on individual segments presented in random order, rather than on entire texts. Consequently, they fall short of simulating real-world translation contexts.

Most importantly, existing productivity studies compare PE to translation from scratch or to TM-assisted translation. The latter is not equivalent to a revision task, because revisers usually work on finalised translations, while fuzzy matches are translation suggestions. Furthermore, many CAT tools highlight the differences between suggestions in the TM and the segment presented to users. This feature aids in segment processing and reduces cognitive load, offering valuable support during the translation process. Obviously, in authentic revision tasks, no such highlighting occurs.

Remarkably, the study of productivity in revision tasks has been largely neglected. While there is strong empirical evidence that PE is faster than translation from scratch, there are no studies that compare PE to revision in terms of temporal effort. More in general, little is known about the time spent on revising a text.

As highlighted by Läubli et al. (2013) and by Macken, Prou, and Tezcan (2020), PE productivity should be measured within authentic, real-world contexts, as opposed to controlled experimental environments, since the latter scenario carries the potential of overestimating achievable time gains. In our collaborative project with the CILS partner, we conducted a pilot study to verify the feasibility of replacing the traditional workflow translation plus revision with a single-step PE approach. Our approach to assessing productivity differs from previous studies in that it focuses on PE and *revision*, rather than translation from scratch or TM-aided translation. Given that the CILS partner outsourced the majority of its translation tasks and performed revision internally, the outcome of this test held potential for strategic resource allocation. Indeed, adopting full PE in house as the default workflow for time-sensitive projects could have led to quicker turnaround times and reduced supplier costs.

Since time cannot be dissociated from quality when it comes to calculating productivity gains, a quality check was conducted on the final texts produced during the experiment, to verify whether the production method has any (positive or negative) impact on the outcome. It is worth noting that such quality check is uncommon in practice, since revision typically constitutes the final phase of the workflow. Furthermore, while prior research has considered the correction of TM fuzzy matches solely as a component of the translation task, we propose that correcting TM matches should be recognised as an integral aspect of the PE task as well. Finally, an additional dimension that lends significance to our work is the focus on language pairs such as *de-fr* and *de-it*, which are notably underrepresented in MT literature. Indeed, most research outcomes predominantly involve English as either the source or target language.

### 5.4.2 Research questions and hypotheses

All of the above considered, this study seeks to answer the following research question:

> **RQ3**: *When PE and revision tasks are carried out under the same conditions, how does linguists' productivity compare in these tasks?*

Since, in our study, productivity is understood as encompassing both linguists' temporal effort and adequacy of the final text (Section 5.1), this research question can be further divided into two specific sub-questions:

> **RSQ3.1**: *How do PE and revision compare in terms of temporal effort for CILS linguists who perform these tasks?*

The null hypothesis ($H_0$) states that there is no difference between the time spent to process segments translated by humans and those translated by the machine.
The alternative hypothesis ($H_1$) posits that post-editing segments translated by the machine requires more time than revising human-translated segments. This hypothesis is based on previous research suggesting that, for experiments conducted in authentic workplace settings, PE is not faster than translating using TM matches (Skadiņš et al., 2014; Läubli et al., 2019). Additionally, we assume that this hypothesis will apply to both PE of SMT and NMT content.

> **RSQ3.2**: *Does the quality of segments resulting from a PE task compare with the quality of segments resulting from the traditional workflow involving translation plus revision, as assessed by CILS evaluators?*

The null hypothesis ($H_0$) states that segments resulting from a PE task will not contain more issues requiring reworking than segments resulting from a revision task.
The alternative hypothesis ($H_1$) posits that post-edited segments will contain more issues needing reworking than revised segments. This hypothesis is based on the premise that a translation workflow involving two distinct professionals (four-eyes process) would ensure a better control over the quality of the final text than a single-step translation approach. Nonetheless, we hypothesise that, on average, fewer segments resulting from post-editing of NMT output will be flagged as needing further reworking, than segments resulting from post-editing SMT output.

## 5.5 Methods

This study adopts a workplace-based approach (Ehrensberger-Dow and Massey, 2020), delving into the natural work environment of CILS translators, in contrast to a controlled laboratory setting typically employed for experimental purposes. While workplace-based studies offer the advantage of contextual authenticity, they pose challenges for researchers due to the increased complexity in controlling experimental variables.

The main goal of our productivity test was to compare the PE and revision task in terms of processing speed and quality of final texts. The data collection was conducted in two separate sessions (henceforth *test rounds*, TR), with a five-month interval between them. In each session, the participants were provided with two slightly different translations of the same source text and were instructed to "correct" them in their usual CAT tool environment. The MT suggestions for the first test round were generated by a customised SMT engine (Step 1), while for the second test round, the MT suggestions came from a customised NMT engine (Step 4, detailed in Section 5.2.2).

In this section, we will present the methodology used to conduct our experiment. We will start by describing how the customised MT engines were trained, followed by an overview of the materials used and the participants involved. Finally, we will provide details on how the data were collected and analysed.

The study is presented with the caveat that it is a pilot, as it features a limited number of participants and a limited number of segments. However, it has been conducted with high ecological validity. This has been ensured in many aspects: for instance, by letting translators work in their usual CAT tool and work environment and employing their customary language resources; by using screen recording rather than eye-tracking; by presenting translators with complete source texts from domains they are experienced in.

### 5.5.1 Custom MT engines

Two custom MT engines trained with the exact same data were used in the experiment. Table 5.1 reports the amount of data used for training. TM segments were extracted from the main translation memory of the CILS, which in turn consisted of several aggregated TMs. Additionally, approximately 2,200 glossary entries were included in the training data.

| | ***de-fr*** | ***de-it*** | ***de-en*** |
|---|---|---|---|
| TM segments | 2,558,148 | 1,929,530 | 417,817 |

Table 5.1 – Training data, per language pair.

The SMT engine was trained on the Moses platform, following the default training process, consisting of corpus tokenisation, training of language and translation models, tuning and testing on a disjoint set from training. Language models were trained using KenLM (Heafield, 2011) on 4-grams. An encoder-decoder NMT model, specifically the Transformer model (Vaswani et al., 2017), was then trained using OpenNMT-tf (Klein et al., 2017). Default hyper-parameters were used for this model. Infrequent words were segmented into sub-word units using the byte pair encoding (BPE) approach (Sennrich, Haddow, and Birch, 2015). Further technical information on the training process can be found in Bouillon et al. (2018) and in Mutal et al. (2019).

### 5.5.2 Participants

Seven in-house linguists participated in the productivity study: three participants for the language combination *de-it*, and two participants each for the language combinations *de-fr* and *de-en*. Unfortunately, due to issues during data collection in the second test round, one participant for the *de-fr* language combination had to be excluded from the study.

At the outset of the project, the linguists voluntarily joined the testing team, with those who displayed an open-minded attitude towards the MT task being more likely to sign up. All but one participant (who was an intern) had a minimum of four years of experience working at the CILS. All participants had undergone the same PE training, and had taken part in the evaluation of MT quality in the months leading up to the productivity test. Three additional in-house, experienced linguists (one for each language) were involved in the assessment of the final text quality.

### 5.5.3 Text materials

We used two different German source texts (which were not part of the training material of the MT engines): one for the language combinations *de-fr* and *de-it*, and another one for the language combination *de-en*. The text for *de-fr* and *de-it* was extracted from the process manuals for post offices. This subject area had received the best evaluations by the in-house linguists as being the most suitable for PE (Section 5.2.2). Typically, the manuals describe new products or services that Swiss Post is offering or new processes that are relevant for the point of sale. These manuals are used only internally at Swiss Post, and the intended readership is employees in Swiss post branches. Since new manuals are periodically published or existing ones are updated to newer versions, this subject area is one of the most frequent domains in which the CILS linguists work. The text chosen for the experiment described the process and conditions to execute international payment orders. It is worth noting that we manipulated the original text by excluding certain segments. The purpose of this manipulation was to ensure that a comparable number of exact and fuzzy matches was included in all the target languages considered (see Table 5.3). Despite these modifications, the resulting document remained a coherent piece of text, maintaining its overall sense and context.

For the *de-en* language pair, we used a text extracted from the annual reporting documents, since this domain was the only one considered for this language pair (as explained in Section 5.2.2). The annual report provides information on the company's financial performance and the implementation of its strategy [9]. It is supplemented by separate reports, particularly the Financial report (which comprises the Management report and Corporate governance section, as well as the annual financial statements for the Group), and the Sustainability report. These documents are typically published online in March every year and are intended to reach a diverse readership, ranging from the financial community – such as retail shareholders and analysts – to company employees, journalists and the general public. The specific text used in the experiment was extracted from the Financial report 2018 and served as a general introductory text included in the Management report section.

The choice of using different texts for the language pairs involved in the project warrants some consideration. While selecting the same text from the annual reporting documentation for all three language pairs would have enhanced the comparability of results across language pairs, we decided to prioritise a text type that is more frequently encountered in the translation workflow for the two language pairs with the highest translation volume at the CILS. Indeed, the annual report is only available once a year. Furthermore, the internal quality assessment results for French and Italian were not particularly favorable for the annual reporting documentation. Therefore, we considered it more meaningful to test linguists' productivity in the best possible conditions, focusing on text types that are most suitable for PE and leveraging MT on those texts only.

9. `https://www.post.ch/en/about-us/news/annual-report`. Accessed 26 June 2023.

Table 5.2 provides a description of the source texts used for the experiment. While the text used for *de-fr* and *de-it* contains more segments, the one used for *de-en* has a higher word count and greater average sentence length. This discrepancy is attributed to the nature of the texts. The process manual used for *de-fr* and *de-it* mainly consisted of brief instructions for the post office staff. On the other hand, the text from the annual reporting documentation used for *de-en* contained syntactically complex sentences, thus leading to fewer segments but a higher word count.

| | **de-fr, de-it** | **de-en** |
|---|---|---|
| Text type | Process manual | Annual report |
| Segments, total | 44 | 30 |
| Words, total | 406 | 413 |
| Sent. length, *M* | 9.23 | 13.77 |
| Sent. length, *Mdn* | 8.50 | 14 |
| Sent. length, *SD* | 6 | 7 |
| Sent. length, min. | 1 | 1 |
| Sent. length, max. | 25 | 26 |

TABLE 5.2 – Description of source texts used for the experiments.

The texts had been previously translated by external linguists during authentic translation assignments, and then revised by CILS linguists (one reviser per target language) who were not involved in the experiment. Only the unrevised version of these texts was retained and used in the experiment. To obtain a machine-translated version of the source texts, the customised SMT and NMT engines were employed. It is important to clarify that, while these texts were not part of the training data for the MT engines, they were not entirely new to the engines either. The reason is that the CILS' TM contained some exact and high-quality fuzzy matches for these texts, which were also included in the test set used in the experiment to simulate a realistic scenario (as explained later).

To create the test sets for our experiment, we combined human and machine-pre-translated segments. We generated two target versions of the same source text, namely version A and version B, by interleaving segments of different origin. As a result, we obtained two similar pre-translated versions of the same source text [10].

To ensure the most authentic conditions for comparing the revision and PE tasks, we carefully designed the test sets to include TM exact and fuzzy matches. Exact matches were included in both versions, as these segments are typically present in both authentic revision and PE assignments. On the other hand, fuzzy matches have only been added to the MT-pre-translated version. The rationale behind this decision is that in authentic revision assignments, the reviser typically does not encounter fuzzy matches to repair, as the translator has usually already processed them. Conversely, in a PE assignment, the post-editor is responsible for correcting both exact and fuzzy matches. Since previous studies had

10. In other words, in version A, one segment came from the TM, while the following came from the MT engine. In version B, it was the opposite: one segment came from the MT system, while the following came from the TM.

shown that editing low-quality fuzzy matches (below 85% match) is not effective (Sánchez-Gijón, Moorkens, and Way, 2019; Parra Escartín and Arcedillo, 2015a), we only included high-quality fuzzy matches in the test set. Table 5.3 describes the two versions of the test set in terms of segments that were presented to the participants [11].

| | **FR** | | **IT** | | **EN** | |
|---|---|---|---|---|---|---|
| | *A* | *B* | *A* | *B* | *A* | *B* |
| EM (100%) | 7 | 7 | 7 | 7 | 4 | 4 |
| FM (90-99%) | 2 | 4 | 2 | 2 | 3 | 0 |
| HT | 18 | 19 | 19 | 18 | 14 | 12 |
| MT | 17 | 14 | 16 | 17 | 9 | 14 |
| **Total** | 44 | 44 | 44 | 44 | 30 | 30 |

TABLE 5.3 – Number of segments in each test version (A, B), per segment origin. HT = human pre-translated; MT = machine pre-translated; EM = TM exact match; FM = TM fuzzy match.

At first sight, the number of MT suggestions appears slightly lower compared to the segments coming from human translation. However, it is essential to consider that fuzzy matches replaced MT whenever available. To gain a clearer understanding, the number of fuzzy matches should be added to the count of MT segments. When this is done, we can observe that the resulting number of segments from HT and MT are reversed between versions. Additionally, exact matches were repeated in each test version.

Table 5.4 reports the total number of source and target words in each test version, categorised by target language and segment origin. The totals provided at the bottom of the table represent the overall number of target words in each test version, regardless of the origin of the segments.

### 5.5.4 Experimental setup

A within-subjects experimental design was implemented, in which the same subjects were repeatedly tested under different conditions. Indeed, participants were required to correct two pre-translated versions of the same source text. This task closely resembled revision tasks that they routinely perform at the CILS (as explained in Section 5.2.1). However, due to the text material design, participants worked both on a fully human pre-translated and on a machine pre-translated version of the source text. While this setup introduces a sequence effect, where participants are likely to be faster in the second version, this condition was consistent for all participants and sentences, whether they were machine-translated or human-translated, which helps mitigate any potential bias.

Participants carried out the experiment in their usual (physical and digital) work environment. The experiments took place at the CILS' premises during office hours, therefore as

11. Although we treated TM matches as a separate category, it is important to recognise that they are, in essence, human translations. However, while participants in our study were not aware of the actual origin of human and machine-pre-translated sentences, TM matches appeared in the CAT tool with metadata indicating their provenance.

| | FR | | IT | | EN | |
|---|---|---|---|---|---|---|
| | *A* | *B* | *A* | *B* | *A* | *B* |
| **EM (100%)** | | | | | | |
| source | 19 | 19 | 19 | 19 | 6 | 6 |
| target | 22 | 22 | 21 | 21 | 8 | 8 |
| **FM (90-99%)** | | | | | | |
| source | 32 | 49 | 25 | 32 | 53 | 0 |
| target | 31 | 47 | 26 | 28 | 49 | 0 |
| **HT** | | | | | | |
| source | 193 | 194 | 194 | 193 | 206 | 201 |
| target | 212 | 240 | 223 | 218 | 238 | 220 |
| **MT** | | | | | | |
| source | 162 | 144 | 168 | 162 | 148 | 206 |
| target, SMT | 194 | 164 | 172 | 174 | 157 | 210 |
| target, NMT | 192 | 151 | 195 | 172 | 167 | 223 |
| **Total** | | | | | | |
| source | 406 | 406 | 406 | 406 | 413 | 413 |
| target (test round SMT) | 459 | 473 | 442 | 441 | 452 | 438 |
| target (test round NMT) | 457 | 460 | 465 | 439 | 462 | 451 |

TABLE 5.4 – Number of words in each test version (A, B), per segment origin. HT = human pre-translated; MT = machine pre-translated; EM = TM exact match; FM = TM fuzzy match.

part of linguist's daily work. Each participant was provided with a laptop equipped with a screen recording software, BB Flashback Express v.5.1, which also featured a key-logging function [12]. The laptops provided were connected to the participants' usual screens, keyboards, and mouse appliances [13]. Participants worked within their regular CAT tool, XTM Cloud v.11.1 and 11.3 [14], had access to an internet browser and to their usual linguistic resources, such as TMs, terminology database and online dictionaries. However, to ensure the integrity of the experiment and prevent access to official translations, the specific translation memories containing the texts used for the experiment were excluded. Before the main experiment, the setup underwent a preliminary test with a CILS linguist who did not take part in the final test.

Participants were made aware of the goals of the experiment (detailed instructions can be found in Appendix E), although they were not provided with specific information regarding the test set design. Rather, they were informed that they would be working on "two similar texts" resulting from a blend of human and machine translations. They were also aware of the subject area of the text. The only available metadata in the CAT tool pertained to the

12. The decision to use the built-in key-logging function in the BB Flashback Express software was made to avoid any interference with the experimental setting. Furthermore, using a separate keylogger like Inputlog (Leijten and Van Waes, 2013) could have introduced unwanted complexities in the data analysis, as results from keyloggers are often challenging to interpret (do Carmo, 2021b).

13. The company makes use of shared desks, so participants worked in an open space and had the flexibility to choose a different desk to sit at each day. All tables were equipped with one or two screens to which linguists could connect their own laptops, keyboards, and mouse appliances. To minimise distractions from other colleagues, a smaller area within the open space was booked exclusively for carrying out the experiment.

14. New versions of the tool were released during the period of the productivity tests.

translation memory, specifying the percentage of fuzziness. The instructions provided to participants were to make use of the pre-translated material as much as possible, while also generating a final text that met the standard quality requirements for this particular text type. Additionally, they were informed that their screen activities would be recorded and were requested to refrain from searching for the official translation published online [15]. Participants provided their consent by signing an informed consent form (available in Appendix F).

Once the participants had completed the task, an additional CILS linguist from each respective target language team (hereafter referred to as the "evaluator") assessed the quality of the final texts. Similar to the participants involved in the productivity test, the evaluators were kept unaware of the origin of the pre-translated segments. They were asked to indicate whether each sentence required any further correction. They were provided with a spreadsheet file containing the source text (one segment per row) along with multiple columns, each corresponding to a text processed by a different participant. Next to each translation, a blank column was left for the evaluators to document any issues identified in the respective segment. Detailed instructions for the evaluators (available in Appendix G), were sent via email. These instructions explicitly emphasized the importance of assessing each translation in isolation, without drawing comparisons among different translations. Additionally, evaluators were explicitly instructed to identify instances where the translation was accurate, yet presented fluency issues. Each evaluator worked independently on the spreadsheet file during office hours and submitted the file via email. It is important to acknowledge that, although the texts within the spreadsheet file were not associated with the participants who processed them (identities were anonymised), absolute anonymity could not be guaranteed, as the evaluators were aware of who was taking part in the experiment. However, considering that the participants had been working at the CILS for many years and occasionally reviewed each other's work, any potential negative repercussions of this setting were not anticipated. Similar considerations extended to the less experienced participant, who, as an intern, had become accustomed to being revised by more experienced colleagues. Therefore, this aspect was not deemed to raise ethical concerns.

The experiment was conducted over two different iterations: the first in April 2018 – using the SMT engine – and the second in September of the same year – using the NMT engine. Both test rounds involved the same participants and evaluators. Participants were presented with same source texts and slightly different target version, resulting from the shuffling of HT and (statistical or neural) MT. Although the possibility of a learning effect between the two test rounds exists, we posit that this likelihood was mitigated by the five-month interval between the rounds. Furthermore, the results of the quality assessment was not shared with the participants during this period.

15. This instruction applied particularly to the English group, given that the annual reporting documentation was accessible online. In contrast, the text used for the French and Italian groups was part of internal documentation and not available online.

Participants who shared the same target language were exposed to the same source texts and target versions for their respective tasks. They worked sequentially on the two test versions, with a brief break between tasks. All the participants processed the test versions in the same order: first test version A, then test version B. Regrettably, the limited number of participants precluded us from implementing a counterbalancing strategy for the task order. Given that the participants were working in their customary CAT tool interface, and on their usual task and text type, an initial training phase to familiarise themselves with the test environment was deemed unnecessary.

### 5.5.5 Data collection

As a result of our experimental design, we collected four different versions of the source texts processed by each participant: on the one hand, one (fully) revised and one (fully) post-edited text originating from the SMT; on the other hand, one (fully) revised and one (fully) post-edited text originating from the NMT engine.

Screen recordings were analysed (approximately 11 hours of recordings in total) to establish the time spent on each segment. This has been done using a spreadsheet file template, in which we marked the start and end times for editing each segment, drawing on BB Flashback's in-built keylogging feature that allows to see the specific point in time when the user selects the segment using the mouse or the keyboard. Having the start and end time available, it was easy to automatically calculate the time spent on the segment, in seconds. This included the time spent reading the segment, as well as subsequent visits of the same segment during the checking phase (Moran, Lewis, and Saam, 2014). Following a strategy proposed by Bundgaard (2017a), the time spent for concordance searches, dictionary consultations and other internet queries was added to the total time spent on the active segment. This has been done because the active segment is presumably responsible for such queries, except if the term searched pertains to another segment. In this case, the time for the query is attributed to the segment in which the term is located. In the spreadsheet file template, we also noted whether such queries had occurred, at what time and which terms were researched. Differently from previous work that took into account processing thresholds (see for instance Federico, Cattelan, and Trombetti, 2012), we did not discard segments based on the time spent on them because they simply signal time spent on research, rather than translator's behaviour (such as distractions or pauses) or issues with the tools (as we verified during process observation and while viewing the recordings that there were no issues).

Data on processing time, as well as the results of quality checks carried out by CILS evaluators on the final texts (Section 5.5.4) were organised as a table in a new spreadsheet file, and then transferred in the SPSS software for statistical analysis. The table contained the following information:

— *Segment ID*: the ID of the segment in the source text (from 1 to 44);

— *Target language*: the target language of the task (FR, IT or EN);

— *Participant*: the ID of the participant, from P01 to P07 (P05 had been discarded, as explained in Section 5.5.2);

— *Segment origin*: the origin of the pre-translated segment (human translation, HT; machine translation, MT; TM exact match, EM; TM fuzzy match, FM);

— *Test version*: either version A or version B (Section 5.5.3);

— *Test round*: whether the segment has been processed in the first (SMT) or second (NMT) test round;

— *Source sentence length*: the length of the source segment, in words [16];

— *Target starting length*: the length of the pre-translated segment, in words;

— *Target final length*: the length of the final segment, in words;

— *Processing time*: the time spent on the segment, in seconds;

— *Modifications*: whether or not the pre-translated segments had been modified;

— *Reworking*: whether or not the segment needed reworking, as resulting from the quality check step;

— *Issue(s) type*: the type(s) of quality issue(s) flagged in the segment during the quality check step;

— *Time-per-word*: seconds spent on average on one source word (time normalized to the length of the source segment).

Overall (across all languages, test versions and test rounds), we collected 944 segments.

### 5.5.6 Data analysis

Our analysis is primarily quantitative. Our dependent variables are processing time and final quality, while our independent variable is segment origin. For the analysis of collected data, we started by measuring total time per task and per segment origin (i.e. processing time).

We also report processing speed, obtained dividing the time spent on the segment by the number of words in the source segment. We used the source word count as this remains consistent between SMT and NMT, while choosing the starting target length would have lowered the comparability of the results. When calculating average processing speed, we decided to exclude exact matches, and this for two reasons: first, they were the main responsible for outliers in the dataset, since sometimes they record zero seconds as some participants did not click on them. Second, they are typically very short segments (two to three words on average, as many of them are paragraph titles) and since we are considering seconds per word, the results for these segments are skewed. We report the average processing speed of individual translators in each translation condition and per segment origin for the whole dataset.

16. The length of the string in cell A1 was calculated using the following Excel formula (in French): =NBCAR(SUPPRESPACE(A1))-NBCAR(SUBSTITUE(SUPPRESPACE(A1);" ";""))+1

Apart from descriptive statistics (mean, median and standard deviation), we tested for the statistical significance of differences in our groups of data. As the data regarding time were significantly non-normal [17], we employed a Wilcoxon signed-ranks test to assess whether the data from each of the two dependent groups were statistically different ($p<0.05$). Regarding quality checks performed on the final texts, we created crosstabs with assessment and segment origin, then we verified the statistical significance of the distribution through a Chi-square test.

## 5.6 Results

In this section, we present the results of our experiment. Firstly, we analyse data on temporal effort (processing time and processing speed) in the two test rounds. Secondly, we focus on the quality checks of texts processed by our participants.

### 5.6.1 Processing time

Figure 5.2 shows the total time spent by participants on the task in each test round, as well as the average time per test round across all participants [18].

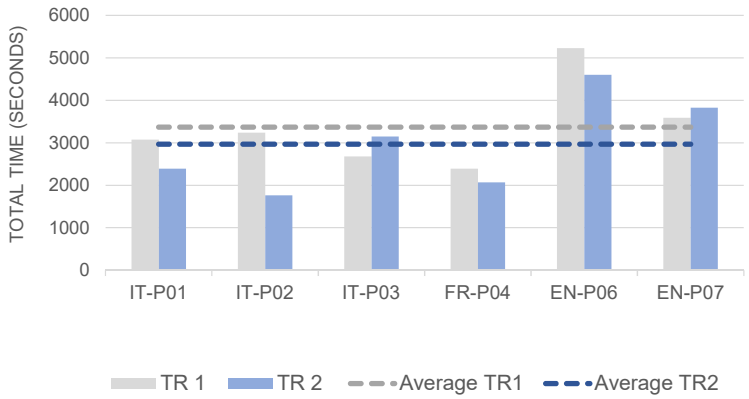



FIGURE 5.2 – Participants' total processing time, per test round (SMT versus NMT).

On average, participants spent 56 minutes on the first test round (Median: 53 minutes) and 49 minutes on the second one (Median: 46 minutes). Overall, we observe that the average time is reduced by 13% in the second test round. We argue nonetheless that it is unlikely that participants' processing time could be affected from the first test, as the second one has taken place after five months in which the participants have carried out several other assignments in the same format, as part of their daily work. Only one linguist (P02) has been extremely faster in the second round, while the other participants have been

17. We tested the data for normality using Kolmogorov-Smirnov test: processing speed for segments in the revision task D(400)=.243, $p$=<.001; processing speed for segments in the PE task D(400)=.166, $p$=<.001.

18. In this section, we report raw time, including the time spent on TM exact matches, which have then been excluded from further analysis.

only slightly faster. Two linguists (P03 and P07) have even taken slightly more time in the second test round than in the first one.

We notice that the difference between the two text types used in the experiment is also reflected in the total time to process them. Indeed, participants who worked on the process manual (P01 to P04) spent on average less time on this text, compared to participants who corrected the text from the annual reporting document (P06 and P07). Individual processing time varies from 29 minutes to one hour and 27 minutes. Remarkably, the least experienced translator (P06) spent significantly more time than the others to complete the task.

Table 5.5 details the total time each participant spent on each test version and test round.

| | **Test round (TR)** | | | | | |
|---|---|---|---|---|---|---|
| | **TR1** | | | **TR2** | | |
| Participant | *A* | *B* | *Total* | *A* | *B* | *Total* |
| IT-P01 | 00:25:02 | 00:26:13 | 00:51:15 | 00:22:56 | 00:16:59 | 00:39:55 |
| IT-P02 | 00:27:39 | 00:26:19 | 00:53:58 | 00:13:16 | 00:16:07 | 00:29:23 |
| IT-P03 | 00:28:31 | 00:16:10 | 00:44:41 | 00:31:49 | 00:20:42 | 00:52:31 |
| FR-P04 | 00:20:19 | 00:19:35 | 00:39:54 | 00:20:08 | 00:14:23 | 00:34:31 |
| EN-P06 | 00:51:58 | 00:35:12 | 01:27:10 | 00:39:35 | 00:37:10 | 01:16:45 |
| EN-P07 | 00:33:54 | 00:25:57 | 00:59:51 | 00:34:58 | 00:28:53 | 01:03:51 |

TABLE 5.5 – Total time to complete the task (hh:mm:ss), per participant, test version (A, B) and test round (TR1=SMT, TR2=NMT).

As expected, test version B required systematically less time to be completed, compared to test version A, except for participant P02 in the second test round. On average, version B required 25% less time than version A in the first round (SMT) and 21% less time than version A in the second round (NMT). However, the final texts produced in the two versions are never exactly the same, meaning that the participants tried to leverage the pre-translated material as much as possible.

### 5.6.2 Processing speed

To calculate processing speed, we normalized the time spent on each segment – expressed in seconds – dividing it by the number of source words in that segment. Figure 5.3 shows the average processing speed of segments during revision and PE across all participants and language pairs, in the two test rounds. We recall that the PE task includes the processing of TM fuzzy matches, while the time spent on TM exact matches has been excluded from the analysis for the reasons explained in Section 5.5.6.

Overall, we observe that linguists are faster during revision than during PE. In the first test round, participants take on average one second less per source word when processing human pre-translated segments compared to MT pre-translated and TM fuzzy matches. In the second test round, we observe a slight reduction in the average processing speed for both tasks (revision and PE). Nonetheless, we notice that the delta between average processing

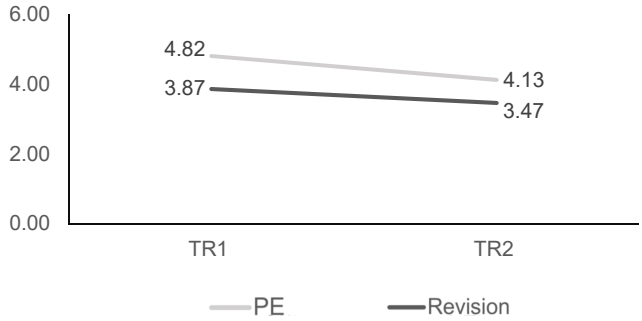



FIGURE 5.3 – Average processing speed (seconds/word) during revision and PE, per test round (TR1=SMT; TR2=NMT), for all participants combined.

speed during revision and PE is reduced in the second round, which suggests that the impact of NMT in speeding up the processing of MT segments is greater compared to SMT. A Wilcoxon signed-rank test indicated that the processing speed of NMT segments is significantly higher compared to the processing speed of SMT segments ($z$=-2.38, $p$=.017) [19]. Conversely, the difference between the processing speed of human pre-translated segments in the two test rounds is not statistically significant ($z$=-.904, $p$=.366). This would seem to suggest that conducting the experiment again after several months has not influenced its outcome.

Tables 5.6 and 5.7 detail the average processing speed per task (revision and PE) across target languages in the first and second test round, respectively.

| Group/task | $n$ | $M$ | $Mdn$ | $SD$ | $z$ | $p$ |
|---|---|---|---|---|---|---|
| IT | | | | | −4.32 | <.001 |
| PE | 111 | 4.43 | 3.06 | 4.29 | | |
| Revision | 111 | 3.59 | 2.00 | 6.32 | | |
| FR | | | | | −2.39 | .017 |
| PE | 37 | 3.89 | 2.71 | 3.15 | | |
| Revision | 37 | 2.78 | 1.33 | 2.93 | | |
| EN | | | | | −1.06 | .290 |
| PE | 52 | 6.30 | 4.57 | 5.75 | | |
| Revision | 52 | 5.25 | 3.45 | 4.52 | | |
| Overall | | | | | −4.54 | <.001 |
| PE | 200 | 4.82 | 3.28 | 4.61 | | |
| Revision | 200 | 3.87 | 2.19 | 5.44 | | |

TABLE 5.6 – Processing speed (seconds/word), per task and target language group, test round 1 (SMT).

The high standard deviations (compared to mean values) denote high variability among different segments. The differences between the tasks in each target language group are statistically significant, except for the English-speaking group.

19. For this test, we only considered machine-pre-translated segments. However, it is worth noting that we run another test including TM fuzzy matches and the results are still significant ($z$=-2.16, $p$=.031).

| Group/task | $n$ | $M$ | $Mdn$ | $SD$ | $z$ | $p$ |
|---|---|---|---|---|---|---|
| IT | | | | | −2.97 | .003 |
| PE | 111 | 3.56 | 2.57 | 2.87 | | |
| Revision | 111 | 3.20 | 1.83 | 4.73 | | |
| FR | | | | | −2.77 | .006 |
| PE | 37 | 3.53 | 2.45 | 3.55 | | |
| Revision | 37 | 2.02 | 1.42 | 1.54 | | |
| EN | | | | | −0.63 | .527 |
| PE | 52 | 5.77 | 4.18 | 4.45 | | |
| Revision | 52 | 5.08 | 4.45 | 3.47 | | |
| Overall | | | | | −3.51 | <.001 |
| PE | 200 | 4.13 | 2.85 | 3.58 | | |
| Revision | 200 | 3.47 | 2.12 | 4.12 | | |

TABLE 5.7 – Processing time (seconds/word), per task and target language group, test round 2 (NMT).

### 5.6.3 Final quality

We move on to discuss the results for the quality checks performed by CILS' evaluators on the texts processed during the experiment. Figure 5.4 reports the proportion of segments that need further correction and their origin on the overall test set across the two test rounds. Exact matches have been excluded from the count and will be discussed later.

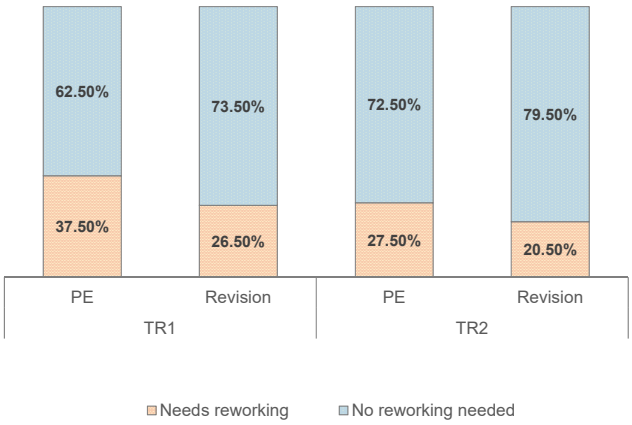



FIGURE 5.4 – Percentage of final segments needing further correction, per task (PE or revision) and test round ($N$=400 in each test round).

We observe that a higher percentage of MT and TM fuzzy matches processed by the linguists requires further correction, compared to human pre-translated segments. This applies to both test rounds, although we notice an improvement in the second test round. Indeed, 37.5% of TM fuzzy matches and SMT segments still needs some reworking, compared to only 27.5% of TM fuzzy matches and NMT segments. These frequencies were significantly different in the first test round, $\chi^2$ (1, n=400) = 5.56, $p$=.018. However, this did not apply to the second test round, where it seems that there is no statistically significant differences between sentences requiring further correction, depending on their origin, $\chi^2$ (1, n=400) = 2.69, $p$=.101.

The extent to which this is due to the impact of the NMT system, rather than to the fact that we have repeated the experiment with the same source text, is unclear. Indeed, we observe that in the second test round the percentage of sentences needing further correction is also reduced for human pre-translated segments, which are exactly the same between test rounds.

Figure 5.5 details the individual results of participants in each test round.

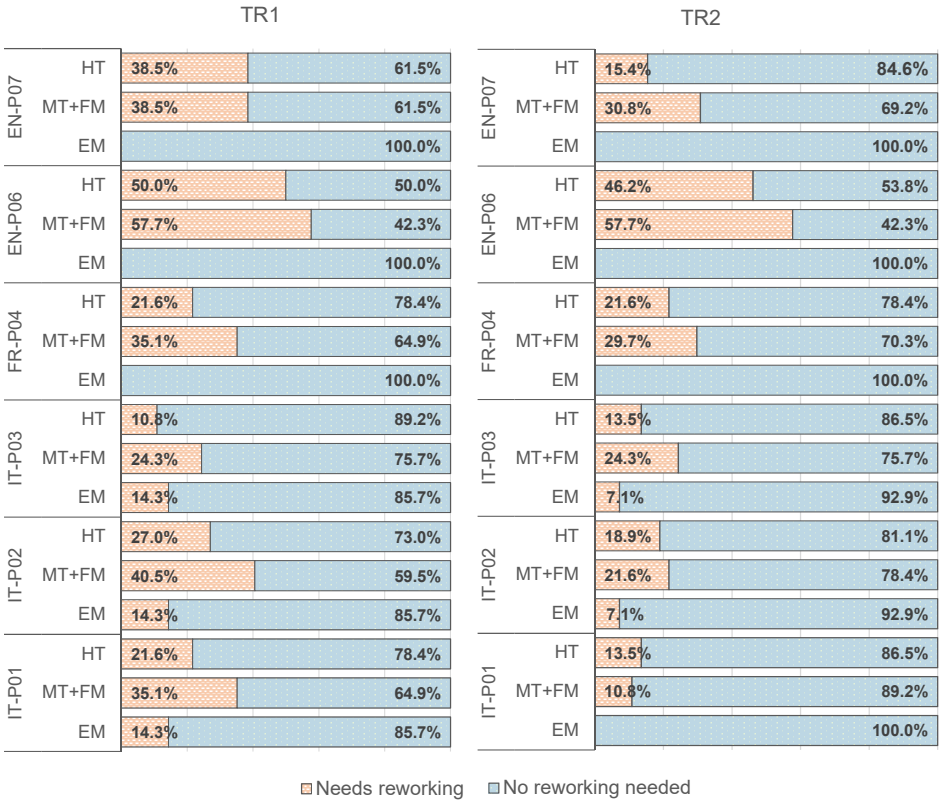



FIGURE 5.5 – Percentage of final segments needing further correction, per participant and test round.

In the first test round, we observe that the percentage of sentences needing reworking is consistently higher in PE tasks, except for participant P07, for whom this percentage is the same among human pre-translated sentences (revision task). In the second test round, the percentage of MT and TM fuzzy pre-translated sentences is reduced for the majority of participants, but remains nonetheless slightly higher compared to human pre-translated segments – except for participant P01. Participant EN-P06 produced the text needing the most reworking, regardless of segment origin. This was expected, since this participant was the least experienced of our pool of linguists and was not yet accustomed to the text types and the CILS requirements in terms of quality. Surprisingly, this applied to both test rounds, with a marginal improvement on human-pre-translated sentences in the second test round.

Most participants left exact matches untouched; however, this approach was erroneous for Italian-speaking participants. The issue centered around two paragraph subtitles that, in the original source translation, were in the plural and singular forms, respectively. These

headings preceded a list of requirements for validating international payments. A discrepancy arose because the first heading referred to a single condition, while the subsequent one encompassed multiple conditions. This issue originated from our own oversight when selecting the segments to include in the final test set. Italian-speaking linguists either overlooked this issue or deliberately chose to adhere to the source material. In the second test round, only one participant corrected both paragraph headings, while the other IT participants addressed only one of the two headings.

## 5.7 Discussion

In Section 5.4, we hypothesised that linguists would require more time to complete PE tasks than revision tasks. Our findings seem to substantiate this hypothesis, since processing speed is indeed reduced in PE. We also observed individual differences in total processing time, a pattern consistently echoed across previous productivity studies, as discussed in Section 5.3 (see for instance Plitt and Masselot, 2010; Guerberof Arenas, 2014a; Bundgaard, 2017a).

NMT content seems to enhance processing speed compared to SMT content – which confirms previous findings reported by Castilho et al. (2017a). While this shows that the improvements in NMT quality provide MT sentences that are more similar than before to human-produced ones (similarly to what Yamada, 2019 and Van Brussel, Tezcan, and Macken, 2018 have shown), the study should be replicated with a different design to verify that a sequence effect has not biased the outcome.

In terms of quality, we found that, on average, a higher number of sentences in PE tasks required further corrections compared to human-pre-translated segments in revision tasks. This percentage was, however, slightly reduced in the second test round, both for post-edited and revised segments. These findings provide support to our hypothesis that NMT output would have required fewer corrections than SMT output but do not explain why fewer issues were found in revised segments in the second test round. A first explanation might be attributed to the fact that segments of different origins were intertwined, and since NMT provided better suggestions than SMT, these sentences are likely to have influenced the overall text quality, prompting more corrections in human-pre-translated segments, too. A second explanation could be that, following the first test round, linguists might have revisited official translations from their archives to validate their work. While the time lapse between the two test rounds should have mitigated the influence of memory cues, we cannot exclude the possibility that participants, having previously worked with the same source text, may have retained some recollection of how to resolve specific issues. Therefore, comparisons between the two test rounds should be approached with appropriate caution.

Fischer and Läubli (2020) found a similar percentage of errors in HT and MT, particularly related to wrong terminology, omissions and typographical issues. These authors also used a design similar to ours, in the sense that MT and HT segments were interleaved for quality assessment, and they note that this "may [have introduced] disfluencies that would not [have occurred] if all segments stemmed from either MT or – particularly – HT" (*ibid.*,

p. 220). In our case, however, CILS evaluators did not comment on issues in coherence or cohesion.

We also acknowledge that text presentation in the CAT tool could have played a role in determining the quality of the final text. Läubli et al. (2020b) tested the impact of text segmentation and orientation in CAT tools on translators' speed and accuracy. Their findings showed that the revision task benefits from document-level presentation, particularly for anaphoric relations, rather than segment-by-segment interfaces. At the same time, segmentation allows for faster within-sentence error identification. Furthermore, while top-bottom orientation is preferred by many, the author found that left-right orientation allows for faster revision. The CAT tool used in our experiment, XTM, employs top-bottom orientation, but the text is segmented. We did not explicitly examine issues in anaphoric relations and other coherence and cohesion markers, nor the evaluators pointed out such issues. However, it is plausible that these aspects were not addressed by the evaluators due to the assessment being conducted on a segmented text within a spreadsheet file, rather than in its final, formatted version.

The fact that some exact matches have been overlooked can be attributed to an excess of trust in TM material, as pointed out by Guerberof Arenas (2014a) and Teixeira (2014b). The only metadata in our testing framework pertained to TM matches, thus this substantiates the likelihood of this scenario. This phenomenon is also congruent with prevailing practices in professional contexts, where explicit instructions often stipulate refraining from modifying exact matches. In professional settings, this is also linked to clear instructions to avoid modifying exact matches. In fact, the modification of such matches is occasionally prohibited, given that these instances often do not warrant compensation for translators. Similar observations regarding issues in TM exact matches find resonance in the research of Daems and Macken (2019).

The implications of our findings concerning linguists' temporal effort in PE and revision tasks hold direct relevance for professional practice. It is evident that a PE task cannot compare, in terms of temporal effort, with a revision task. Consequently, the time allocated to the PE task in professional workflows should be calibrated accordingly.

While our study contributed to the understanding of temporal effort in professional PE and revision tasks, it is important to acknowledge the limitations that impact the scope of our findings. First, the limited number of participants involved in the experiment constitutes a severe drawback. For instance, the data for *de-fr* reflect the performance of only one participant. This prevented us from forming treatment and control groups, and it is a recognised criticism in the existing literature, as it does not make possible to generalise the findings. A second limitation, somehow related to the first one, is that we relied on one single evaluator for each target language. However, as previously noted by Läubli et al. (2019), some domain-specific settings such as CILS do not allow for collecting a large number of participants. While our CILS partner employs more than 20 linguists in house, the productivity study was part of a pilot project involving only a limited number of linguists. Another limitation is related to the methodology employed to record the time spent on the task. Screen recording – and subsequent manual time counting of the time spent on

the segment – ensures ecological validity in settings where in-built plugins are not available for the CAT tool in use (as in our case), but may lack the precision of more advanced approaches like eye-tracking. The latter allows researchers to verify where participant focus their attention at each time of the process. However, an eye-tracking device, too, has its own array of challenges and constraints, notably the difficulty associated with handling longer texts in which the participants need to scroll on the screen (O'Brien, 2009; Hvelplund, 2014).

Future research endeavors could significantly benefit from expanding the pool of participant. Replicating the study with the entire in-house linguist team at the CILS, for example, would yield insights that are more representative of real-world scenarios and enhance the generalisability of the findings. Furthermore, given our primary focus on quantitative data, our study's assessment process lacked a structured evaluation grid for categorising specific types of issues found in the text. Employing a standardised assessment framework in future studies could provide valuable insights into specific issues affecting final texts.

Finally, an aspect that remained unexplored in our productivity study is linguists' attitudes towards technology and MT in particular. Recent work by Briva-Iglesias, O'Brien, and Cowan (2023, p. 6) advocated for the inclusion in productivity studies of MT user experience (MTUX), defined as "[a] person's perceptions and responses resulting from the use and/or anticipated use of MT". The authors measure professional translators' performance (both quality and productivity) when working with static and interactive PE. They found no statistically significant impact of the modality on overall performance (translators are only slightly faster with interactive PE). However, participants reported that interactive PE is more attractive, allows for faster processing and for an increased control of the interaction – a preference also expressed by participants in a study by Daems and Macken (2019). However, there was no correlation between MTUX levels and translators' performance. This finding matched those of Plitt and Masselot (2010) and Läubli et al. (2013), but Sánchez-Gijón, Moorkens, and Way (2019) found that participant' perception of NMT PE effort matched their performance. Undoubtedly, this question deserves further investigation.

## 5.8 Summary of the chapter

In this chapter, we investigated the third key aspect defined under Goal 2: assessing linguists' productivity in performing PE and revision tasks. To achieve this, we conducted a small-scale productivity test within an authentic professional context. This test focused on three language pairs and involved seven in-house linguists.

Unlike prior research that compared PE to translation from scratch, our study contrasted PE with revision. We mainly focused on linguists' temporal effort in completing the assignments. Additionally, the quality of the final texts was assessed to determine whether the traditional two-step translation-plus-revision process could be replaced by a single PE task without compromising quality. The research was undertaken in collaboration with a Swiss CILS in which PE and revision were carried out under the same conditions. Specifically, PE tasks were carried out on fully pre-translated texts and as the last step in the

workflow. Furthermore, CILS linguists were accustomed to working with pre-transled texts, since outsourced translation and in-house revision were common practice.

After reviewing existing literature on the topic (which is abundant for PE tasks and extremely scant for revision tasks), we formulated two research subquestions and corresponding hypotheses. Our first hypothesis posited that post-editing segments translated by the machine would have required more time than revising human-translated segments. We also hypothesised that post-edited segments would have contained more errors compared to revised segments, with NMT possibly mitigating this issue in PE tasks.

Our findings substantiated these hypotheses. We observed that linguists were faster during revision than during PE. Nonetheless, the use of NMT in the PE task appeared to enhance processing speed. Indeed, linguists were significantly faster in post-editing NMT content than SMT content. The quality evaluation performed by CILS evaluators on the final texts revealed that a slightly higher proportion of post-edited sentences in the PE task required further correction compared to the sentences that underwent human revision.

From the linguist's perspective, these findings indicate that performing PE tasks instead of revision tasks does not necessarily allow for time savings. When viewed from a management perspective, our results cast light on the potential quality-related risks of bypassing the verification step performed by a second linguist during PE tasks.

In the upcoming chapter, we will delve into another study conducted within the same CILS context, with a specific focus on the modifications made to pre-translated texts during PE and revision tasks.

# Chapter 6

# Editing pre-translated texts: a corpus-based study using authentic PE and revision data

## 6.1 Overview

In Chapter 2, we have seen that static PE and the use of the batch pre-translate function are common practice in Swiss CILS, and that most often post-edited texts remain unrevised. Hence, it could be said that, in these professional contexts, MT and human pre-translated texts are handled under the most similar conditions. Drawing on these findings, in this chapter we move on to investigate the last key aspect defined under Goal 2, namely how PE and revision of human translation compare in terms of *edits* performed on a pre-translated text. As defined by do Carmo (2021b, p. 3), *editing* is the

> technical dimension of the writing task performed by translators when they apply four actions to previous text: deleting, inserting, replacing, and moving words and groups of words.

Nonetheless, the view of editing that we propose in this chapter is more holistic: it builds upon and further extends the definition provided by do Carmo to encompass *the quantity, the type and the nature of the modifications applied to a pre-translated text.*

Using authentic PE and revision data gathered between April and August 2020 in two language combinations (German to French and German to Italian) from our CILS partner [1], we will examine how the edits made by in-house linguists to raw NMT output and to unrevised translations differ. Following approaches previously used in Laflamme (2009), Parra Escartín and Goulet (2021), and Koponen and Salmi (2017), we will look at the data from different angles, specifically the type of edit performed (i.e., insertion, deletion, substitution and movement), the parts of speech affected (e.g. nouns) and the linguistic dimension involved (e.g. syntax). To complete our analysis, edits will be evaluated in

1. As described in Chapter 5.

terms of their necessity and effectiveness: the former will help us determine whether edits correspond to the correction of actual errors in the two tasks, while the latter will allow us to examine to what extent linguists' interventions improve pre-translated texts' quality. Our study is unique in that it is the first corpus-based investigation of authentic PE and revision texts produced in the same professional setting and within the same conditions by experienced in-house linguists. We maintain that such comparative study of editing yields both didactic and practical relevance.

The remainder of this chapter is structured as follows. Section 6.2 serves as the foundation for our study, delving into the theoretical underpinnings and relevant literature surrounding the topic of editing in Translation Studies and MT research. In Section 6.3, we introduce our research question, along with the variables that will guide our study. Section 6.4 delves into the methodology employed to address our research question. This encompasses a detailed explanation of how our corpus was compiled, a description of the sample used for analysis, as well as an overview of the quantitative and qualitative analyses conducted on our data. In Section 6.5, we present the findings, which will be further discussed in Section 6.6. Lastly, Section 6.7 provides a summary of the chapter.

## 6.2 Editing in revision and PE

This section introduces the topic of editing and provides an overview of relevant literature on editing in revision and PE. We start by explaining what we mean by 'editing' and how the product of various steps in the translation workflow can be used to estimate the underlying process. Then, we review how corpora have been used in Translation Studies and in MTPE research to investigate the changes made to a pre-translated text. Finally, we discuss previous experimental studies that analysed and categorised edits in terms of their necessity and correctness [2]. In these subsections, we first discuss relevant literature from the revision and the PE domain separately, then, whenever available, we analyse studies that focused on the two tasks at the same time.

### 6.2.1 Defining editing

The terminology and specific processes related to editing may vary depending on the context and industry in which it is being used. For instance, in the publishing sector, editing may refer specifically to the process of preparing a manuscript for publication, which can involve tasks such as copyediting, proofreading, and fact-checking. In the film industry, editing may refer to the process of selecting and combining raw footage to create a final video product. In Translation Studies, however, a distinction is usually made between 'editing'

2. Koponen, Salmi, and Nikulin (2019, p. 71) define *correctness* as "[the] accuracy of meaning as well as the grammaticality of the target language". As we will explain later (Section 6.4.5), we adopt a different approach and investigate *effectiveness*, a concept that extends the previous definition to encompass the effect that the edit has on the sentence compared to its unedited version – whether it improves, degrades or has no effect on the sentence. To the best of our knowledge, previous studies on the evaluation of edits have not used the concept of effectiveness.

and 'revising', where the former indicates the correction process made to original texts, and the latter refers to translated texts (Mossop, 2020).

Recently, do Carmo (2021b, p. 5) has provided a comprehensive definition of editing and clarified how this concept relates to translation, revision and PE tasks:

> We define editing as a writing task by which the translator applies one of four editing actions [...] to words or groups of words. The four actions are deleting, inserting, replacing, and moving. Editing is performed in a translation process, for example, when the translator modifies a fuzzy match presented by a translation memory; editing also occurs in a revision process, when the reviser changes a sentence by the translator; and editing is also part of PE, when the translator improves or corrects a suggested translation produced by an MT system.

While 'translating' means generating a text, 'editing' implies modifying an existing piece of text. Although closely related to both revision and PE, do Carmo highlights that editing is not synonymous with these tasks, as it only refers to their technical aspect – the one that involves using the keyboard and the mouse to apply the four editing actions to a (pre)translated text. Our own view of editing subscribes to and at the same time extends the one of do Carmo, in that we use the term 'editing' to refer to all the changes ('edits') performed on a pre-translated text, both from a quantitative and a qualitative perspective. Indeed, the study of editing that we propose in this chapter includes investigating not only the number of edits and the distribution of editing actions performed, but also describing the nature of the edits and evaluating their necessity and effectiveness (cf. Section 6.3.2). Nonetheless, do Carmo's definition provides an excellent starting point to understanding the value of editing in the study of the similarities and differences between revision and PE.

As do Carmo notes, editing is distinct from PE in that the latter typically involves retranslating entire segments from scratch. To distinguish between PE and translation, do Carmo (2021b) and do Carmo and Moorkens (2021) have suggested to set an editing threshold of 25%, similarly to what happens with TM matches. The idea is that, if we modify more than 25% of a machine-translated sentence it means that the usefulness of the MT suggestion is limited and we should consider that we are translating, rather than post-editing, that sentence. We believe that the same line of reasoning could be applied to human-translated sentences during traditional revision. However, to the best of our knowledge, a similar threshold has never been suggested to differentiate revision from translation, probably because, although highly discouraged (Mossop, 2020), retranslation is considered part of the revision process. The amount of editing performed on a sentence, paramount to defining an editing threshold, is usually estimated by comparing the minimum distance between two text strings, the so-called 'edit distance' (Damerau, 1964; Levenshtein, 1966). Perhaps surprisingly, while MT research has made extensive use of edit-distance-based metrics, research on revision in Translation Studies has rarely applied such metrics to revised products (an exception is the study by Macken et al., 2022).

### 6.2.2 Exploring editing through an edit-distance-based metric: from TER to HER

Edit-distance-based metrics have often been used to investigate the technical dimension of PE (Alvarez, Oliver, and Badia, 2020; Koponen et al., 2012; Scansani and Mhedhbi, 2020; Tatsumi, 2010; Temnikova, 2010). Although less accurate than keylogging tools, which detect keyboard and mouse operations (Alves et al., 2016; Lacruz, Denkowski, and Lavie, 2014), edit-distance-based metrics have been extensively used in the literature and in industry settings alike, as they represent a fast and straightforward method to quantify the amount of editing made to raw MT, simply comparing this text with its final, post-edited version.

One of the most used automatic metrics based on editing distance is Translation Edit Rate (TER, Snover et al., 2006), which measures the minimum amount of editing that should be performed on raw MT output to transform it into a reference translation or, most often, into its post-edited version – in this case, the metric is often referred to as Human-targeted TER, HTER. A terminological clarification is due on this point: as do Carmo (2021b, pp. 11–12) explains, there is much confusion around the use of the terms TER and HTER. Indeed, while HTER is described by its authors (Snover et al., 2006) as a human-annotated version of TER, it has been received by the scholarly community as representing post-edited texts. However, the annotation performed to obtain HTER does not correspond to PE, as in HTER annotators are monolingual users instructed to optimize the distance between a reference and the MT output while also preserving semantic relationships. Traditional PE, usually carried out checking the source text, could eventually increase (rather than minimise) the edit distance to the MT output (do Carmo, 2021b, p. 12). We agree with the author that this is a misunderstanding of the original application of this metric and advocate for a use of TER – not HTER – to describe the amount of editing work performed on post-edited texts as well.

Let us now look at how the metric works: TER compares the final text (called *reference*) with the raw MT output (called *hypothesis*) and estimates the amount of changes performed, declined in the four editing actions previously mentioned, namely insertions, deletions, replacements and movements. Its formula is the following:

$$\text{TER} = \frac{\#\text{ of edits}}{\text{average }\#\text{ of reference words}} \tag{6.1}$$

The score results from the total number of operations (edits) performed on the sentence, divided by the number of tokens in the reference (or the average number of tokens, if more than one reference is used). The higher the score, the more the reference differs from the hypothesis[3]. However, if one wants to use TER to study editing, do Carmo (2021b) has

3. It is worth noting that TER only considers exact matches at word level. To address this flaw, the same authors (Snover et al., 2009) have also proposed TER-plus (or TERp), which makes use of stemming, synonymy and paraphrase to evaluate the similarity between the hypothesis and the reference. While TERp has been found to better correlate with human judgement than its predecessor, the features implemented as improvements are not relevant for our study of editing through an edit-distance-based metric.

shown that this metric does not reflect the actual editing operations performed by the linguist, since it is computationally designed as an error rate that focuses on the hypothesis. The author explains this difference as follows (*ibid.*, p.12):

> Referring to edits or errors is not just a terminological difference, but one of perspective, purpose, and method. An error rate looks for what went wrong in the MT output, when compared to the reference, so its focus is on the hypothesis. An edit rate looks at the editing process necessary to correct that error, so its focus is on the reference.

Indeed, the TERcom script [4] estimates the number of editing actions to go from the reference (PE) to the hypothesis (raw MT), while the post-editor does precisely the opposite. Since insertions and deletions are sensitive to the direction of comparison of the sentences, in TER the number of these two editing actions in particular is always reversed. Consider the following pair of sentences:

**Hypothesis (raw MT)**: This is a sentence

**Reference (post-edited)**: This is a modified sentence

Although the post-editor has clearly inserted a word ('modified') in the second sentence, the TERcom script will start its analysis from the reference and will thus record the deletion of one token. To remediate for this issue, do Carmo (2021b) has proposed a new way to use the script, inverting the order of the reference and the hypothesis to obtain a more accurate description of the editing actions performed. Using the edited text as the hypothesis and the raw MT output as the reference, insertions and deletions are correctly estimated by the metric. The count of substitutions and shifts is not sensitive to the direction of comparison and thus it is not affected by this change. The researcher names the result of this new procedure *Human Edit Rate* (HER) and explains that, as opposed to TER, HER can be considered an edit rate, since it provides an estimate of the edits that were applied to the MT raw output to produce the post-edited version. The score indicated by TER and HER for a given sentence are different, since there is a change in the denominator (i.e., the number of tokens in the reference, which is now the raw MT output). To avoid any possible confusion, the author also suggests replacing the term "reference" with "unedited" and hypothesis with "edited".

It is worth noting, however, that HER does not address all the shortcomings of TER. In HER reports, too, there is a high quantity of wrong information related to issues with the edit count and the representation of the edited words. For instance, the script records editing actions performed at the word level, meaning that if the linguist selects two contiguous words and then presses the Backspace or the Delete button on the keyboard — thus performing *one* deletion of two words — the script will record *two* deletions instead. The same happens for substitutions and insertions, as the calculation does not consider linguistically motivated changes; therefore, whenever one word is replaced with a sequence of two words (as in 'did'→'have done') the script will rather record one substitution and one insertion. To address this issue, Blain et al. (2011) proposed to identify *PE actions*, which are minimal

4. Available at `https://www.cs.umd.edu/~snover/tercom/`. Accessed 20 March 2023.

and logical edits performed by post-editors, as opposed to mechanical edits identified by TER/HER. This approach takes into account edit propagation, i.e., changes that are a consequence of another edit, as in the case of modified determiners or adjectives when a noun is modified. PE actions provide a better description of the modification process, but they require manual annotation and are also more complex to identify automatically compared to mechanical edits.

Other issues in the calculation of HER scores concern shifts: unlike insertions, deletions and substitutions, shifts are counted as single operations, regardless of the number of contiguous words shifted. The TERcom script indicates this information in a separate category (*words shifted*, or WdSh), but this value does not always correspond to the actual number of words shifted. To compute the score, the script considers the sum of insertions, deletions, substitutions and shifts – not WdSh –, thus it is difficult to interpret HER as the actual percentage of words modified. Another issue is that the application of HER instead of TER implies a change in the position of the word shifted. Although this does not affect the number of shifts performed, it could nevertheless represent an issue if one uses results from the TERcom script to describe the editing process more in detail. Finally, as noted by do Carmo (2021b, p. 26), "the problems in identifying the edits grow as the edit scores increase, especially above 20%". The author (*ibid.*, p. 26–27) demonstrated that HER overestimates the overall number of edits and that it is also biased towards replacement and against movement.

Before concluding this section, it is important to note that several authors have argued that edit-distance-based metrics provide only a rough estimation of the actual editing process, which can be better observed using keystroke logs (Cumbreño and Aranberri, 2021; Daems et al., 2017a; Elming, Winther Balling, and Carl, 2014; Koponen et al., 2012). Indeed, as do Carmo (2021a), the simple, 'static' comparison of two products cannot capture the complexity of a dynamic process such as the editing one. However, tools such as Translog (Carl, 2012) or Inputlog (Leijten and Van Waes, 2013) — extensively used in Translation Process Research — produce reports that are often difficult to interpret in terms of editing actions and also difficult to apply at scale (do Carmo, 2021b). Furthermore, as do Carmo has shown (*ibid.*), Inputlog disables some functions in Microsoft Word, such as drag and drop and the selection of several words using the Ctrl key, affecting the editing process and potentially undermining the ecological validity of a study. In this scenario, do Carmo (*ibid.*, p. 4) advocates for a view of editing as "an element that could help bridge the gap between [Translation Process Research] and MT". Although with several shortcomings, "[HER] presents an improved perspective on actual editing, and when researchers aim at improving process estimation methods, [anything] that brings them closer to that process should be preferred" (*ibid.*, p.31).

### 6.2.3 Quantifying editing actions

#### Revision

As noted earlier, revision scholars have rarely relied on edit distance-based metrics to analyse the revision product. In general, it could be argued that editing as a technical dimension has been largely overlooked in revision.

Horváth (2009) was one of the few researchers who described editing actions in revision products. The author manually identified five revision operations (i.e., insertion, deletion, replacement, rearrangement and annotation) finding that substitutions are the most frequent category, followed by additions and deletions. Using the same categorisation of revision operations, Robin (2014b) found that the most frequent operation carried out by revisers is deletion – contributing to increasing lexical density – and the second most frequent one is insertion – in order to improve text cohesion.

Different results have been obtained by Pontrandolfo (2017), who quantified the revision actions performed by different types of revisers on a legal text. Following a categorisation established by Parra Galiano (2015) that includes six revision actions (i.e., insertions, deletions, replacements, paraphrase, marking and annotation), the author found that substitutions are the most frequent ones, followed by insertions and paraphrases, while deletions are not often performed by revisers in this case study. The author reports that the analysis of editing actions has been carried out manually on the final product, without using any semi automatic metric or statistical software [5]. A manual identification of the revision actions performed by professional revisers was also carried out by Laflamme (2009), who, focusing on lexical modifications, found that words are most often replaced, deleted or inserted, while they are only very rarely shifted in a different position. Finally, in an experimental study on the acquisition of revision competence by translation students, Robert, Remael, and Ureel (2017) used the four main editing actions (insertions, deletions, substitutions and shifts) to classify instances of preferential changes ("hyperrevisions"). The authors found that substitutions are the most frequent type of hyperrevision, followed by insertions, shifts and deletions.

In conclusion, it appears that most researchers in the field of revision tend to agree that substitutions constitute the most common editing action in revised products. However, there is less consensus when it comes to the distribution of other editing actions. One might contend that the language of the revised text could have influenced these observations, but it is noteworthy that even researchers focusing on the same target language (such as Horváth, 2009 and Robin, 2014b for Hungarian, or Laflamme, 2009 and Robert, Remael, and Ureel (2017) for French) do not come to the same conclusions. Therefore, we could attribute this lack of consensus to the different data analysis methods employed by the researchers.

5. "The analysis is both quantitative and qualitative and has been carried out manually by reading the material, i.e., no statistical tools or semi-automatic analysis software have been used". [El análisis realizado es de tipo cuantitativo y cualitativo y se ha llevado a cabo manualmente mediante lecturas del material, es decir, no se ha empleado ninguna herramienta estadística ni programas de análisis semiautomáticos.] Pontrandolfo (2017, p. 122), our translation.

**Post-editing**

Within PE research, TER has commonly been utilized to examine the distribution of editing actions in post-edited texts. According to do Carmo et al. (2021, p. 122), well-balanced APE systems using TER as automatic metric consistently report the following percentages of edit types: around 54% substitutions, around 20% deletions and 15% insertions, while shifts are usually below 10%. Depending on the MT system architecture, the language pair considered, and the method used to compute the number and type of editing actions performed, studies on human PE have reported different results. For instance, in an experience with Finnish translation students, Koponen, Salmi, and Nikulin (2019) found statistically significant differences in the distribution of edit types among different MT systems: while replacements are the most frequent edit type overall, deletions are most often implemented to the output of RBMT systems, and insertions are the second most frequent edit type in the output of SMT engines.

In a study on authentic texts post-edited by French-speaking professional translators (Mutal et al., 2019), we reported the same distribution of edit types between SMT and NMT output, namely replacements, deletions, insertions and shifts, but recorded a higher proportion of deletions (i.e., words deleted) in SMT. In this study, we employed TER to grasp the distribution of edit types in post-edited texts, but a different picture is depicted when counting the editing actions manually. For example, Aranberri (2017) reported that professional translators working into Basque with RBMT and SMT generic systems carry out replacements and shifts most often than insertions and deletions. Conversely, in two studies with non-translators who carry out PE into English, both Goulet et al. (2017) and Parra Escartín and Goulet (2021) reported that insertions and deletions are the second and third most frequent edit types, respectively.

While findings from the abovementioned studies agree that replacements are the most frequent type of change, results concerning the other editing actions are divergent. In this regard, the different text types, languages and post-editors'/revisers' profiles considered play certainly a role, as does the lack of a common methodology to identify editing actions. In fact, while the majority of revision scholars do not provide details on how editing actions are determined, Laflamme (2009, p. 76) explains that addition and substitutions can apply to "characters [or to] lexical unit[s]"[6], thus, it is unclear how misspelled words are categorised. In this context, the use of an automatic metric such as HER to categorise editing actions could provide an objective basis for the comparison of the editing work performed by post-editors and revisers.

6. "An insertion is an addition of (one or two) characters or of a lexical unit where there was none before". [L'addition est un ajout de caractères (un ou deux) ou d'une unité lexicale là où il n'y avait rien auparavant.] Laflamme (2009, p. 76), our translation.

### 6.2.4 Investigating editing through product data

The analysis of corpora of edited texts can provide valuable insights into the patterns of changes made during revision and PE tasks. In thesis, we understand *corpora* broadly, as collections of texts produced in real-life contexts (therefore, data with high ecological validity) and analysed both manually and automatically [7].
In this section, we review how corpora have been used in revision and PE studies (as well as in studies focusing on both tasks) to investigate the modifications made to pre-translated texts.

#### Revision

Several revision scholars have used corpora of revised texts to quantify and classify revisers' interventions. For instance, Rega (1999) examined the changes made on a small corpus of texts from three domains [8], and classified them as stylistic subjective, stylistic objective and domain-specialist interventions. These kinds of changes do not correct proper errors, but rather aim at improving the text, making it more readable or compliant with the requirements of a specific genre. Although the author did not provide quantitative data on the distribution of these categories, she did provide several examples of each.

Solum (2018) reported some figures on the number of changes made to translated texts and quantified the comments made by different Norwegian copy-editors to 26 excerpts from 13 literary texts, finding high individual variation. The focus of her research, however, was on determining the influence of copy-editors in literary translation. Therefore, the researcher discussed how many of the changes proposed by copy-editors are accepted, modified or rejected by translators – while the type of intervention proposed was not further discussed.

A different perspective was adopted by Popič (2014), who used corpora to delve into language policy issues. The researcher compiled and analysed a corpus of changes made to authored and translated texts to unveil the pitfalls of *Lektura*, the translation correction procedure most widespread in Slovenia, which is carried out most often by people who are not translators and without the source text. This latter point was also brought about by Mc Donough Dolmaya (2015), who analysed a corpus of 29 articles translated from Spanish and French into English and revised by voluntary, often non-professional revisers in the Wikipedia platform. Focusing on transfer and language/style issues in pre-translated texts, the author found that Wikipedia users detected and solved the majority of language errors, more often than transfer errors (especially omissions), suggesting that source texts are not always consulted during revision. However, while these findings shed light on the quality of crowdsourced translations, they do not say much of typical revised texts, since in Wikipedia

7. For a discussion on the ambiguity of the definition of *corpus* in Translation Studies and in Corpus Linguistics, the reader can refer to Granger and Lefer (2022, pp. 16–17).

8. Since the size of this corpus is quite small, Robert (2008) included this paper among case studies of the revision product. However, we decided to include it in our review of corpus-based studies as it does involve more than one text or text type.

"the translations can never be considered definitive versions, but rather evolving texts [...]" (*ibid.*, p.29).

Among corpus-based studies that compared revised and unrevised text, the one carried out by Robin (and reported in several publications, namely Robin, 2014a, 2018, 2019) is particularly interesting, in that it considers both the type of revisional modifications made to the text and their effect on the so-called 'translation universals' (Baker, 1993, 1996), i.e., the distinctive features of translated texts that set them apart from texts originally drafted in the same language [9]. The author analysed, both quantitatively and qualitatively, a parallel corpus of ten English novels translated into Hungarian, to investigate to what extent translators and revisers apply explicitation and implicitation strategies. The quantitative analyses consisted in comparing the subcorpora in terms of text length, average sentence length, lexical variety (expressed as type-token ratio, TTR – the quantity of different words on the total number of words in the text) and lexical density (the quantity of content words on the total number of words in a text). Differences between subcorpora have been tested for significance via parametric and non-parametric tests. A subset of data has also been analysed from a qualitative perspective, through a manual identification of the type of transfer and revisional operations carried out by translators and revisers – for instance, determining whether they concern grammar or lexis. All interventions were then classified as instances of explicitation or implicitation and categorised as mandatory or optional.

The results showed that, beside correcting translators' grammatical and spelling mistakes, revisers edit the text independently, performing several explicitation and implicitation operations. Revised texts resemble each other more than draft translations, and a tendency towards avoiding grammatical and lexical redundancy is noted. Considering the above, the researcher went further claiming that translation universals can rather be viewed as the result of editing strategies. It is worth noting, however, that these considerations were mainly based on the qualitative analysis and that statistical analyses conducted on the data had not detected any significant differences between draft translations and revised texts. This was considered expected by the author herself, since revisers are professionals who know they need to intervene as little as possible. Therefore, the researcher claimed that finding significant differences would have meant that revisers had completely retranslated the content (Robin, 2019, p. 143).

9. Relatively few scholars investigated the impact of editorial interventions on the universal features of translation (Bisiada, 2017, 2018a,b; Kruger, 2017; Serbina, Bisiada, and Neumann, 2021), obtaining rather inconsistent results – probably due to differences in study designs, corpus composition and language pairs considered. Furthermore, in many of these studies it is unclear what the editorial process really entails, i.e., who is the editor and whether he or she checks the source text during the process. For instance, as Bisiada (2018a) points out, editors usually intervene after translation revisers, thus modifying a text that results from the work of two different people — see also Moe, Juznic, and Zigon (2021) for a discussion of the role of different actors in the publishing process.

### Post-editing

In PE research, corpora have rarely been used to analyse post-editors' corrections. Rather, corpora of post-edited texts have been employed for two main purposes: first, to train APE systems (cf. do Carmo et al., 2021 for a review). To reduce human PE effort, these systems learn from the corrections made by humans to the raw MT output during the PE process and use this information to improve the MT output, namely fixing systematic errors or replacing words with more specific ones in a given context.

Second, corpora of machine-translated and post-edited texts have been investigated in the context of the so-called post-editese research (Daems, De Clercq, and Macken, 2017). This fairly recent strand of work has shed light on the linguistic features that differentiate post-edited output from human translations, such as average sentence length, lexical variety and lexical density. The term "post-editese" has been conceived along the lines of translationese, or "[the] set of features common to [human-translated] texts that differentiates them from [original texts]" (Aranberri, 2020, p. 93). Since many of the features that characterise post-editese were already typical of translationese [10] (cf. Volansky, Ordan, and Wintner, 2015), the former has been defined as "an exacerbated translationese" (Toral, 2019, p. 1).

Researchers working in this area have found that, while readers are unable to distinguish between MTPE of publishable quality and human translation (Daems, De Clercq, and Macken, 2017), computational analyses allow us to unveil and quantify the influence of the production mode on the final product. Indeed, nowadays we know that post-edited texts are in general simpler – in terms of lexical density and lexical variety – than human translations (Martikainen and Kübler, 2016; Toral, 2019). Furthermore, since post-editors are primed by MT suggestions (Green, Heer, and Manning, 2013; Čulo and Nitzke, 2016), post-edited texts are found to be more similar to their source texts in terms of sentence length and POS sequences (Toral, 2019). For the same reason, terminology in post-edited texts has been found more similar to raw MT output than to human-translated texts (Čulo and Nitzke, 2016). Overall, these features [11] lead to simpler texts with a greater interference from the source language, causing concerns about the implications for the target language in the long run (Toral, 2019).

It should be noted that much of the research on post-editese has been conducted on parallel texts specifically created for experimental purposes (cf. Castilho, Resende, and Mitkov, 2019; Farrell, 2018; Castilho and Resende, 2022), which are usually not considered corpora as such. However, ongoing research is now focusing on authentic, comparable corpora from in-house language services (Volkart and Bouillon, 2022).

10. Stemming from results in translationese and post-editese studies, researchers further investigated the characteristics of raw MT output, finding evidence of the so-called "machine-translationese" (Loock, 2018; Webster et al., 2020; Vanmassenhove, Shterionov, and Way, 2019; Vanmassenhove, Shterionov, and Gwilliam, 2021; De Clercq et al., 2021; Brglez and Vintar, 2022; Luo and Li, 2022).

11. It is worth noting, however, that studies on post-editese features have reported mixed results, depending on the corpora used, language pairs, PE level and MT architectures considered.

### Revision and post-editing

Although corpora have been used in Translation Studies and, to a lesser extent, in PE research, contrastive corpus-based analyses of revision and PE are extremely limited. To the best of our knowledge, the only study of this kind is the one by Vardaro and colleagues (Vardaro, Schaeffer, and Hansen-Schirra, 2019b,a), who analysed a parallel corpus of 24 English source texts (902 segments, 17925 source tokens) and their translations produced at the German Department of the DGT. The corpus consisted of raw MT segments (both NMT and PBMT), NMT post-edited (NMTPE) and NMT post-edited and revised segments (NMTpeREV). Using both automatic and manual evaluation methods to identify the changes made to pre-translated texts, the researchers compared raw MT segments against NMTPE segments, and NMTPE against NMTpeREV, respectively. After a first automatic error annotation using Hjerson (Popović, 2011), a more fine-grained manual annotation using the Multidimensional Quality Metrics (MQM) [12] framework has been carried out on a subset of errors. Texts have also been evaluated using automatic metrics, such as BLEU (Papineni et al., 2002). In addition, subcorpora were analyzed for number of tokens, average sentence length, TTR and lexical density. Results for automatic metrics showed improvements for NMT segments compared to PBMT ones. This was expected, since all raw MT segments have been compared to post-edited segments produced by the NMT engine; therefore, raw MT segments produced by the PBMT engine were clearly penalized. It is worth noting that many automatic scores for MT evaluation, especially BLEU, are meaningless if taken alone and make sense mostly in balanced comparisons (Callison-Burch, Osborne, and Koehn, 2006; Ananthakrishnan et al., 2007). Furthermore, automatic measures used on NMTPE segments are not fully comparable to other reported automatic scores, since different reference texts have been used.

In the same study, the number of tokens was found to be smaller for revised segments than for source texts, and only slightly smaller for revised segments compared to raw or post-edited segments. Average sentence length, measured in words, is shorter for revised segments than for source, raw MT or NMTPE segments, while lexical density remains constant across the subcorpora. Regarding TTR, it is consistent between raw NMT, NMTPE and NMTpeREV, but it is overall higher in German than in source texts, meaning that target texts "are slightly more specialized than the corresponding English [source texts]" (Vardaro, Schaeffer, and Hansen-Schirra, 2019b, p. 2). The automatic annotation is carried out on raw MT and on the PE subcorpus to identify five error categories taken from Vilar et al. (2006), namely extra word, missing word, inflection, reordering and lexical error. Compared to PBMT, NMT system makes fewer errors in all categories, but especially with inflectional errors, while only slight improvements are reported for missing and extra word. For both raw and post-edited segments, the most frequent category was lexical errors, with missing word in third position. Reordering and extra word were the second most frequent

12. `https://themqm.org/error-types-2/typology/`. Accessed: 4 February 2023.

categories in raw and post-edited segments, respectively. However, as the authors themselves report (Vardaro, Schaeffer, and Hansen-Schirra, 2019a, p. 8), a main drawback of Hjerson is that it includes punctuation issues into the lexical error category. That is why the authors decide to carry out a manual annotation of errors in the lexical, missing and extra word category, i.e., the most frequent error category, and the two categories where only a slight improvement compared to PBMT was noted.

Results of manual annotation showed that error distribution in raw and post-edited segments is similar, meaning that post-editors and revisers correct the same types of errors, mostly register errors (called "stylistic changes"), function words, mistranslations and terminology errors. Nevertheless, the authors clarify that they do not know whether reviser changes affect the same words that post-editors changed (meaning that post-editors' corrections were not satisfying), nor whether revisers modifications could be considered essential or preferential (*ibid.*). It would have been interesting to verify these points, and also to see whether and in which cases revisers came back to the original NMT proposal. Furthermore, errors were manually annotated per sentence, but it is not reported whether annotators (one or many) had access to some more context in order to classify the modifications. This is relevant, since it has been shown that taking isolated segments for manual assessment, i.e., out of their original context, can significantly alter the results (Läubli et al., 2020a). As additional analysis, the authors use simple linear regression models with the same error category as predictor and dependent variable to verify whether errors in raw NMT output can influence the number of errors in post-edited segments, per sentence. Results showed that the effect of both stylistic changes (i.e., register errors) and mistranslations (including terminology errors) between raw NMT and post-edited NMT is positive and significant ($p<0.05$), meaning that post-editors are primed by the NMT output — as already showed by Čulo, Hansen-Schirra, and Nitzke (2017).

### 6.2.5 Evaluating edits

Both in revision and in PE research, the modifications made to pre-translated texts have been also investigated through experimental and observational studies. To better understand the nature of these modifications, scholars have developed various taxonomies to categorise edits, often with a focus on determining their necessity and correctness.

#### Revision

Revision scholars have proposed to distinguish between necessary – i.e., justified – interventions, "underrevisions" (for overlooked or failed necessary changes), "overrevisions" (for changes that introduce an error) and "hyperrevisions", for unnecessary changes (Brunette, Gagnon, and Hine, 2005; Künzli, 2007; Robert and Van Waes, 2014).

A typology of revisional modifications that considers not only the necessity of the intervention, but also its motivation, has been outlined by Robin (2018). It comprises four categories: the first one is called "rule-based" and involves mandatory interventions, motivated by grammar, syntax and semantics. Two categories deal with optional modifications,

which can be either "norm-based" or "strategy-based" (also called "editorial" operations): the former include interventions "motivated by language customs and stylistic preferences", while the latter category involves modifications "motivated by general communicative principles and text building strategies [that promote] the processability and readability of the text" (*ibid.*, p. 159). Finally, the fourth category includes preferential, unnecessary modifications. While rule-based modifications *correct* the translation, and optional modifications *improve* it, preferential interventions do not have any positive effects on the target language text and therefore should be avoided. In her corpus-based study of revisional modifications in literary texts, the author found that revisers mainly performed strategy based operations, which "[optimise] the lexical and grammatical redundancy, [create] a balanced text and [enrich] the vocabulary" (*ibid.*, 160). While Robin's taxonomy can be considered a valuable tool to justify the changes made to a translation, it does not assess the correctness of the modifications applied.

Conversely, Rensburg (2017) proposed an instrument to assess revisers' changes across three dimensions, namely (1) the necessity of the intervention, (2) the importance of the error to correct, and (3) the effect that errors have on translation accuracy, target language usage (i.e., fluency) and the function of the target text (i.e., register, level of formality and genre conventions). For example, the correction of an omission constitutes "a necessary change made to correct a major error in translation accuracy" (*ibid.*, p. 74). In addition, the author proposes a formula to quantify the quality of the revision product (drawing on previous proposals by authors such as Arthern, 1987; Künzli, 2007; Robert, 2012; Mossop, 2020). In an empirical study with 30 professional revisers, the author tried to determine whether there is a relationship between revisers' profile (in terms of qualifications and experience) and the quality of the revised product but could only find a statistically significant correlation between translation experience and necessary changes that correct fluency errors. Surprisingly, the findings shed light on a marked tendency to overlook errors – and, to a lesser extent, also to make unnecessary changes and introduce errors – which applies even to experienced revisers.

Other researchers engaged in a comprehensive description of the modifications carried out by professional revisers. For instance, Laflamme (2009), proposed to categorise lexical modifications considering the revision operation performed (as discussed in Section 6.2.3), the linguistic unit concerned and the linguistic dimension involved, i.e., whether the change has to do with morphology, semantics, syntax, style or typography. In an observational study with six revisers working on authentic revision assignments, the researcher found that, on average, modifications of lexical units account for 44% of the total modifications made to translated texts. In the study, however, 'lexical units' are intended in the broad sense, meaning that they comprise not only nouns, but also verbs, adjectives, adverbs, pronouns, prepositions, determiners and conjunctions [13]. The findings show that lexical modifications

13. "I have deliberately considered the lexicon in its broadest sense, i.e., 'all the lexical units - or words - of a given language' [...]. Thus, both grammatical and lexical words can undergo lexical modifications". [*J'ai volontairement considéré le lexique dans son acception la plus large, c'est-à-dire « ensemble des unités*

mainly concern – in order – verbs, nouns and adverbs, and primarily the linguistic dimension of semantics, followed by syntax, morphology and style. Through the analysis of revisers' retrospective verbalizations and interviews, the researcher has also identified the motivation behind the changes performed by the professionals – although some of the changes could not be sufficiently justified by the revisers themselves. While the perspective adopted to describe revisers' interventions in this investigation is new compared to previous studies, neither the necessity, nor the quality of the changes performed are considered.

**Post-editing**

Studies in the PE domain have often examined the necessity of post-editors' interventions. However, an important factor influencing the distinction between necessary and preferential changes in PE is the type or level of PE being performed. PE guidelines aim to determine what should and should be not corrected in different PE assignments, but can be challenging to implement for MT users (Flanagan and Christensen, 2014; Hu and Cadwell, 2016). In general, linguists working with MT are instructed to minimise the number of edits, reusing the MT output as much as possible and avoiding purely stylistic modifications (Massardo et al., 2016).

In her PhD thesis, de Almeida (2013) investigated the changes made during PE by linguists with varying levels of translation and PE experience. Edits were classified as essential, preferential, essential errors not implemented or errors introduced. Essential changes are those that, if not implemented, leave the sentence grammatically incorrect, unintelligible or semantically inaccurate (*ibid.*, p. 100). Although PE guidelines for the task explicitly stated that preferential changes should be avoided, all participants implemented this type of changes. However, the study does not evaluate whether these preferential changes lead to any (desirable) improvements. Additionally, all participants overlooked certain errors that required correction and even introduced some errors during the PE task. To classify PE changes, the researcher devised a customised typology based on widely used MT error taxonomies in the localisation domain (such as LISA QA 2009). The focus of de Almeida's typology was on describing the corrections made, rather than the specific errors being addressed. The findings revealed that the most frequent edits made by post-editors in the study were at the language level (that is, grammar and syntax). This is quite logical, given that the study makes use of an SMT engine, which is known to produce several fluency errors (Castilho et al., 2017b; Klubička, Toral, and Sánchez-Cartagena, 2017).

This finding is echoed in a pilot study by Koponen and Salmi (2017), who investigated the quality of PE changes made by translation students during a light PE task. Drawing on the results from HTER reports, changes were categorised as form changed, word changed, deletions, insertions, order changes and multiple changes. The authors found that the majority of edits implemented consisted in modifying the morphological form of the

*lexicales – ou mots – d'une langue donnée » [. . . ]. Ainsi, autant les mots grammaticaux que les mots lexicaux peuvent être la source de modifications lexicales.*] (Laflamme, 2009, p. 49, our translation).

word and most of them were deemed necessary. Conversely, word-order changes and word substitutions were often considered unnecessary. Overall, 38% of the edits were categorised as 'unnecessary' (*ibid.*, p. 144). While nearly all changes were correct, 9% of the changes introduced an error in the final version.

In a follow up study, Koponen, Salmi, and Nikulin (2019) found that unnecessary edits were prevalent, particularly for the RBMT output, where they accounted for 61% of the total number of edits performed. Although the participants were translation students with limited PE experience, the majority of their edits were deemed correct, with only up to 6% of edits made to the output of each MT system were labelled as incorrect (*ibid.*, p. 75).

Aranberri (2017) has also demonstrated that, when post-editing the output of an SMT system for the first time, professional translators tend to primarily make essential changes, but also implement several preferential edits.

Some interesting work on the different types of edits performed during PE has also been carried out with non-translators (Goulet et al., 2017; Parra Escartín et al., 2017; O'Brien, Simard, and Goulet, 2018; Parra Escartín and Goulet, 2021). Parra Escartín et al. (2017) and Parra Escartín and Goulet (2021) investigated the use of MT as an aid for academic writing. Spanish-speaking physicians were asked to draft an excerpt of an article in their native language and field of expertise, and then to self-post-edit the English translation obtained through Google Translate. Subsequently, a professional English proofreader specialised in the medical domain was recruited to correct the texts. The researchers compared the edits made by non-translators during self PE with the edits made by the proofreader on the final versions. Following the methodology devised by Laflamme (2009) for revised texts, edits were analysed in terms of the editing operation performed, the linguistic unit affected, and the linguistic dimension involved. Additionally, all the changes were labelled as essential or preferential, and overlooked and introduced errors were also analysed. The findings revealed that the proofreader made a greater number of edits compared to the medical practitioners, indicating that the latter had overlooked several errors. Despite the guidelines for the proofreading task specifying to avoid stylistic changes, the professional proofreader mostly implemented preferential changes of this kind and even introduced some errors. Edits most frequently involved phrases, nouns or verbs and mainly concerned the linguistic dimensions of style or syntax.

#### Revision and post-editing

Few researchers have focused on a comparative evaluation of edits performed during revision and PE. One such study was conducted by Nitzke and Gros (2021), who investigated overediting instances (i.e., preferential changes) in revision and PE, reanalysing some datasets from previous studies, one dealing with revision and two with PE. The researchers found that over-editing instances are linked to stylistic and lexical preferences and mostly occur during revision, probably because "less obvious mistakes could be found in [this task]" (*ibid.*, p. 31). However, the contexts considered are hardly comparable and it is therefore difficult to draw any definite conclusions.

In another experiment by Daems and Macken (2021), overediting instances were estimated using TER. The researchers determined a baseline score reflecting only necessary changes and then compared against this baseline TER scores obtained by professional translators during PE and revision tasks. They found that several preferential edits were performed in both conditions, but also that "more editing generally leads to higher quality" (*ibid.*, p. 64).

More recently, Macken et al. (2022) investigated the changes performed on an English novel machine-translated into Dutch using DeepL, post-edited and then revised by two different professionals (one per task). Using a selection of automatic evaluation metrics as well as manual error annotation on a subset of sentences, the authors aimed at "[measuring] the (dis)similarity between the consecutive versions and [analyse] the linguistic characteristics of the three translation variants" (*ibid.*, p. 101). Automatic metrics showed that the post-edited version was linguistically more similar to the raw MT version than it was to the revised version, meaning that the reviser made a great contribution towards enhancing the lexical richness and lexical variation of the text. The fine-grained manual error annotation step also let the authors determine the type and purpose of interventions performed by the linguists during PE and revision. Edits were labelled following a linguistic typology that includes four macro-categories, namely lexico-semantic, spelling and punctuation, style, syntax and morphology, each with several subcategories. Edits performed by the post-editor were also annotated as MT error correction, consistency, preferential and undesirable modification. During the revision step, these edits were complemented with the PE error correction category, to account for errors introduced by the post-editor and amended by the reviser. Finally, all errors remaining in the final texts were also marked during the annotation process.

Results showed that, while the majority of post-editor's interventions are aimed at correcting MT errors, most of the changes made by the reviser are considered preferential (as in de Almeida, 2013 and in Robin, 2019). The vast majority of edits in both PE and revision pertained to the lexico-semantic category, but to different subcategories. Compared to the post-editors, the reviser also performed several structural changes, explicitations and more sentence splitting. Finally, although most of the MT errors were amended during the PE step, a small number of errors persisted in final texts. Perhaps surprisingly, the reviser made overall more changes than the post-editor, most of them aimed at improving the readability of the text – which confirms findings from previous studies carried out in the publishing sector (cf. Kruger, 2017).

It must be noted that, while PE was carried out in a CAT tool, revision was performed in Microsoft Word and mainly as a monolingual process in which the reviser consults the source text in case of issues in the target. Some criticism can be directed at the error taxonomy used to annotate the texts: the four macro categories used are concise but comprehensive enough to describe the edits performed on both human and machine translated texts. However, the corresponding subcategories seem unsuited to capture the linguistic aspects modified in post-edited and in revised texts, as the majority of edits are classified in what should be residual categories (i.e., "Other"). As the authors themselves recognise, the major limitation

of this study is that it involved only one reviser and one post-editor, and the latter had limited experience in the literary domain. Thus, the results mirror the distribution of edits of two different people, and also cannot be considered typical of the revision and of the PE task, since the revision is carried out on a post-edited text and not on a traditional human translation.

## 6.3 A corpus-based study of editing in authentic revision and PE assignments

In this section, we outline the rationale for the study and introduce the research question that will guide our investigation, along with the variables that will be examined.

### 6.3.1 Rationale

As it can be inferred from what has been discussed thus far, several studies have already examined the modifications made to pre-translated texts, but they tend to focus mostly on PE or revision independently. Some studies have looked at the two activities together, yet they are either based on data collected through experiments – meaning that they examine only one or very few texts and text types (as in Daems and Macken, 2021 or in Macken et al., 2022) — or consider revision as a further step in the PE workflow (as in Vardaro, Schaeffer, and Hansen-Schirra, 2019b), which does not allow for a direct comparison of the two tasks.

In our work, we propose an analysis that investigates these two tasks simultaneously, i.e., taking place at the same level and in the same workplace setting. Indeed, in our study PE and revision are carried out as separate services and both represent the last step in a QA workflow. The same quality standards and the same guidelines apply to the correction of human translated and machine-translated texts (as explained in Chapter 5). For instance, both tasks involve the bilingual reading of source and target. Finally, texts included in our study have been post-edited or revised by highly experienced professionals as part of their usual workflow. This setting guarantees that the results are directly comparable and allow us to single out the variability across activities that is due to the intrinsic differences between PE and revision, leaving out the variability that could be traced back to the differences between the nature of texts and professional contexts. Lastly, individual variability is mitigated by the fact that we included texts from several professionals with the same background and similar work experience.

Although our analysis is based on the product of PE and revision, our aim is to get some insights into the process as well. To this end, we will use a relatively-new edit-distance-based metric from MT research, Human Edit Rate (HER, do Carmo, 2021b), which provides a more accurate estimate of insertions and deletions than its original counterpart, TER (Snover et al., 2006), as detailed in Section 6.2.4. As Gaspari (2022, p. 52) argued, "there is ample opportunity for [corpus-based translation studies] to apply its methodologies and techniques in the process-oriented investigation of editing, revision and post-editing". We

would claim that the opposite is also promising, as long as new techniques are found to analyse the process in large amounts of textual data. Applying the HER metric on corpora of post-edited and revised texts can serve as a first step toward this objective. Although we are well-aware of the shortcomings of using an edit-distance-based metric to describe the process, we find that its main advantage lies in its capacity to simplify a complex process as the editing one (do Carmo, 2019). In this regard, HER provides us with a simple, but more accurate estimation of the editing process compared to its predecessor. Used as a common metric to investigate revision and PE, HER can serve as a key to shed new light on the modifications made to a pre-translated text during these two tasks.

We contend that a comparative study of editing in PE and revision is relevant for both didactic and practical purposes. Analysing the kinds of changes that professional linguists perform on human-translated and on machine-translated texts provides valuable information to translation instructors, who can adapt or develop new strategies to raise students' awareness towards issues in different types of pre-translated texts. Practitioners, too, can benefit from a comparative description of modifications in PE and in revision. In particular, results could prompt other CILS and in-house translation teams in different organisations to initiate a critical reflection on the effectiveness of their interventions on pre-translated texts, as well as on the implications of working with texts produced by humans or machines. Finally, as investigating the changes made by linguists to pre-translated texts implies delving into the characteristics of these texts, with our study we are contributing to existing knowledge on the differences between PE and revision and, by extension, between NMT and human translation.

### 6.3.2 Research question and variables

The study described in this chapter seeks to answer the following research question:

**RQ4**: *When PE and revision are carried out under the same conditions, how do PE and revision compare in terms of edits made to pre-translated texts?*

In order to answer the abovementioned question, we compiled and analysed two comparable corpora of texts resulting from PE and revision assignments, respectively. We therefore adopted a product-based approach, investigating the textual outcome of PE and revision tasks (Saldanha and O'Brien, 2013). Compared to controlled experiments, the analysis of corpora has the main advantage of allowing us to investigate a wide range of authentic texts. It is important to note that the texts included in our corpora were post-edited or revised by professional CILS linguists as part of their daily workflow, and none of the linguists involved knew that their work would later become part of this study. Therefore, the material investigated represents naturally occurring data and guarantees high ecological validity. The independent variable (ID) considered in this study is the text production mode (namely, *MT plus PE* or *human translation plus revision*), and we investigate whether the latter has an impact on the following dependent variables (DV):

**DV6.1** — *Amount of editing performed*: this variable refers to the total number of edits made to the texts, as recorded by the HER metric (do Carmo, 2021b).

**DV6.2** — *Distribution of editing actions*: this variable will also be measured using the HER metric, but instead of focusing on the overall score, we will look at the amount and type of editing actions performed, namely insertions, deletions, substitutions and shifts of words or groups of words.

**DV6.3** — *Parts of speech (POS) affected*: this variable concerns the parts of speech that are most frequently affected by editing.

**DV6.4** — *Linguistic dimension involved*: this variable points to the linguistic dimensions that are most frequently affected by linguists' edits during PE and revision, such as semantics, syntax, and style. The taxonomy used for the linguistic annotation of edits will be presented in Section 6.4.4.

**DV6.5** — *Necessity of the edits*: this variable refers to the extent to which the edits were necessary to correct errors, to improve the quality of pre-translated texts or to comply with in-house guidelines.

**DV6.6** — *Effectiveness of the edits*: this variable relates to the extent to which the edits improved, degraded or had no effect on pre-translated texts.

The results of our analyses will be presented in Section 6.5. The analyses will be conducted on the entire corpus; however, special attention will be given to a specific subsample of sentences, i.e., those falling in the HER range 1-25. This is because one of our objectives is to test the editing threshold proposed by do Carmo and Moorkens (2021) to distinguish "proper" PE from translation. Further details regarding the data and the methodology employed to address our research question will be provided in the next section.

## 6.4 Data and methods

To analyse the products of PE and revision, we compiled two corpora, one consisting of authentic PE jobs and the other of authentic revision jobs carried out at the CILS partner. In this section, we start by describing the CILS partner, then we detail how the corpora have been constructed and what kind of analyses have been performed.

### 6.4.1 Background

The corpora have been collected from Swiss Post (cf. Chapter 5). After an 18-month testing phase, at the time of data collection (September-October 2020), the customised NMT system had been used in production for five months. During these first months of MT implementation in production, CILS customers (that is, company employees) could request PE as a service via the CILS internal purchase platform. In this phase, the CILS had implemented MT in some specific projects only, which were also entirely managed by in-house resources. It is also worth noting that PE assignments were most often left unrevised, since MT had been introduced primarily to speed up the workflow. Nonetheless, the MTPE workflow was intended to provide exactly the same final quality as a traditional workflow

involving human translation plus revision. The only difference was the time allocated to the PE which, following the productivity test detailed in Chapter 5, was set at 1.5 of the time usually allocated to revision.

The MT engine was fully integrated through an API into the CILS' usual CAT tool interface. MT suggestions appeared whenever no TM matches were available, or if such matches were below an 85% similarity threshold. Most often, linguists were presented with a fully pre-translated text containing one suggestion per sentence, coming either from the TM or from the NMT engine.

### 6.4.2 Corpus design

This section details the process set up to collect the data for our corpora of edits in revision (henceforth 'REV') and in PE. These corpora can be classified as both parallel and comparable: their internal structure is the one of a parallel corpus (Lefer, 2020), as they include original texts and their translations in (one or) two languages. Two translated versions are available for each source text: first, the machine or the human-translated version (henceforth a 'pre-translation') in the PE or the REV corpus, respectively. Second, the final, published version, which has been post-edited (in the case of NMT pre-translations) or revised (in the case of human pre-translations). The PE and the REV corpora are designed to be comparable with each other [14]. Comparability has been ensured using various criteria, as it will become clear in the next paragraphs. Finally, the corpora are monodirectional, meaning that only one translation direction (from one source language into two target languages) is represented (Lefer, 2020).

The corpora were constructed with opportunistic principles (McEnery and Hardie, 2012, p. 11), meaning that they "represent nothing more nor less than the data that it was possible to gather for a specific task". Nonetheless, in designing our corpora we have adhered to a rather rigorous sampling frame – which, according to McEnery and Hardie (2012, pp. 11–13), is a missing aspect in opportunistic corpora. The two corpora cover the same time-span and two language combinations, namely German into French and German into Italian. After signing a data protection and confidentiality agreement with the CILS management [15], we have been granted access to the CILS translation management platform and to the list of all PE and REV assignments completed between May and July 2020 for the two language pairs considered. The REV list contained 2,735 assignments, while the PE list for the same time-span included only 133 PE assignments. The considerable difference in the number of assignments shows that MT is not applied on all text types, but rather on a tiny percentage of assignments.

14. McEnery and Hardie (2012, p. 20) define a comparable corpus as "a corpus containing components that are collected using the same sampling method, e.g. the same proportions of the texts of the same genres in the same domains in a range of different languages in the same sampling period".

15. Obtaining informed consent from the linguists was deemed unnecessary in this context, as the intellectual property of product data under examination belonged to the company, rather than to the individuals. Additionally, it is essential to note that all data related to the linguists who contributed to the texts has been anonymized.

To reduce the huge disparity in corpus size and to lower data collection effort (since the download of files could not be automatized), we took the following decisions:

— For PE, we added assignments completed in April and in August 2020 (total: 157 PE assignments); this led to a total of 290 PE assignments;

— For REV, we chose to focus on jobs completed during the first five working days of each month, on a three-months' time-span (May-June-July). We preferred to cover whole weeks rather than to randomly choose three days per month, in order to account for linguists' fixed days off (as not all salaried translators work five days per week, but most often have one or even two days off). This process let us reduce the initial dataset to 665 REV assignments.

From these extended lists, we inspected the corresponding PE and REV assignments on the CILS platform and manually downloaded those who complied with the following requirements:

1. *Assignments were not particularly confidential.* Several assignments were marked as "strictly confidential" on the translation management platform and could not be included in our corpus.
2. *Assignments were carried out by in-house staff.* We chose to retain only texts that had been post-edited or revised by in-house linguists. While the vast majority of revision assignments were carried out in house, a small percentage of texts were revised by freelance translators via external partner agencies. In this case, however, it was impossible to retrieve the name of the linguist who had revised the text. Furthermore, in the first phase of MT implementation in the production workflow, partner agencies had no access to the custom NMT engine and PE assignments were exclusively managed by in-house linguists. Hence, retaining only texts revised (or post-edited) by in-house staff allowed us to focus on a more comparable pool of linguists and to exclude a potential confounding variable.
3. *For PE assignments, MT segments accounted for at least 50% of the text.* Several PE assignments included only one or two segments coming from the NMT engine, while the rest of the text was translated using TM matches. This was the case, for instance, of documentation updates. Given that we were only interested in the segments coming from the MT, these assignments would have provided us with a rather small quantity of usable data, and – what is more problematic – with very little context to evaluate the modifications. Therefore, we decided to retain only PE assignments in which at least 50% of the text (source words) had been translated with the NMT engine. This choice let us collect a fair quantity of MT sentences in context, which yields more useful insights into the work of linguists with an NMT engine [16].

16. As Läubli, Sennrich, and Volk (2018, p. 4793) pointed out, document-level MT evaluation "exposes translation errors related to discourse phenomena which remain invisible in a sentence-level evaluation". We argue that this has direct implications for linguists, who are called to ensure coherence and cohesion at textual level.

| | PE | | | | REV | | | |
|---|---|---|---|---|---|---|---|---|
| | source segments | MT-segments only | MT modified during PE | % modified on tot. MT segments | source segments | seg. modified during REV | % modified on tot. REV segments | human-only segments modified in REV |
| *de-fr* | 4,183 | 2,774 | 2,202 | 79.38 | 46,629 | 5,166 | 11.08 | 2,939 |
| *de-it* | 3,704 | 2,494 | 1,993 | 79.91 | 41,309 | 3,729 | 9.03 | 2,030 |
| **TOT** | **7,887** | **5,268** | **4,195** | **79.63** | **87,938** | **8,895** | **10.12** | **4,969** |

TABLE 6.1 – Triples in corpora, before and after sample selection, per task and language combination.

Once all the non-compliant assignments had been discarded, we obtained a final selection of 169 PE and 278 REV assignments. This reduction in the number of usable assignments has different reasons, depending on the workflow considered: for PE assignments, the majority of source texts were translated using TM, rather than the MT engine; while among REV assignments, we encountered a huge number of strictly confidential texts, which drastically lowered the number of available assignments.

As assignments can contain more than one text to post-edit/revise, the PE corpus consisted of 174 source texts [17] — with 140 overlapping source texts, i.e., machine-translated and post-edited both in French and in Italian – while the REV corpus consisted of 338 source texts [18], with 257 texts overlapping between target languages [19].

From the translation management platform, we manually downloaded a .zip file for each assignment, consisting of the source text, the final (target) text, pre- and post-task analysis files containing assignment metadata (such as percentage of TM and MT segments, word count, name of the translator/post-editor/reviser), and a table detailing all the modifications made per segment.

Since our aim was to compare the edits made to MT and human-translated segments, we extracted from each assignment only the segments that had been modified during the PE or the REV task. We collected in a spreadsheet file three versions of each segment, namely the source segment, its pre-translated version (either machine-translated or human-translated, for the PE and the REV corpus, respectively), and its final post-edited or revised version. We will refer to the three different versions of the same segment as *triple*. Table 6.1 describes the corpus size at various stages of our data collection process.

The PE corpus consisted of 7,887 triples — or 4,183 and 3,704 triples, for the language pairs German into French and German into Italian, respectively. From this collection, we first excluded all segments coming from the translation memory (n=2,619), as we were only interested in those coming from MT. We were therefore left with 2,774 (*de-fr*) and 2,494

17. 162 for the language pair *de-fr*, and 152 for the language pair *de-it*.
18. 321 German source texts translated in French, and 274 German source texts translated into Italian.
19. However, it is worth noting that, even in overlapping assignments, source segments could not be the same for both languages as the TMs are different and some segments could have matches in the TM for one language, but not in the other.

(*de-it*) segments coming from the NMT engine. Then, we focused on machine-translated segments that underwent modifications during the PE task (n=2,202 for DR-FR and n=1,993 for *de-it*), that is 79.38% (*de-fr*) and 79.91% (*de-it*) of the MT segments in each corpus. This means that 20.62% (*de-fr*) and 20.09% (*de-it*) of all the usable machine-translated segments in the PE corpora have been validated without further modifications.

As for the REV corpus, the original collection consisted of 87,938 triples – 46,629 and 41,309 triples for the language pairs German into French and German into Italian, respectively. From this list, we extracted segments that were modified during the revision step (n=5,166 for de fr and n=3,729 for *de-it*), as automatically recorded by the CILS' CAT tool and stored in Excel tables downloaded from the translation management platform. Initially, we had decided to include segments coming from the TM, as they result from human translation and a rereading of these segments, too, represents an integral part of REV assignments [20]. However, after closer inspection of the data, we found that, due to the inclusion of TM matches, our corpus contained several duplicates, both among source and target (revised) segments.

Indeed, for source segments with matches in the TM, the majority of changes made by the reviser to the pre-translated output fell into one of these two scenarios: either the translator had selected a TM match and this was modified by the reviser, or a TM segment was originally discarded by the translator and then brought back by the reviser with a simple click to replace the translator's proposal. In both cases, modifications made by the reviser were automatically propagated over source segments duplicates throughout the text. Considering these issues, and in an effort to ensure maximum comparability between our corpora, we decided to remove TM matches from the REV corpus, too, and to focus on segments translated from scratch and then modified by the reviser in the revision step (in Table 6.1 these are referred to as *human-only segments*). Finally, we cleaned this collection removing segments containing only links to webpages, telephone numbers, addresses or people names, in order to focus on meaningful changes only. As we did for the PE corpus, we also merged some contiguous source segments when the translators had merged them in their translation. In the next section, we will describe the final selection of segments used in our analyses. We will refer to *corpus* or *corpora* to indicate our final selection of PE and REV triples, while we will use *subcorpus* to refer to the collection of triples from a specific language pair.

### 6.4.3 Corpus description

In this section, we summarise the characteristics of our PE and REV corpora, in terms of number of segments, tokens, text domains and linguists who carried out the tasks. Each final version of the triples included in these corpora contains at least one modification to its

20. In the TS literature, there is less consensus on whether editing TM matches in PE assignments should be considered "post-editing" or rather "revising" (see for instance the discussion in do Carmo and Moorkens, 2021). In the present study, we solve the terminological confusion making clear that we refer to PE only when MT segments come into play.

(human or machine) pre-translated version.

— **Assignments, texts, segments (triples)**

Table 6.2 describes the number of assignments, source texts and triples included in each corpus (*all*) and in individual target language subcorpora (*de-fr* and *de-it*).

| Task | | Assignments | Source texts | Triples |
|---|---|---|---|---|
| **PE** | *de-fr* | 156 | 160 | 2,150 |
| | *de-it* | 146 | 150 | 1,939 |
| | ***all*** | **169** | **174** | **4,089** |
| **REV** | *de-fr* | 227 | 251 | 2,920 |
| | *de-it* | 183 | 204 | 2,013 |
| | ***all*** | **248** | **283** | **4,933** |

TABLE 6.2 – Number of assignments, texts and triples in corpora, per task and language combination.

The final PE corpus contains 4,089 triples, namely 2,150 for *de-fr* and 1,939 for *de-it*, with 1,553 source segments overlapping between the two languages. The segments come from 169 different PE assignments and 174 individual source texts [21], from which we extracted a minimum of one to a maximum of 74 source segments (Median=10) from each text in the target language subcorpora. Since we were interested only in segments that underwent some modifications during the PE task, we could extract on average 58% [22] of segments from each individual text included in the corpus.

The REV corpus comprises 4,933 triples, i.e., 2,920 for *de-fr* and 2,013 for *de-it*, including 760 overlapping source segments, meaning that a translation is available in both target languages. In this corpus, segments originate from 248 assignments, corresponding to 283 individual source texts, from which we extracted a minimum of one to a maximum of 276 source segments (Median=4) for each target language. Since we were only interested in human-translated segments that underwent modifications during the revision phase, we could extract on average 19% of segments from each text [23]. The difference between the average percentages of original source segments included in each corpus can be attributed to the retention criteria imposed during data collection. While for the PE corpus we retained only texts in which at least 50% of the source words originated from MT, for REV assignments we did not impose a minimum percentage of words or sentences translated from scratch in each text. It is

21. Overlapping assignments between target language subcorpora: PE n=133, REV n=9. Overlapping texts: PE n=136, REV n=172. However, the reader should bear in mind that assignments can contain more than one text; therefore, an overlapping assignment between target languages could simply mean that different texts from the same assignment have been retained in each subcorpus. Similarly, overlapping texts do not mean that the same source segments have been extracted for the two subcorpora. Therefore, only overlapping segments (reported in the text) should be taken as a meaningful means of comparison.

22. Since assignments could contain more than one text, this figure correspond to 52% of the total number of segments in the selected PE assignments.

23. This corresponds to 6.68% of the total number of segments in the selected REV assignments.

also worth noting that source texts included in the PE corpus are in general shorter, in terms of number of segments (ca. 25 segments on average, Median=13), than the source texts included in the REV corpus (ca. 162 segments on average, Median=44). Nonetheless, the latter contained several TM matches. It is important to bear in mind that the final REV corpus contains only sentences translated from scratch by human translators, i.e., without the aid of TM suggestions, and then modified by revisers. That is why only a relatively small percentage (on average) of original source segments has been included in the REV corpus.

— **Tokens**

Table 6.3 details our PE and REV corpora (and corresponding subcorpora) in terms of number of tokens, punctuation included.

| **Task** | | **Source text** | **Pre-translated text** | **Target text** |
|---|---|---|---|---|
| PE | *de-fr* | 33,561 | 42,252 | 43,633 |
| | *de-it* | 30,527 | 34,990 | 36,904 |
| | ***all*** | **64,088** | **77,242** | **80,537** |
| REV | *de-fr* | 49,093 | 62,806 | 63,464 |
| | *de-it* | 33,192 | 38,448 | 39,233 |
| | ***all*** | **82,285** | **101,254** | **102,697** |

Table 6.3 – Number of tokens in corpora, per task and language combination.

The higher number of source segments in the REV corpus also corresponds to a higher number of source tokens (+18,197) compared to the PE corpus. This difference is even greater if we compare texts translated by the machine and by human translators (+24,012 tokens in the REV corpus), while it slightly decreases (+22,160 tokens) between target texts in the PE and the REV corpora [24].

The largest subcorpus corresponds to human-translated and revised texts in the language combination *de-fr*, followed by texts that have been machine-translated and post-edited in this same language combination. Both the PE and REV subcorpora for the language combination *de-it* are smaller in size than their counterparts, which is in line with the workflow of the CILS partner, where *de-fr* is the first language combination for number of annual assignments.

We now move on to describe our PE and REV corpora in terms of source segment length distribution (Figure 6.1). Source sentence length goes from a minimum of one token in both tasks to a maximum of 69 source tokens in PE, and 103 tokens in REV.

24. This would suggest that post-editors add more words to the pre-translated text compared to revisers. In this section, however, our goal is to simply describe our corpora and corresponding subcorpora. We will provide further interpretation of the data in Section 6.5.

Although for a fairer comparison of the corpora it would be better to have the same maximum length, we decided not to exclude sentences longer than 69 tokens in REV (n=5), as these cases were interesting to analyse from a qualitative perspective and also because their exclusion would have not significantly affected the total average sentence length of the corpus. Nonetheless, in order to visualize the data more effectively, in Figure 6.1 we have excluded extreme data points that would have distorted the representation. From a closer inspection of longer source segments, we found that 51 segments (20 in PE and 31 in REV) are actually made up of two or many merged sentences.

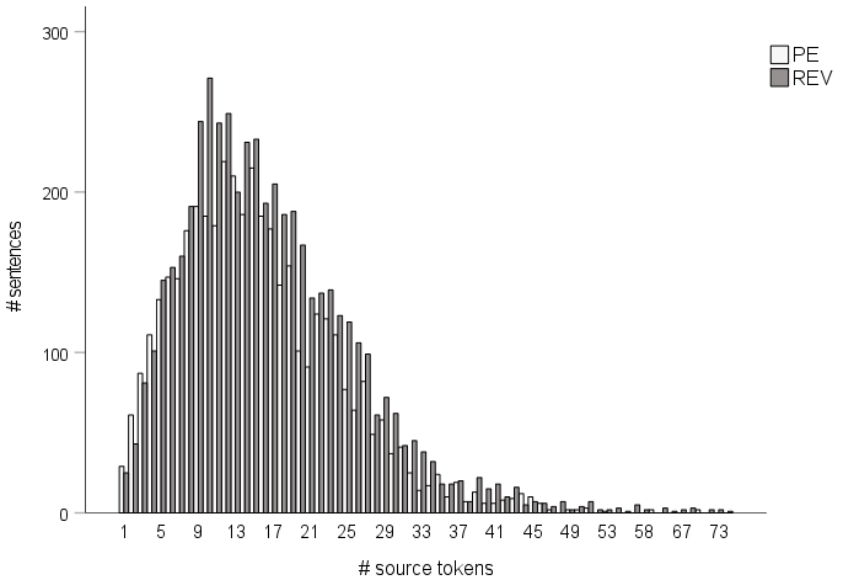



FIGURE 6.1 – Source segment length distribution, per task.

Regarding the minimum source sentence length, it could be argued that segments of just one or two tokens could not be particularly representative of the editing work, as most of the times these represent wrongly-translated words or terms in a list that linguists simply need to replace. However, these short German segments usually correspond to well-formed phrases in French and in Italian. We found that these phrases are not always wrong, and their modification is often driven by contextual needs, as in (1):

(1) **Source (de)** *Angebotsarten*
'Types of offer'

**Pre-translation (it)** *Tipi di offerte*
'Kinds of offers'

**Final (it)** *Tipo di offerta*
'Kind of offer'

Therefore, we decided to also retain source segments of one token (n=54).

— **Domains**

CILS' customers — the ones requesting translations to the language service — are corporate employees working in various corporate departments. Translation assignments are recorded in the translation management platform using the name of the corporate department, since specific TMs and other translation resources such as glossaries are associated to specific departments. Translations requested from the same departments loosely cover the same topic or related topics, although text types can vary. Drawing on the information stored in the translation management platform, we describe our PE and REV corpora in terms of corporate departments that requested the translation. We will refer to "text domain" to indicate this information.

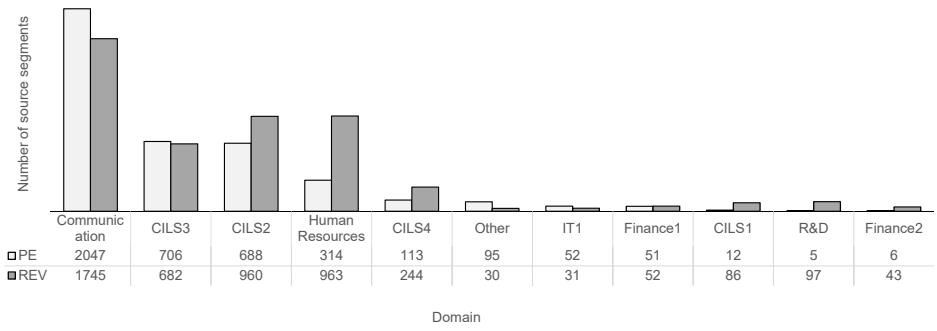



FIGURE 6.2 – Distribution of text domains in our corpora (number of source segments, per task).

Our PE and REV corpora contain texts extracted from 17 different domains. As shown in Figure 6.2, the most frequent domain for both tasks is Communication, followed by specific corporate domains (CILS3 and CILS2) whose name has been anonymised. The Other category includes seven domains for which we could not collect any segments in one of the two tasks. In total, this category contains 95 PE and 30 REV segments. As noted by Loock (2018), text domain can greatly influence the results of MT systems, depending on the data used to train the engine. We recall that the NMT engine in use at the CILS partner had been trained using all the available CILS' TMs, and therefore on all the text domains represented in Figure 6.2.

Since the number of segments do not necessarily correlate with the number of tokens, in Figure 6.3 we show the percentage of corpus source tokens included in each domain, per task.

Texts coming from the Communication domain represent almost 51% of source tokens in the PE corpus and 35% of the REV corpus. This is by far the largest domain in terms of source words in both corpora. Slight differences between corpora are shown for the other domains, but percentages are significantly lower compared to the Communication domain. For instance, the second largest domain is CILS3 in the PE corpus (17.11%) and Human Resources in the REV corpus (18.96%). It is also worth mentioning that segments in the domain Finance2, as well as the various domains included in the Other category correspond to less than 1% source tokens on the total tokens in

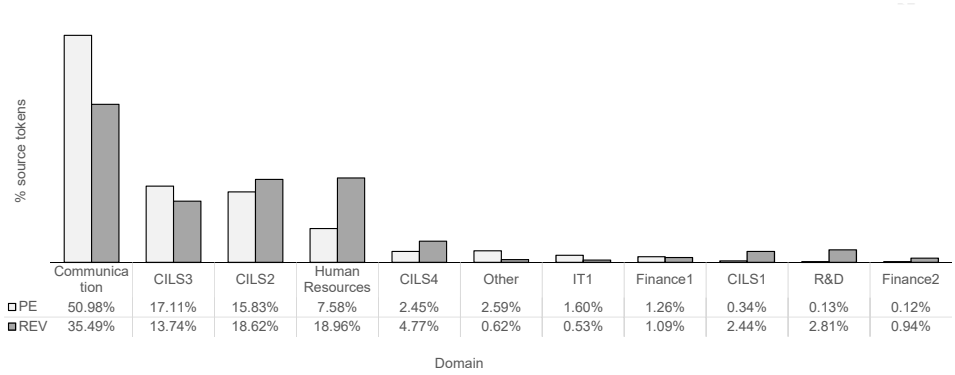



FIGURE 6.3 – Distribution of text domains in our corpora (percentage of source tokens on the total, per task).

each corpus. The reader should note that the description of source domains is given only to evaluate the comparability of our corpora, and for the sake of transparency and completeness of information. Since we were not interested in the impact of different domains on the quantity and type of corrections performed on pre-translated texts, we have not controlled for this variable during data collection and corpus design. Therefore, the number of segments included in some domains is unbalanced between corpora. Nonetheless, the three domains that, taken together, account for 84% and 69% of the segments in the PE and REV corpora, respectively — namely Communication, CILS3 and CILS2 — are quite balanced between the corpora, both in terms of segments (+/- 40%) and in terms of source words (+/- 51%).

— **Linguists**

The segments in our corpora have been revised or post-edited by 23 linguists (L) in total, 12 for the language pair *de-fr* and 11 for the language pair *de-it*. All the linguists who carried out the PE and the revision tasks were employed in house, while the human translations came mainly from two external partner agencies, therefore, we do not know how many translators have translated the texts in the REV corpus. Figure 6.4 shows the number of source segments processed by in-house linguists in each task. Except for two revisers (L-FR013 and L-IT012) and one post-editor (L-IT009) for whom we do not have data from the opposite task, all the other linguists are represented in both corpora. However, the number of source segments processed in each task varies significantly for each linguist. Indeed, ten out of twenty linguists have modified more than the double of the segments in one of the two tasks compared to the opposite one. For instance, linguist L-FR010 post-edited 159 segments in our PE corpus, but the same linguist revised only 4 segments among those included in the REV corpus.

We also report the number of source tokens post-edited or revised by each linguist on the total of source tokens in each corpus (Figure 6.5).

As with the domains, linguists were not a considered a variable during data collection.

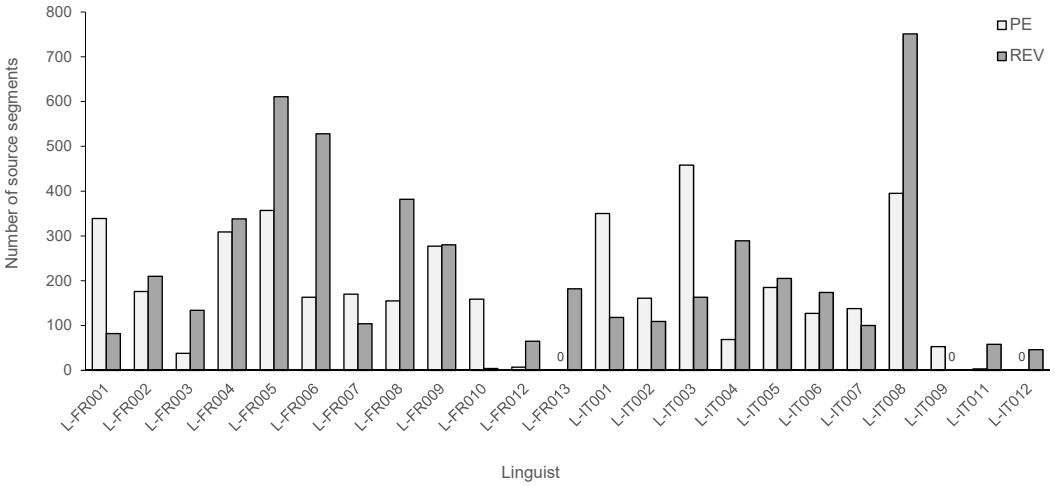



FIGURE 6.4 – Number of source segments handled by in-house linguists, per task.

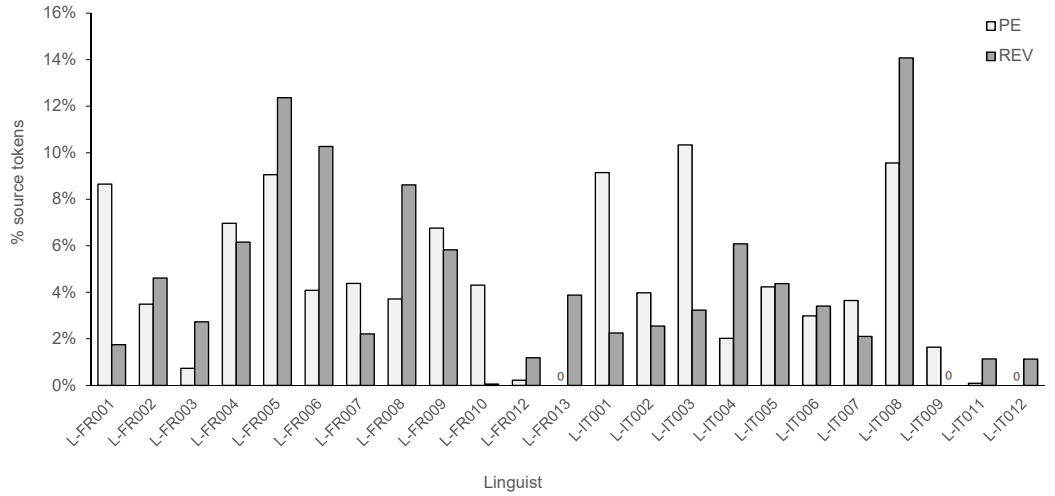



FIGURE 6.5 – Percentage of source tokens handled by individual linguists in each task.

Therefore, in our corpora we find that some of them have performed one task more than the other. Nonetheless, the corpora represent a homogeneous and varied sample of linguists who carried out the tasks. For eight linguists out of 23, the number of segments processed in PE and in REV is highly comparable (+/- 39%).

### 6.4.4 Data processing and analysis

In this section, we move forward to describing how the study data were processed and analysed. After completing the corpus design process (Section 6.4.2), corpora have been saved in .txt files, one file per task and per category in the workflow, namely source, pre-translation (automatic or human) and target text – we will refer to each one of these three collections as "subsets". It is worth mentioning that, as we had signed a confidentiality agreement with the CILS partner to obtain the data for this study, pre-processing of the data and subsequent analyses were carried out using standalone software, *ad hoc* scripts written in Python or Java and NLP pipeline libraries. We therefore avoided any Web-based tools that would not guarantee a secure transfer and processing of data. The files have been pre-processed (i.e., tokenised and POS-tagged) using spaCy (v3.4), a modern NLP library which offers several pre-trained statistical models for a variety of languages [25]. Custom tokenisation rules were applied to take into account special cases, such as abbreviations (e.g. *incl.* and *p. ex.*). Punctuation marks were counted as tokens. Following the pre-processing step, corpus data were collated in a spreadsheet tool and organized as a table in which each row contained one triple, i.e., a source segment, with its pre-translated and final (post-edited or revised) version, as recorded in the CILS' TM. The table rows also included the following information:

— *ID*: a general identifier assigned to the triple;

— *LANG*: the target language;

— *TASK*: either PE or REV;

— *TM ID*: an identifier containing the segment and job number, as recorded in the CILS TM;

— *LINGUIST*: the code assigned to each linguist who post-edited or revised the segment;

— *SOURCE, PRE-TRANS and TARGET TOKENS*: the number of tokens in the segment's source, pre-translated and final version;

— information on the original job (*CUSTOMER* for the domain, as shown in Section 6.4.3, *T-GROUP* for the target group, such as company employees or company customers, and *MEDIUM* for the dissemination channel, for instance intranet, website, leaflets etc.).

— *POS-tagged version* of the segments' pre-translated and final version;

— number of *POS modified* during PE and REV, per POS category;

25. `https://spacy.io/usage/models`. Accessed: 10 February 2023

— *HER score* per segment; and

— number of *editing actions* per type, as recorded by HER.

To facilitate the analysis of edits, we highlighted the differences between pre-translated and final segments using a custom tool. An example is displayed in Figure 6.6, where the verb *terminer* ('to end') was replaced during the revision task with the verb *quitter* ('to quit'):

| ID | LANG | TASK | TM ID | SOURCE | PRE-TRANSLATION | FINAL |
|---|---|---|---|---|---|---|
| 4668 | FR | REV | 20.04.1368_1_94 | Um diesen Modus wieder zu beenden, klickst du wieder auf das gleiche Symbol oder drückst die «ESC-Taste» auf deiner Tastatur. | Pour *terminer* ce mode, clique à nouveau sur le même symbole ou appuie sur la «touche ESC» de ton clavier. | Pour *quitter* ce mode, clique à nouveau sur le même symbole ou appuie sur la «touche ESC» de ton clavier. |

FIGURE 6.6 – Screenshot from the study Excel table showing the difference between a pre-translated segment and its final version.

Data have been analysed both quantitatively and qualitatively, combining methods used in MTPE research (see for instance Parra Escartín and Goulet, 2021 and Koponen, Salmi, and Nikulin, 2019) and in studies on the product of revision (Laflamme, 2009). The quantitative analysis consisted in calculating (i) the amount of editing performed, (ii) the distribution of editing actions and (iii) POS tokens affected by the edits. A more detailed analysis was performed on sentences falling in the HER category 1-25. The qualitative analysis has been carried out on a smaller portion of data from this same category. It includes (i) a linguistic evaluation of the edits performed, as well as an evaluation of their (ii) necessity and (iii) effectiveness.

Statistical analyses were performed in the IBM SPSS software (v. 28.0.1.1) and in R Statistical Software (v. 4.3.0). Descriptive statistics were employed to provide an overview of the findings. To assess the significance of differences between the PE and the REV corpora, inferential methods such as correlations, t-tests, Chi-square tests of independence and Mann-Whitney U tests were employed. For a technical explanation of these methods, we refer the reader to Mellinger et al. (2017).

### Quantitative analyses

The quantitative analyses performed on our corpora involved examining the amount of editing performed, the distribution of editing actions and the POS affected by edits during PE and REV. HER scores were calculated on the PE and the REV subcorpora to obtain an estimate of the edits quantity and edit types performed during PE and REV. Starting from the TERcom script (v. 0.7.25), we inverted the reference and hypothesis used to compute TER [26]. The input used for the TERcom script [27] were tokenised sentences obtained during the pre-processing step, meaning that we skipped the normalisation step provided by the

26. We used the following command to launch the script: `java -jar tercom.7.25.jar -s -r pretranslated.txt.sgm -h final.txt.sgm -n HER`

27. `https://www.cs.umd.edu/~snover/tercom/`. Accessed: 10 February 2023.

TER script [28]. Capitalization was taken into account as a modification. Edit types were automatically calculated by the HER script on each edited word, e.g. three insertions correspond to three words inserted.

The tokenised version of the corpus was also used to calculate the average sentence length of the main corpora, subcorpora and corresponding subsets. We indicate mean values, median, standard deviation and coefficient of variation. The latter indicator is a ratio of the standard deviation to the mean that allows for a more straightforward comparison of samples with different means. POS categories were automatically assigned to tokens in the Italian and in the French pre-translated subcorpora. Drawing on spaCy (v. 3.4) language-specific statistical models, we employed the most accurate pipelines [29] to parse our subcorpora, namely `fr_core_news_lg` for French and `it_core_news_lg` for Italian. According to spaCy's internal evaluations [30], POS taggers used in these pipelines provide an accuracy level of 97% and 98% for French and Italian, respectively. Nonetheless, when highlighting textual differences between POS-tagged subsets to obtain modified tokens with their POS, we found that in ca. 11% of the sentences (970 out of 9,022 sentences) some of the tokens were tagged with two different POS categories between their pre-translated and final versions.

To avoid extracting false POS-tagged tokens, we decided to manually inspect and correct sentences with a POS category mismatch. Finally, we linked modified, POS-tagged tokens with the editing action performed on them. However, extracting only highlighted POS tokens in pre-translated sentences formally excludes insertions, which are not visible in the pre-translated subsets. Indeed, while deletions, replacements and shifts are computed on pre-translated texts, insertions are calculated on the target texts. Simply counting insertions in the target texts, however, does not provide an accurate picture of the words inserted, as HER computes insertions by observing the number of tokens in target texts that exceed the counting of tokens in pre-translated texts. This includes cases of substitutions 1:n, as in *lors*→*à l'occasion*, which HER identifies as substitution of one token and insertion of two tokens, while we should rather consider it as the simple substitution of a preposition (*lors*) with a prepositional phrase (*à l'occasion*). Since we were interested in determining which POS tokens are most often inserted during the PE and the REV step, we decided to inspect sentences in the HER 1–25 category to manually identify *true insertions*, i.e., insertions of tokens that do not match a substituted element and therefore have truly no correspondence in pre-translated texts. In this particular case, the choice to only correct sentences in the HER 1–25 category is also motivated by the fact that POS affected by edits are most meaningful when the modification can be isolated and easily identified, rather than

28. We did so to identify the modifications made to specific parts of speech in a subsequent step of the analysis.

29. These models are reported to prioritize accuracy over efficiency. See `https://spacy.io/models`. Accessed: 13 March 2023.

30. `https://spacy.io/models/fr#fr_core_news_lg`. Accessed: 13 March 2023.

| | | | | | |
|---|---|---|---|---|---|
| **ADJ** | adjective | **INTJ** | interjection | **PUNCT** | punctuation |
| **ADP** | adposition | **NOUN** | noun | **SCONJ** | subordinating conjunction |
| **ADV** | adverb | **NUM** | numeral | **SYM** | symbol |
| **AUX** | auxiliary | **PART** | particle | **VERB** | verb |
| **CCONJ** | coordinating conjunction | **PRON** | pronoun | **X** | other |
| **DET** | determiner | **PROPN** | proper noun | | |

Table 6.4 – Universal POS tags used by spaCy.

in partially of fully retranslated sentences where entire structures are substituted. For the categorisation of POS tokens, spaCy makes use of the Universal POS tagset [31] (Table 6.4) [32].

### Qualitative analyses

The qualitative analyses involve a tripartite annotation of the edits made during PE and REV, consisting of a linguistic categorisation of the edits and an evaluation of their necessity and effectiveness. These three annotation processes have been carried out on the same subset of data. It is worth mentioning that the evaluation of necessity and effectiveness of the edits has been carried out with external annotators and, in part, with some CILS linguists. Therefore, a separate section will be devoted to the description of the methodology used for this evaluation. In this section, we first describe the subset used for the qualitative analyses, then we present the taxonomy used for the linguistic categorisation of edits.

The qualitative analyses have been performed on a subset of 300 sentences for each task, evenly distributed between the two language combinations considered [33]. In order to exclude potential confounding variables, we had to impose some restrictions on the choice of segments to be included in our test subset. First, we decided to keep only texts coming from the *Communication* domain, which is the domain with the largest number of segments and source tokens in both tasks, as shown in Section 6.4.3. Second, since we were interested in evaluating specific modifications rather than cases of retranslation, we focused on segments containing a limited and well-identifiable number of edits. Therefore, we decided to evaluate only segments that fall into the 1-25 HER category. Third, to ensure comparability of the segments and also to avoid over representation of certain linguists, we tried to maintain some variety in terms of the linguists who post-edited or revised the segments included in the test set. This was done by taking random subsamples of segments from each linguist in the PE dataset (the smallest dataset) and then randomly selecting a similar number of segments in the REV dataset. However, since some linguists had done more PE or REV than the opposite task, and some had not done PE or REV at all (cf. Section 6.4.3), it

31. `https://universaldependencies.org/u/pos/`. Accessed: 13 March 2023.
32. Adpositions include prepositions and postposition. However, since postpositions are not used in the target languages considered in our study, we will only refer to prepositions.
33. Further information on this subset is available in Section 6.4.5.

was not possible to include exactly the same number of segments from each linguist in both tasks.

It is also worth noting that, as the subset was intended to be evaluated for the necessity of edits by four CILS linguists (two per target language), we excluded segments that had been previously post-edited or revised by these linguists. Although we could not include entire texts in the test set, we tried to prioritize contiguous sentences, in order to provide some context for the evaluation of edits. Finally, although we chose the same number of segments for each task, a different number of modifications per sentence can be counted. The structure of the test set used for the qualitative analyses is detailed in the Appendix section (Appendix H).

The purpose of the linguistic annotation is mainly descriptive and only partially interpretative: we do not discuss what post-editors and revisers *should have corrected* (i.e., we do not annotate all raw MT or translator errors), but we rather examine what linguists corrected or modified in pre-translated texts and why. Popović (2018, p. 141) has pointed out that "post-editing can be viewed as implicit error annotation, since each edit operation is actually a correction of a translation error". We only partly agree with this claim. Both in PE and in REV, linguists are called not only to correct errors, but also to make other modifications that improve the final text, or that are required by in-house guidelines and customer needs. Therefore, the notion of translation error is quite restrictive in this context. Since we do not annotate errors in pre-translated texts in the first place, the present annotation can also be seen as a study of the issues that CILS linguists most often detect in raw MT and in human translation. This implies that other errors could have been left unnoticed.

#### Defining an edit annotation taxonomy

In order to find a taxonomy that would fit our purposes, we reviewed various existing taxonomies developed for the annotation of errors and edits in human or machine-translated texts (Vilar et al., 2006; Laflamme, 2009; Farrús et al., 2010; de Almeida, 2013; Daems, Macken, and Vandepitte, 2013; Costa et al., 2015; Lommel, 2018; Tezcan, Hoste, and Macken, 2018; Mossop, 2020; Parra Escartín and Goulet, 2021; Granger and Lefer, 2021). While the majority of taxonomies developed for MT engines consider MT as a final product and lie on the fundamental distinction between accuracy and fluency (see for instance Lommel, 2018; Tezcan, Hoste, and Macken, 2018), we sought a linguistically motivated taxonomy applicable to both machine translated and human translated texts. A taxonomy of this kind allows us to focus on the edits made during PE and REV, which do not necessarily correspond to errors in pre-translated texts.

We devised our taxonomy drawing upon three main sources, since none of them could be used as such in our study. Our first source is the linguistic typology developed by Laflamme (2009) to categorise lexical edits in revision. This taxonomy has been enriched by Parra Escartín and Goulet (2021) and successfully reused by these authors in their study to categorise edits performed during PE and proofreading. Laflamme's taxonomy is

however tailored for lexical modifications and many of the subcategories [34] considered are shared among different linguistic dimensions [35]. The version proposed by Parra Escartín and Goulet (2021) is too coarse-grained and does not provide enough information on the type of edit performed.

Our second source is the linguistic taxonomy devised by Macken et al. (2022) to categorise edits made during PE and subsequent revision. This taxonomy combines some linguistic dimensions that are sometimes difficult to distinguish (such as lexis and semantics) and includes several subcategories to describe the type of change performed. However, as already mentioned in Section 6.2.3, these subcategories appear inadequate in capturing the linguistic elements that have been altered in pre-translated texts, given that the majority of modifications fall in residual subcategories. Furthermore, the taxonomy is only briefly presented in the original paper, and the annotation guidelines are not made available; therefore, there are no examples of edits included in each subcategory.

Finally, our third source is the TAS taxonomy (Granger and Lefer, 2021), developed in the framework of the Multilingual Student Translation project (Granger and Lefer, 2020) to categorise human translation errors, but also applicable for annotating MT errors (Lefer, Piette, and Bodart, 2022). The TAS taxonomy includes multiple macro-dimensions and finer-grained categories and provides users with a detailed annotation manual, but it was originally designed for translation and with a double pedagogical purpose in mind: to describe typical translation student errors – such as "Word category confusion" – and to mark particularly felicitous translations. Therefore, not all the subcategories apply to our classification of professional edits in post-edited or revised texts.

Building on the three taxonomies presented earlier, we put forward a proposal to classify edits — and their corresponding editing actions — into one of the following five linguistic dimensions: (i) Lexis and semantics; (ii) Syntax and morphology; (iii) Discourse and pragmatics; (iv) Style; and (v) Mechanics. A sixth category ('Other') was included to account for cases where the linguistic dimension is not clear. The taxonomy is shown in Table 6.5 and a detailed explanation of the subcategories is provided in what follows.

Edits falling in the **Lexis and semantics (LEXSEM)** category remedy a mistranslation of the source text or a misselection of lexical content in the target language. It includes the following subcategories:

— **Content distortion/lexical misselection**
  All kinds of modifications performed on words or word strings to amend a lexical misselection or a distortion of the source content in the pre-translated text, i.e., incomprehensible, incorrect, inexact, illogical or potentially ambiguous meaning.

34. Laflamme (2009) does not use the word 'subcategories' but refers to 'issues detected' [*problèmes détectés*]. Nonetheless, these are linked to several linguistic dimensions (*ibid.*, p. 228).

35. For instance, the category "Unknown issue" can refer to the following linguistic dimensions: syntax, semantics, typography and 'undetermined'.

| LEXIS AND SEMANTICS (LEXSEM) | STYLE (STYLE) |
|---|---|
| Addition removed | Degree of informality/register |
| CILS terminology | Explicitation |
| Collocations and idioms | Implicitation |
| Content distortion/lexical misselection | Redundancy avoidance |
| Omission amended | Rephrasing/lexical choice |
| Translated untranslatable | Style guide change |
| Untranslated translatable | |
| **SYNTAX AND MORPHOLOGY (SYNMORPH)** | **MECHANICS (MECH)** |
| Conjunction | Extra/unbreakable space |
| Determiner, preposition, pronoun | Spelling |
| Inflectional morphology and agreement | Punctuation |
| **DISCOURSE AND PRAGMATICS (DISCPRAG)** | **OTHER** |
| Cohesion | Other |
| Constituent reordering | |
| Sentence merging | |
| Sentence splitting | |

TABLE 6.5 – Taxonomy used for the annotation of edits in pre-translated texts.

(2) **Source (de)** *Ich betreibe ein eigenes Kosmetikstudio*
'I run my own beauty salon'
**Pre-translation (it)** *Gestisco un salone di bellezza personale*
'I run a personal beauty salon'
**Final (it)** *Gestisco in proprio un salone di bellezza*
'I run my own beauty salon'

(3) **Source (de)** *...auch für kleinere Unternehmen*
'also for smaller businesses'
**Pre-translation (fr)** *...également pour les PME*
'for SMEs too'
**Final (fr)** *...également pour les entreprises plus petites*
'also for smaller businesses'

(4) **Source (de)** *der Corona-Pandemie*
'the Coronavirus pandemic'
**Pre-translation (it)** *la pandemia Corona*
'the Corona pandemic'
**Final (it)** *la pandemia di coronavirus*
'the Coronavirus pandemic'

— **Omission amended**
Insertion of words or word strings that appear in the source text but were omitted in the pre-translation. This subcategory also includes cases of partial omissions, which involve a substitution of lexical content (as in example 6).

(5) **Source (de)**
*Bei Fragen oder Unklarheiten könnt ihr jederzeit auf eure/n Vorgesetzte/n zugehen.*
'If you have any questions or doubts, you can contact your supervisor at any time.'
**Pre-translation (fr)**
*...vous pouvez vous adresser à votre supérieur(e).*
'...you can contact your manager'
**Final (fr)**
*...vous pouvez à tout moment vous adresser à votre supérieur(e).*
'...you can contact your supervisor at any time.'

(6) **Source (de)** *in einem aussergewöhnlich herausfordernden Jahr.*
'in an extremely challenging year.'
**Pre-translation (it)** *in un anno straordinario.*
'in a remarkable year.'
**Final (it)** *in un anno straordinariamente impegnativo*
'in an extremely challenging year.'

— **Addition removed**
Deletion of words or word strings that appear in the pre-translation but not in the source text. This category includes the deletion of repeated elements deriving from (MT) mistranslations or oversights.

(7) **Pre-translation (fr)** *une seule seule opportunité*
'a single single opportunity'
**Final (fr)** *une seule opportunité*
'a single opportunity'

— **Untranslated translatable**
Substitution of source content left untranslated in the pre-translation.

(8) **Pre-translation (it)** *Fase 1 und fase 2*
'Phase 1 und phase 2'
**Final (it)** *Fase 1 e fase 2*
'Phase 1 and phase 2'

— **Translated untranslatable**
Substitution of untranslatable translated content (e.g. a proper noun that should not be translated).

(9) **Source (de)** *Storyline*
'Storyline'
**Pre-translation (it)** *Scénario*
'Storyline'
**Final (fr)** *Storyline*
'Storyline'

— **Collocations and idioms**
Substitution of words or word strings with collocations or idiomatic expressions.

(10) **Source (de)** *Das ist nicht mein Bier*
'That's not my business'
**Pre-translation (it)** *Non è la mia birra*
'That's not my beer'
**Final (it)** *Non sono affari miei*
'That's not my business'

(11) **Pre-translation (it)** *è capitata al momento giusto*
'it happened at the right moment'
**Final (it)** *casca a pennello*
'it hits the nail on the head'

— **CILS terminology**
Substitution of words or word strings with their equivalent from the CILS' terminology database.

The **Syntax and morphology (SYNMORPH)** category includes edits related to sentence grammar, word grammar and inflectional morphology.

— **Inflectional morphology and agreement**
Substitution of words or word strings with alternative inflected forms of the same words, including different conjugated forms of the same verb.

(12) **Pre-translation (it)** **i clienti è tornato*
*'customers is back'
**Final (it)** *i clienti sono tornati*
'customers are back'

(13) **Pre-translation (fr)** *il a*
'he has'
**Final (fr)** *il avait*
'he had'

— **Determiner, preposition, pronoun**
Substitution or deletion of incorrect determiners, prepositions or pronouns or insertion of such elements where their absence introduces issues in the sentence structure. This category does not include the substitution of articles with demonstratives (see DISCRPRAG – Cohesion).

(14) **Pre-translation (fr)** **à les* clients
*'to customers'
**Final (fr)** *aux clients*
'to customers'

(15) **Pre-translation (it)** **quel ristorante vicino mare*
*'that restaurant near sea'
**Final (it)** *quel ristorante vicino al mare*
'that restaurant by the sea'

— **Conjunction**
Insertion or substitution of conjunctions operating at segment level[36], including the substitution of a conjunction with punctuation marks, such as commas and colons,

36. Conjunctions are local cohesion devices, i.e., they operate within sentences, while the category DISCPRAG – Cohesion refers to connectives between sentences.

and vice versa. N.B. This category does not include cases where a full stop has been substituted with a conjunction (or vice versa) in order to merge or to split two contiguous segments (see DISCPRAG – Sentence merging/splitting).

(16) **Pre-translation (it)** *cioè per cinque anni*
'i.e., for five years'
**Final (it)** *ovvero per cinque anni*
'that is, for five years'

(17) **Pre-translation (it)** *Non ho pranzato, non mi va*
'I haven't had lunch, I don't feel like it'
**Final (it)** *Non ho pranzato e non mi va*
'I haven't had lunch and I don't feel like it'

The **Discourse and pragmatics (DISCPRAG)** category groups together edits related to textual cohesion and information structuring at textual level, such as sentence merging and sentence splitting, or at segment level, such as constituent reordering.

— **Cohesion**
Insertion of linkwords and pronominal references which are not present in the source segment [37], in order to improve cohesion between segments. This subcategory also includes substitution of articles with demonstratives, the substitution of cohesive markers operating beyond the sentence level, as well as substitution of (function or content) words and word strings with information retrieved from previous segments.

(18) **Source (de)** *Der Projekt*
'The/This project'
**Pre-translation (fr)** *Le projet*
'The project'
**Final (fr)** *Ce projet*
'This project'

(19) **Source (de)** *Er hat sich noch nicht entschieden.*
'He has not yet decided.'
**Pre-translation (it)** *Non ha ancora preso una decisione.*
'He has not yet decided.'
**Final (it)** *Tuttavia, non ha ancora preso una decisione.*
'However, he has not made a decision yet.'

37. Elements present in the source text and absent in the pre-translation fall into the LEXSEM-OMISSION subcategory.

(20) **Source (de)**
*...hat entschieden, vorläufig keine Velos zu transportieren.*
'...has decided not to carry any bicycles for the time being.'
**Pre-translation (fr)**
*...a décidé de ne pas transporter de vélos pour l'instant.*
'...has decided not to carry any bicycles for the time being.'
**Final (fr)**
*...a décidé de ne pas les transporter pour l'instant.*
'...has decided not to carry them for the time being.'

— **Constituent reordering**
Shifts of words or word strings in a different position in the sentence.

(21) **Pre-translation (fr)** *Tous les cours sont enregistrés.*
'All courses are recorded. '
**Final (fr)** *Les cours sont tous enregistrés*
'Courses are all recorded.'

— **Sentence merging/splitting**
Insertion, deletion or substitution of punctuation marks or conjunctions to merge or to split contiguous segments.

(22) **Pre-translation (it)**
*Il progetto dura due anni. La fase pilota comincia il 15 marzo.*
'The project will last two years. The pilot phase begins on March 15th.'
**Final (it)**
*Il progetto dura due anni, la fase pilota comincia il 15 marzo.*
'The project will last two years, the pilot phase begins on March 15th.'

The **Style (STYLE)** category refers to edits made to comply with the CILS in-house style guide, to avoid redundant information, as well as to change the register and degree of (in)formality. This category also includes edits that provide a synonym or rephrase longer strings in the pre-translation. Finally, it also encompasses edits made to explicitate information that was left implicit (or vice versa) in both the source text and the pre-translation.

— **Style guide change**
All kinds of modifications performed on words, word strings, punctuation marks, accents, whitespaces and other symbols to comply with CILS stylistic and typographic guidelines.

(23) **Pre-translation (it)** *La temperatura ideale è 30°.*
'The optimal temperature is 30°.'
**Final (it)** *La temperatura ideale è 30°C.*
'The optimal temperature is 30°C.'

(24) **Pre-translation (fr)** *L_´Autorité.*
'The Authority'
**Final (fr)** L'Autorité.
'The Authority'

— **Degree of (in)formality/register**
Substitution of words or words strings to modify the tone or the communicative function of the pre-translation. It also includes the substitution of personal pronouns or word strings to modify the address form (e.g. (fr) *tu<>vous*; (it) *tu<>voi, Lei*).

(25) **Pre-translation (fr)** *Je te conseille de...*
'I advise you to...'
**Final (fr)** *Je vous conseille de...*
'I advise you to...'

(26) **Pre-translation (it)** *Ci sono...*
'There are...'
**Final (it)** Vi sono...
'There are...'

— **Explicitation** [38]
Insertion of information that was left implicit in the pre-translation, although it could be retrieved from the context or from real-world knowledge (i.e., it does not constitute an omission of source text content). This subcategory includes cases where an expanded version is preferred over an abbreviation or an acronym.

(27) **Source (de)** *Bern ist die Hauptstadt.*
'Bern is the capital.'
**Pre-translation (fr)** *Berne est la capitale.*
'Bern is the capital.'
**Final (fr)** *Berne est la capitale de la Suisse.*
'Bern is the capital of Switzerland.'

(28) **Pre-translation (it)** *La FINMA*
'FINMA'
**Final (it)** L'Autorità federale di vigilanza sui mercati finanziari.
'The Swiss Financial Market Supervisory Authority'

38. Baker (1996, p. 176) defines explicitation as "the tendency to spell things out in translation, including, in its simplest form, the practice of adding background information".

— **Implicitation** [39]
Deletion or substitution of information that was made explicit in the pre-translation but can nonetheless be retrieved from the context or from real-world knowledge. This subcategory includes ellipses and cases were an abbreviation or an acronym is preferred over its expansion.

(29) **Pre-translation (it)** *L'Ente Nazionale per le attività Spaziali e Aeronautiche*
'National Aeronautics and Space Administration'
**Final (it)** La NASA
'NASA'

(30) **Source (de)** *Welche Schwerpunkte gilt es zu priorisieren?*
'What are the key priorities to address?'
**Pre-translation (fr)** *Quelles priorités doit-on fixer ?*
'What priorities should we set?'
**Final (fr)** *Quelles priorités fixer?*
'What priorities should be set?'

— **Redundancy avoidance**
Deletion or substitution of words or word strings that constitute repeated information in the pre-translated text.

(31) **Pre-translation (fr)** *...de la troisième question (un mot-clé par question)*
'...of the third question (one keyword per question)'
**Final (fr)** *...de la troisième question (un mot-clé par chacune)*
'...of the third question (one keyword each)'

— **Rephrasing/lexical choice**
Substitution of words with a synonym or substitution of word strings to rephrase pre-translated content, without bringing any major change in meaning.

(32) **Pre-translation (it)** *le mascherine verranno consegnate...*
'masks will be delivered...'
**Final (it)** *le mascherine verranno fornite...*
'masks will be provided...'

39. Vinay and Darbelnet (1995, p. 344) define implicitation as "[a] stylistic translation technique which consists of making what is explicit in the source language implicit in the target language, relying on the context or the situation for conveying the meaning".

(33) **Pre-translation (fr)** *Vous en êtes également responsables*
'You are also responsible' (of something)
**Final (fr)** *Vous en assumez aussi la responsabilité.*
'You also take responsibility' (for something)

Finally, the **Mechanics (MECH)** category covers edits related to punctuation, spelling and typography.

— **Spelling**
Substitution of misspelled words and typographical errors (including accents, diacritics, within-word hyphens and incorrect capitalisation) with their correct equivalent in target language. This subcategory includes cases that are not already specified in the CILS stylistic and typographic guidelines (see STYLE – Style guide change).

(34) **Pre-translation (fr)** **le bâtimnt*
*'the buildng'
**Final (fr)** *le bâtiment*
'the building'

(35) **Pre-translation (it)** *Ne riparliamo a Settembre*
'We'll talk about that in september'
**Final (it)** *Ne riparliamo a settembre*
'We'll talk about that in September'

— **Punctuation**
Insertion, substitution or deletion of punctuation marks. This subcategory does not include cases in which a full stop is replaced with a comma, colon or semi-colon, to merge two contiguous segments (see DISCPRAG – Sentence splitting/ merging).

(36) **Pre-translation (it)** *cliccare su "ON/OFF"*
'click on ON/OFF'
**Final (it)** *cliccare su «ON/OFF»*
"click on ON/OFF'

— **Extra/unbreakable space**
Insertion or deletion of an extra or unbreakable space.

(37) **Pre-translation (it)** *la prossima settimana_.*
'next week .'
**Final (it)** *la prossima settimana.*
'next week.'

The present taxonomy was initially applied to a test set different from the one used for the main evaluation, and then refined accordingly. The annotation was performed by the main researcher and then validated by two University colleagues with experience in the annotation of MT and human translation errors. The *Other* category was chosen whenever multiple linguistic dimensions were possible for the same edit, often for cases where two editing actions were performed – as in the following example where the edit involves an insertion and a shift:

(38) **Pre-translation (it)**
*...disponibili nel comunicato stampa allegato e qui nella posizione ufficiale.*
'...available in the attached press release and here in the official document.'
**Final (it)**
*...disponibili nel comunicato stampa allegato e nella posizione ufficiale disponibile qui.*
'...available in the attached press release and in the official position available here'

We carried out the annotation in an Excel file, organised as a table in which each row contained only one modification. Segments containing more than one modification were therefore duplicated in consecutive rows, as often as the number of modifications contained in the segment. The modified parts in the pre-translated and in the final version were highlighted in blue. Two dependent drop-down menus were used to select the appropriate linguistic dimension affected by the edit, among the six presented above, and the corresponding sub-category.

We manually checked the editing actions pre-determined by the HER metric using a technique already implemented by Mutal et al. (2019), which consists in grouping semantically-related changes in one editing action. In this way, edits that affect two or more contiguous words and are linguistically related (as in *vicino al→sul*) were counted as a single editing unit. As previously mentioned (Section 6.2.2), HER describes this type of edits as one substitution and one deletion, while they would logically represent a string substitution.

As the HER metric does, we also consider editing actions at token level, instead of character level. For instance, *quelque→quelques*, is considered a word substitution, not the simple insertion of the -s character. For the same reason, edits in the *Spelling* subcategory are interpreted as word substitutions. The same holds true for edits in the DISCPRAG dimension that imply sentence merging or splitting, i.e., instead of describing the change as the insertion or substitution of punctuation marks plus substitution of lowercase word with a capitalised word or the opposite, these kinds of edits are described as one substitution overall.

Finally, following the idea of PE change introduced in Blain et al. (2011), edits caused by other edits, such as the change of singular/plural or masculine/feminine of determiners and adjectives, were counted as a single editing action, as they would not have occurred without the modification of the noun.

### 6.4.5 Human evaluation: necessity and effectiveness of edits

We conducted a small-scale, blind evaluation of the necessity and effectiveness of edits in PE and in REV, with human annotators. Our goal was twofold: (1) to determine whether linguists exhibit a tendency to make more necessary or unnecessary changes when working with human-pre-translated or with machine-pre-translated texts, and (2) to investigate whether edits performed during PE and REV consistently improve the pre-translation or whether linguists introduce more errors in one of the two tasks.

In line with previous literature (cf. Section 6.2.5), our study adopts the following definition of edits' necessity: necessary edits are those that, if not implemented, leave the sentence grammatically incorrect, inaccurate in meaning, unidiomatic or non-compliant with in-house guidelines. Unlike previous studies that have typically considered the evaluation of edits in terms of necessity and *correctness*, our study introduces a novel perspective by incorporating the concept of *effectiveness*. We define edit effectiveness as the extent to which the implemented edit impact the quality of the pre-translated sentence (i.e. no impact, positive impact, negative impact).

#### Evaluation of edits' necessity

The evaluation of edits' necessity involved a total of six evaluators, three for each target language. Two evaluators were CILS linguists, referred to as "expert evaluators", who were familiar with the texts, the in house style guide and other guidelines associated with assignments for specific customers. The remaining four evaluators were translation instructors and experienced translators from the Faculty of Translation and Interpreting at the University of Geneva. These will be referred to as "external evaluators".

Upon receiving permission to share the test data from the CILS partner, each evaluator was assigned 300 triples for their respective target language, consisting of 150 PE segments and 150 REV segments. Additionally, to assess intra-rater reliability, two segments in each test set were repeated. It is worth noting that evaluators were unaware of the source of the data, i.e., whether the sentences were pre-translated by human translators or via the MT engine. The evaluation was conducted using an Excel file organised as a table, with each row containing one triple and only one modification between the pre-translation and the final version. Segments with multiple modifications were repeated in consecutive rows, as often as the number of modifications contained in the segment [40]. The modified parts in the pre-translated and final versions were highlighted in blue, enabling evaluators to focus on one edit at a time. A drop-down menu was provided to select the appropriate option for evaluating the necessity of the edits, namely whether the edit was *necessary* or *optional*. A *Comment* section allowed annotators to indicate cases of indecision or provide other relevant information. Detailed instructions were included in another sheet within the same Excel file, specifying that evaluators should focus on the *Pre-translation* column to evaluate the

40. Consequently, the evaluation file consisted of a total of 436 rows for the *de-fr* combination and 437 rows for *de-it*, which included the two additional segments inserted for consistency-checking purposes.

necessity of the edit — except for omissions, which were only visible in the *Final version* (Appendix I).

Prior to the main evaluation, the setting and instructions underwent a pilot test with two colleagues from the University of Geneva. In addition to verifying the feasibility of the task and the clarity of instructions, the pilot test also served to determine an average completion time, which was ultimately set at two and a half hours. Feedback from the pilot evaluators highlighted the challenge of assessing the necessity of edits without familiarity with in-house style guides. In response, for external evaluators, we pre-filled the evaluation of edits falling under the scope of in-house guidelines. This decision was based on the objective nature of assessing these edits, as we were not evaluating evaluators' ability to read and follow in-house guidelines. Expert evaluators, who were already familiar with the style guide, did not require this pre-filling. Pilot evaluators also noted the need for a broader context, such as the previous sentence, the paragraph or even the whole text, to evaluate some of the edits. Previous studies have shown that a larger context is paramount in evaluating MT correctness (see for instance Castilho, Popović, and Way, 2020; Castilho, 2020; Läubli, Sennrich, and Volk, 2018). When compiling the test set for the qualitative analyses, we prioritized contiguous sentences that had been modified in PE or in REV. However, conducting a systematic, document-level evaluation was not feasible for this study, as it would have significantly lengthened the task completion time. Most importantly, while the CILS had agreed to let us share the (small) test set with external evaluators, it would have been less inclined to share several entire texts.

To overcome this limitation, external evaluators were asked to report cases in which a broader context was needed to assess the necessity of the edit. In a subsequent step, external evaluators were provided with the segment in context, which included the three previous and following segments or the entire text, depending on the case. Expert evaluators did not need this additional step. Since the ID column contained the code pointing at the assignment number and segment number, as recorded in the CILS TM, CILS evaluators could refer to it to retrieve the original assignment from their database and resolve any doubts stemming from a lack of textual context [41]. Evaluators were also encouraged to contact the researcher for further information or if they had any doubts regarding the evaluation setting and instructions.

An inter-rater agreement was calculated with Light's kappa for multiple raters (Light, 1971). Italian-speaking raters report a kappa score of 0.386, which can be considered a fair agreement (Landis and Koch, 1977). Percentage agreement among Italian raters was 64% overall, with greater consistency for edits in the PE corpus (69%) than for edits in the REV corpus (58%). Similarly, a fair agreement, with $\kappa$=0.323, was found among French-speaking raters [42]. On this subcorpus, however, percentage agreement is overall lower (47%)

41. Although this was not explicitly requested in the instructions, one out of three CILS evaluators reported to have used this approach.

42. It is worth noting that the expert evaluator intentionally excluded the evaluation of one sentence and inadvertently overlooked the assessment of two additional sentences within the Excel file. Consequently, the

compared to the Italian subcorpus,and approximately the same for edits in the REV corpus (47%) compared to edits in the PE corpus (46%).

**Note on the evaluation of edits' necessity**

It is important to note that the evaluation of edits' necessity was initially intended to be conducted exclusively with CILS linguists to ensure maximal ecological validity. Given their expertise and familiarity with the specific text type, CILS linguists were considered the most appropriate evaluators for assessing the necessity of edits. Therefore, an evaluation was organized specifically for CILS linguists; however, it did not proceed as expected.

Following discussions with CILS management, it was decided to treat the evaluation as a regular assignment, providing linguists with a dedicated timeframe of two and a half hours to complete the assessment. Two in-house linguists per language combination were assigned to the task. The evaluation setup mirrored that of the assessment with external evaluators, with the addition of two additional columns, namely T-group and Medium (as explained in Section 6.4.4). This was done to provide CILS evaluators with as much contextual information on the original assignment as possible. The instructions provided to CILS evaluators differed slightly from those given to external evaluators. Specifically, they were asked to "apply the criteria they would usually apply in their daily assignments" to distinguish between necessary and optional edits. It was also specified that all the segments were taken from assignments in the Communication domain.

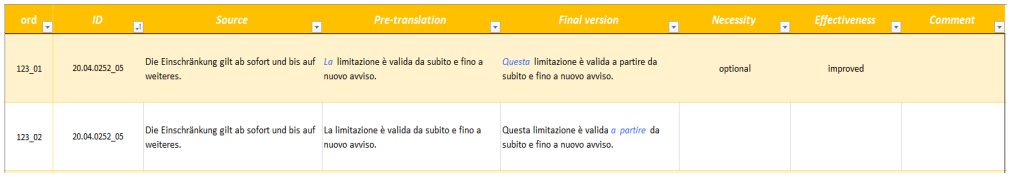
Despite having piloted the evaluation setup to estimate the completion time, only one out of four CILS evaluators managed to complete the assessment within the allocated time frame. The other evaluators were only able to assess half or two-thirds of the test set. Furthermore, when uncertain, only the evaluator who completed the assessment (one of the two French-speaking professionals) had looked up segments in their original assignments for clarification. Hence, it was decided to retain and analyse only the data from this evaluator and to merge them with those collected from the external evaluators for the French test set. However, for the Italian test set, none of the data from the two CILS evaluators could be reused. To address this, an additional expert evaluator, who was a former CILS linguist, was enlisted to conduct the evaluation under the same conditions as the external evaluators. Nonetheless, this evaluation was conducted using the Excel file specifically designed for CILS evaluators.

This evaluation setup allowed us to gather three judgments for assessing the necessity of edits in each target language subcorpus: two from external, professional translators and one from an internal expert CILS evaluator. This approach allowed us to retain a majority judgment on each edit.

computation of Light's kappa for the French test set encompasses 431 sentences instead of the original count of 434.

### Evaluation of edits' effectiveness

The evaluation of edits' effectiveness was conducted by external evaluators only [43], using the same Excel file as the evaluation of necessity (Figure 6.7). The column labelled Effectiveness contained a drop down menu offering three options: improved, neither improved nor degraded, and degraded. For this evaluation in particular, raters were instructed to focus on the final version of the segment in order to check whether or not the solution improved the pre-translation. Additionally, they were explicitly instructed to evaluate only the modification highlighted in blue in each row.



| ord | ID | Source | Pre-translation | Final version | Necessity | Effectiveness | Comment |
|---|---|---|---|---|---|---|---|
| 123_01 | 20.04.0252_05 | Die Einschränkung gilt ab sofort und bis auf weiteres. | La limitazione è valida da subito e fino a nuovo avviso. | Questa limitazione è valida a partire da subito e fino a nuovo avviso. | optional | improved | |
| 123_02 | 20.04.0252_05 | Die Einschränkung gilt ab sofort und bis auf weiteres. | La limitazione è valida da subito e fino a nuovo avviso. | Questa limitazione è valida a partire da subito e fino a nuovo avviso. | | | |

FIGURE 6.7 – Screenshot of the evaluation file for external evaluators.

For the evaluation of edits' effectiveness, inter-rater agreement was calculated between raters in each target language subcorpus using Cohen's kappa (Cohen, 1960). The agreement between raters was considered fair according to Landis and Koch (1977), with a kappa value of 0.346 for Italian-speaking raters and 0.257 for French-speaking raters. The percentage agreement was higher between Italian-speaking raters (82%) compared to French-speaking raters (70%). Overall, there was greater consistency in evaluating the efficiency of edits performed in PE assignments (87% for Italian and 70% for French) compared to REV assignments (76% for Italian and 69% for French).

In order to retain a judgement of effectiveness for each edit while considering the assessment from both raters, an "effectiveness score" was calculated. The *improved* option was assigned one point, *neither improved nor degraded* was assigned 0 points, and *degraded* was assigned a score of -1. The judgments provided by the two raters were summed to determine the score for each edit. The effectiveness score could range between -2, indicating that the edit degraded the pre-translation, and +2, indicating that the edit improved the pre-translation. A score of zero could indicate two different things: either a *truly preferential edit* (in the case of an edit that was originally deemed 'optional') or an *ineffective edit* (in the case of a necessary edit where linguists failed to improve the pre-translation, though without making it worse).

## 6.5 Results and discussion

In this section, we will detail the results from quantitative and qualitative analyses on our PE and REV corpora, following the order of presentation of dependent variables as reported

43. With a view to reducing the complexity and the length of the task — since the CILS management had graciously agreed to allocate internal resources to this study – we decided to prioritize the evaluation of necessity over the one of effectiveness, as the former would have been more critical to evaluate than the latter, which could more easily be assessed by external evaluators.

in Section 6.3.2. Unless otherwise specified, all the analyses have been performed on 4,089 triples for the PE corpus and 4,933 triples for the REV corpus.

### 6.5.1 Amount of editing performed

HER scores show that PE assignments required significantly more edits compared to REV assignments. Specifically, these scores are systematically higher for post-edited sentences than for revised sentences in both language pairs (Table 6.6). Interestingly, we observe a contrasting pattern between the language pairs within each task. In the PE subcorpora, sentences in the *de-it* language combination exhibit a higher HER score than *de-fr* sentences. Conversely, in the REV subcorpora, French-speaking revisers modified a higher percentage of their human pre-translated sentences compared to Italian-speaking revisers.

| Task | Language combination | HER | Total HER |
|---|---|---|---|
| **PE** | *de-fr* | 35.210 | **36.317** |
| | *de-it* | 37.654 | |
| **REV** | *de-fr* | 23.073 | **20.774** |
| | *de-it* | 17.018 | |

Table 6.6 – HER scores, per task and per language combination.

While previous studies have demonstrated that proofreaders and revisers tend to engage in more extensive editing when rereading texts that have been post-edited by non-translators or inexperienced post-editors (Parra Escartín et al., 2017; Macken et al., 2022), it is important to consider the specific context of CILS workflows. In these workflows, in-house linguists primarily focus on revising the work of their colleagues or other external professional translators. As a result, the dynamics differ from those observed in other scenarios. To preserve the integrity of the original translation and respect the work of the professional translators, it is a common practice to instruct revisers to minimise their intervention when revising human-translated content. Post-editors are also encouraged to leverage the potential assistance of the MT engine by reusing its output as much as possible.

In light of these considerations, the observed discrepancy between HER scores in our corpora indicates that the customised NMT engines do not offer the same level of quality as human, professional translators. This outcome was expected: as aptly explained by do Carmo and Moorkens (2021, p. 40), MT should be seen as "a set of suggestions or hypotheses for the translation of a text" rather than a finalised translation.

The discrepancy observed between target languages in the PE subcorpora could potentially indicate that the customised NMT engine employed for translating texts into Italian yields slightly lower quality output when compared to the customised NMT engine used for French. Additionally, the difference in scores between target languages in the REV subcorpora suggests that either French-speaking revisers tend to intervene more frequently compared to their Italian-speaking counterparts, or that there are more issues present in human-pre-translated sentences in French than in Italian pre-translations.

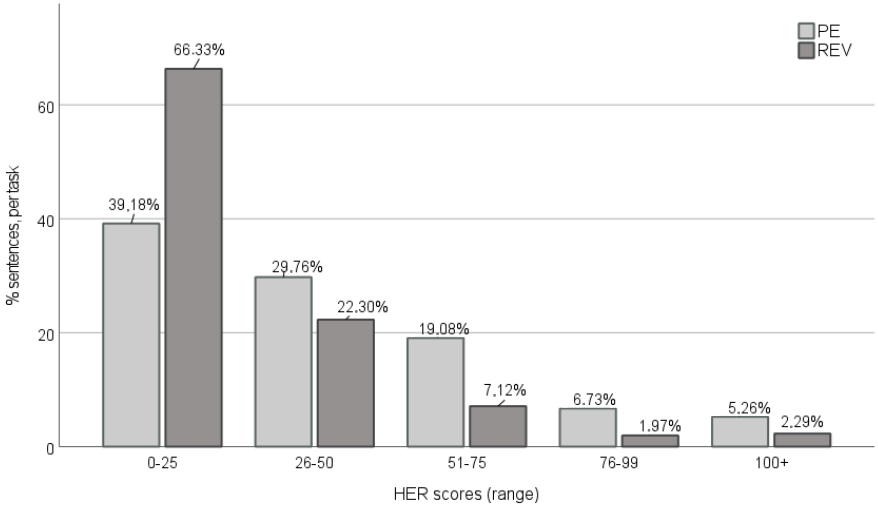



FIGURE 6.8 – Distribution of HER scores, per category (range) and task.

When looking at the distribution of edits, as represented by HER score ranges (Figure 6.8), we observe that 39% of the sentences in the PE corpus underwent some slight modification (i.e., HER scores up to 25). The percentage of the sentences edited during PE constantly decreases of ten points across the first three HER categories, with around 30% of sentences falling in the score range of 26-50 HER and 19% of sentences in which post-editors modified up to two thirds of the tokens (HER score range 51-75). While a smaller percentage of sentences in PE has been heavily or completely modified (around 7% and 5% falling in the two remaining HER categories, respectively), major restructuring is uncommon in revision. In fact, in the REV corpus, we notice a peak in the lowest HER range, meaning that in the vast majority of the sentences (66%) only a small percentage of the tokens has been modified. In 22% of the sentences in the same corpus, revisers have modified up to half of the tokens, while 7% of the sentences falls into the 51-75 HER category. Finally, only around 4% of the sentences in the REV corpus have been heavily or completely modified during revision [44].

Overall, the results suggest that revisers tend to primarily make slight modifications, whereas post-editors are confronted with the need for more extensive interventions across a larger number of sentences. However, it is noteworthy that a significant majority of sentences in the PE corpus undergo modifications of up to 50%, while over three-fourths of sentences in the REV corpus exhibit modifications of up to 25%. Therefore, we can infer that the disparity between the two tasks, in terms of the extent of editing required, is not substantial.

Before we move on with the analysis of our corpus, it is useful to focus on extreme HER scores. Although we are dealing with a corpus of modified sentences, the (inverted) TERcom script fails to record modifications performed on 146 sentences (57 in PE and 89 in

44. Aggregated result for HER categories 76-99 and 100 or above.

REV), which therefore obtain a HER score of zero. These cases concern sentences in which an unbreakable space has been inserted during the correction process, as in (39) [45];

(39) **HT (fr)**
*L'Autriche a levé ses restrictions de voyage avec la Suisse depuis le 4 juin 2020.*
'Austria lifted its travel restrictions with Switzerland on 4 June 2020.'
**REV (fr)**
*L'Autriche a levé ses restrictions de voyage avec la Suisse depuis le 4-juin-2020.*
'Austria lifted its travel restrictions with Switzerland on 4-June-2020.'

or in which an extra space has been inserted or removed, as in (40).

(40) **NMT (it)**
*Luogo : Live Talk (incontro informativo digitale)*
'Venue : Live Talk (digital briefing)'
**PE (it)**
Luogo: Live Talk (incontro informativo digitale)
'Venue: Live Talk (digital briefing)'

This kind of information is easily lost during the transfer of data between different file formats, or during the tokenisation process that foreruns the application of the TERcom script.

On the opposite extreme category, we find sentences that have obtained a HER score above 100, i.e., edited sentences that have more tokens than their unedited version. This is not uncommon in MTPE, especially when post-editing the output of NMT systems, which are known for their fluent mistranslations and the tendency to omit words (Castilho et al., 2017b; Van Brussel, Tezcan, and Macken, 2018; Vamvas and Sennrich, 2022). It is certainly less common in revision, where serious omissions are as unexpected as undesired. In our corpus, 162 sentences – 1.8% of the whole corpus – obtain a HER score above 100; 112 sentences in the PE corpus and 50 sentences in the REV corpus, and these are equally distributed between target language (81 sentences each). A closer inspection of these sentences illustrates our claim; the vast majority of sentences that obtained a HER score above 100 in both the PE and the REV corpus stems from the correction process itself, for instance to correct mistranslations, as in (41):

45. Parts that differ between pre-translated and final version have been highlighted to facilitate the comparison. The source text will be shown when necessary, otherwise excluded to avoid information overload.

(41) **Source (de)**
*Arbeiten, die keine Internet-Verbindung benötigen, wenn immer möglich* offline machen.
'Do work that does not require an internet connection offline whenever possible.'
**NMT (it)**
**Lavorare che non necessitano di un collegamento internet, possibilmente offline.*
*'To work that does not require an internet connection, possibly offline.'
**PE (it)**
Laddove possibile, si consiglia di svolgere offline le attività che non necessitano di una connessione internet.
'Wherever possible, it is recommended to carry out offline activities that do not require an Internet connection.'

and to improve text fluency or style, as in (42):

(42) **Source (de)**
*Mietwagen werden vollgetankt übernommen und müssen vollbetankt* zurückgegeben werden.
'Rental cars are taken over with a full tank and must be returned with a full tank.'
**HT (fr)**
*Les voitures de location sont réceptionnées avec le plein d'essence et rendues .* avec le plein d'essence
'Rental cars are delivered with a full tank of petrol and returned with a full tank of petrol.'
**REV (fr)**
Le plein d'essence a été effectué avant la prise en charge du véhicule de location et doit être effectué avant sa restitution à l'agence.
'The fuel tank has been filled before the rental vehicle is picked up and must be refilled before the vehicle is returned to the rental agency.'

Sometimes, extremely higher HER scores are the result of omissions repaired, particularly in the PE corpus, as in (43):

(43) **Source** *Wir helfen – Sie bleiben zu Hause!*
'We help – You stay home!'
**NMT (fr)** *On va vous aider.*
'We'll help you'
**PE (fr)** *Nous aidons – Vous restez chez vous !*
'We help – You stay home!'

The example sentence (43) obtained a HER score of 160. In our dataset, however, the highest HER score equals 1700 and it is assigned to a sentence where the NMT system failed to output a translation — therefore producing a severe omission — and the post-editor had to translate the source sentence from scratch. This case is shown in (44):

(44) **Source (de)**
4. *Zwingender Reisegrund (z.B. Pendler mit Arbeitsstelle im anderen Land).*
'Mandatory travel reason (e.g. commuter with a job in the other country).'
**NMT (it)**
4
**PE (it)**
4. *Necessità di mettersi in viaggio (ad es. pendolari che lavorano nell'altro paese).*
'Necessity to travel (e.g. commuters working in the other country).'

Omissions are less common, but still present, in the REV corpus, as in (45), where the human translator has left the sentence incomplete:

(45) **Source** (de) *Object: Neue Methodensammlung*
'Subject: new compendium of methods'
**HT (fr)** *Objet : nouvelle*
'Subject: new'
**REV (fr)** *Objet : nouveau recueil de méthodes*
'Subject: new compendium of methods'

Finally, some other scores in the most extreme category are the result of a segmentation issue in the corpus, as in (46), which comes from the PE corpus and obtained a HER score of 400, the second highest score in our dataset:

(46) **Source** (de) *Ob die*
'If the'
**NMT (fr)** *Si*
'If'
**PE (fr)** *Il faudra voir si*
'It remains to be seen whether'

The pre-translated sentence above cannot be considered an omission, as the NMT system produced a correct translation of the source text. Its mismatch with the post-edited version is rather the result of a corrupted segmentation of the original sentence. Table 6.7 summarises the reasons for extremely higher HER scores in the PE and the REV corpus, respectively.

| Reason | PE | REV |
|---|---|---|
| correction | 92 | 42 |
| omission | 17 | 4 |
| segmentation | 3 | 4 |
| **Total** | **112** | **50** |

Table 6.7 – Number of sentences that obtained a HER score above 100 (per task) and the reason for the score.

We now move on to analyse the amount of editing performed on our corpora, observing how edits affect text length. In Section 6.4.3, we have described the corpora in terms of number

of tokens at various steps in their production process. In Table 6.8, we retrieve these data and analyse the variation (in percentage) between versions.

| Task | | Source | Pre-translated | Target |
|---|---|---|---|---|
| **PE** | *de-fr* | 33,561 | 42,252 (+25.9%) | 43,633 (+3.27%) |
| | *de-it* | 30,527 | 34,990 (+14.62%) | 36,904 (+5.47%) |
| | ***all*** | **64,088** | **77,242 (+20.53%)** | **80,537 (+4.27%)** |
| **REV** | *de-fr* | 49,093 | 62,806 (+27.93%) | 63,464 (+1.05%) |
| | *de-it* | 33,192 | 38,448 (+15.84%) | 39,233 (+2.04%) |
| | ***all*** | **82,285** | **101,254 (+23.05%)** | **102,697 (+1.43%)** |

Table 6.8 – Number of tokens in corpora, per task and language combination. Percentages between parentheses indicate expansion rates and refer to the previous version of the text considered (left column).

In the translation step, we observe an overall increase in the number of tokens, which is attributable to systemic differences between the source and the target languages. Nevertheless, we notice that the expansion rate is higher for human-translated texts (+23.05%) than for machine-translated texts (+20.53%). We also notice an expansion in the number of tokens between pre-translated content and final texts, in both tasks, but most prominently in PE (+4.27% versus +1.43% in REV). Finally, the overall expansion rate between source texts and target texts is slightly higher in PE workflows (+25.67%) than in REV workflows (+24.8%). When considering target language subcorpora, we observe that, during the translation step, the subcorpus *de-fr* undergoes the highest expansion in terms of number of tokens. During the correction step, however, this tendency is inverted, and Italian-speaking revisers and post-editors tend to add more words to the pre-translation than their French-speaking colleagues.

Let us now look at the average sentence length of different subcorpora at various steps in the production process (Table 6.9).

| Task | Source | | Pre-translated | | Target | |
|---|---|---|---|---|---|---|
| **PE** | *M* | 15.67 | *M* | 18.89 | *M* | 19.7 |
| | *Mdn* | 14 | *Mdn* | 17 | *Mdn* | 18 |
| | *SD* | 8.84 | *SD* | 10.658 | *SD* | 11.072 |
| | *CV* | 0.564 | *CV* | 0.564 | *CV* | 0.562 |
| **REV** | *M* | 16.68 | *M* | 20.53 | *M* | 20.82 |
| | *Mdn* | 15 | *Mdn* | 18 | *Mdn* | 19 |
| | *SD* | 9.567 | *SD* | 12.078 | *SD* | 12.15 |
| | *CV* | 0.574 | *CV* | 0.588 | *CV* | 0.584 |

Table 6.9 – Average sentence length (*M*), median (*Mdn*), standard deviation (*SD*) in tokens (punctuation included) and coefficient of variation (*CV*) of different corpus versions, per task.

Besides the expected expansion in the translation between the German source and the Italian and French pre-translated texts (+20.55% in PE and +23.08% in REV), we notice that the average sentence length increases in post-edited and in revised texts, but to a

different extent. Indeed, post-edited texts are on average 4.29% longer than machine pre-translated texts, while the average sentence length in revised texts increases on average by 1.41% compared to their human-translated, unrevised version.

The higher coefficients of variation in REV show that there is slightly more variation in the sentence length of revised texts (0.584) compared to the post-edited ones (0.562). It is worth noting that there was already slightly more variation in the sentence length of the German source texts in the REV corpus (0.574) compared to the PE one (0.564). Nevertheless, while in the PE corpus the coefficient of variation remains the same between the source texts and their machine pre-translated version (0.564), in human-pre-translated texts there is slightly higher variation in sentence length (0.588) compared to their corresponding source texts in German (0.574). In both the PE and the REV corpora, the difference between the average sentence length of the pre-translated and target text is statistically significant ($p<.001$), as per the results of a t-test for paired samples [46].

Our results contradict those of Robin (2019), who found a general reduction in the number of words during revision in Hungarian. This has been interpreted by the author herself as "an effort to reduce the effects of overused explicitation [...] that result in a more redundant translation than its source equivalent or authentic texts originally written in the target language" (*ibid.*, p.133). Following the same line of reasoning, our results seem to point to an explicitation tendency of the linguists during the correction stage in both tasks. Obviously, it would be incautious to make such claim at this stage of analysis: first, because we are dealing with different language pairs compared to Robin (2019), and second, because we do not know yet what kind of edits have been performed. Nonetheless, the systematically higher expansion rate in the number of tokens during PE seems to indicate that post-editors, more often than revisers, add more words to the machine-translated sentences in order to achieve a more natural sentence length in final texts. Indeed, considering that the source sentences in the REV corpus were on average one token longer than source sentences in the PE corpus, we notice that the average sentence length of PE target texts somehow matches that of revised texts.

Table 6.10 illustrates the comparison of average sentence length in the different target language subcorpora. In the PE corpus, we observe a marked expansion between source and machine-pre-translation for texts in French (+25.88%) compared to those in Italian (+14.68%). The same happens in the REV corpus, where the language combination German-French undergoes the highest expansion between the source sentences and the human-pre-translations (+27.96%) compared to sentences in the language combination German-Italian (+15.83%). During the correction process, however, this tendency is inverted; in PE, final texts in Italian are 5.43% longer than pre-translations, while final texts in French are 3.26% longer than their machine-translated counterparts. For the REV corpus, revised Italian texts are 2.04% longer than human-pre-translated texts, while, on average,

46. PE $t(4{,}088)=18.195$ ; REV $t(4{,}932)=8.745$.

| Task | Language combination | Source | Pre-translated | Target |
|---|---|---|---|---|
| PE | *de-fr* | *M* 15.61 | *M* 19.65 | *M* 20.29 |
| | | *Mdn* 14 | *Mdn* 18 | *Mdn* 18 |
| | | *SD* 8.797 | *SD* 10.978 | *SD* 11.370 |
| | | *CV* 0.564 | *CV* 0.559 | *CV* 0.560 |
| | *de-it* | *M* 15.74 | *M* 18.05 | *M* 19.03 |
| | | *Mdn* 14 | *Mdn* 17 | *Mdn* 17 |
| | | *SD* 8.889 | *SD* 10.229 | *SD* 10.696 |
| | | *CV* 0.565 | *CV* 0.567 | *CV* 0.562 |
| REV | *de-fr* | *M* 16.81 | *M* 21.51 | *M* 21.73 |
| | | *Mdn* 15 | *Mdn* 19 | *Mdn* 19 |
| | | *SD* 9.739 | *SD* 12.735 | *SD* 12.814 |
| | | *CV* 0.579 | *CV* 0.592 | *CV* 0.590 |
| | *de-it* | *M* 16.49 | *M* 19.10 | *M* 19.49 |
| | | *Mdn* 15 | *Mdn* 17 | *Mdn* 17 |
| | | *SD* 9.311 | *SD* 10.901 | *SD* 10.985 |
| | | *CV* 0.565 | *CV* 0.571 | *CV* 0.564 |

TABLE 6.10 – Average sentence length (*M*), median (*Mdn*), standard deviation (*SD*) in tokens (punctuation included) and coefficient of variation (*CV*) of different corpus versions, per task and per language combination.

revised French texts are only 1.02% longer than their unrevised version. Finally, the coefficients of variation tend to be slightly lower in German source texts than in French and Italian pre-translations, except for the French PE subcorpus. Overall, except for the French PE subcorpus, sentence length variation tends to be slightly reduced during the correction process. In light of these results, we can elaborate on the insights we discussed regarding the number of tokens in language-specific subcorpora. The consistent increase in average sentence length between pre-translations and final texts is particularly pronounced in Italian, and we propose that this phenomenon can be attributed to the customised NMT system's relatively lower quality compared to that of the system used for French. Consequently, Italian post-editors, more frequently than revisers (in both languages) and French post-editors, find themselves needing to incorporate additional words into the machine-translated sentences to attain a more natural sentence length in the final texts.

Paired-samples t-tests were conducted to compare the average sentence lengths in pre-translated and final texts, in each language combination and task. All the differences have been found statistically significant ($p<.001$) [47].

We now go ahead with the analysis of HER score ranges, this time focusing on the average number of tokens per sentence (Figure 6.9). The average number of source tokens decreases as the HER scores increase. Indeed, longer sentences (PE $M$=17 tokens; REV $M$=18 tokens) are in general those that underwent the smallest modifications (i.e., fall in the HER range 0-25), while shorter sentences obtain higher scores. The latter observation, however, does not correspond systematically to a major restructuring of the sentences, but

47. PE fr, $t(2,149)= 10.743$; PE it, $t(1,938)= 15.058$; REV fr, $t(2,919)= 4.539$; REV it, $t(2,012)= 9.984$.

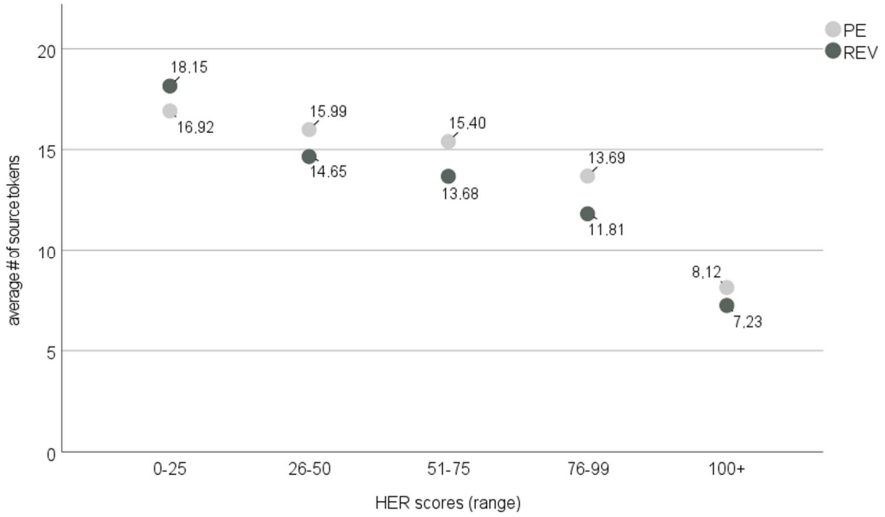



FIGURE 6.9 – Average source tokens per HER score category in each task.

is rather due to the formula used to calculate HER, which is sensitive to the length of the string. For instance, one single edit on a string of three tokens will lead to a higher score compared to the same edit on a string of ten tokens. The same tendency is observable when considering the average number of tokens in pre-translated texts – i.e., the "unedited" segments – in both languages (Table 6.11).

| | | | **HER (range)** | | | | |
|---|---|---|---|---|---|---|---|
| | | | **0-25** | **26-50** | **51-75** | **76-99** | **100+** |
| **# tokens pre-translation (Mean)** | **PE** | *de-fr* | 21 | 21 | 18 | 16 | 8 |
| | | *de-it* | 20 | 18 | 18 | 16 | 9 |
| | **REV** | *de-fr* | 24 | 19 | 18 | 15 | 9 |
| | | *de-it* | 21 | 15 | 13 | 10 | 5 |

TABLE 6.11 – Average number of tokens in pre-translated texts, per HER score (range), in each task and target language.

A Pearson's correlation coefficient was computed to assess the relationship between the number of tokens (in source and in pre-translated sentences) and the amount of editing performed on pre-translated sentences, as represented by the number of edits (NumEr) automatically recorded by the TERcom script. The results of these tests are displayed in Figure 6.10 [48].

In the PE corpus, we found a positive, moderate correlation between the number of source tokens and the number of edits in pre-translated sentences (r=.506, $p<.001$); while for sentences in the REV corpus we detected only a weak, positive correlation between these two variables (r=.334, $p<.001$). The same holds true for the correlation between the number of tokens in pre-translated sentences and the number of edits performed; the correlation is

48. It is worth noting that, for this test, we decided to exclude segments that obtained extreme HER scores, in order to avoid sentences for which no editing was recorded by the HER script, as well as sentences with segmentation issues that would distort the results.

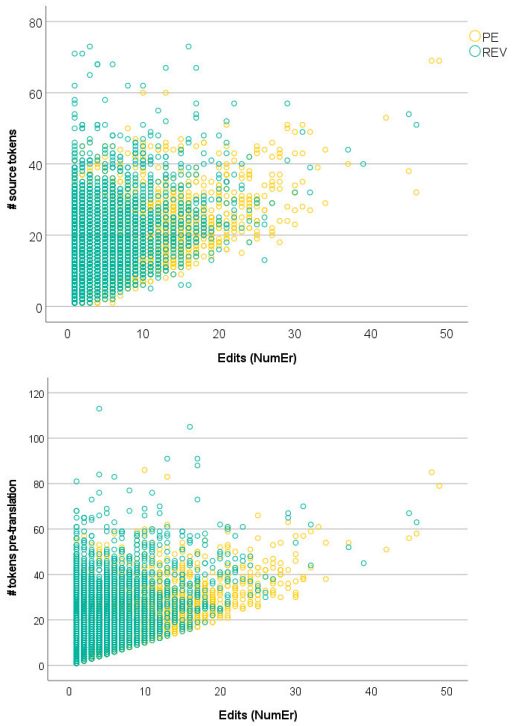



FIGURE 6.10 – Correlations between number of tokens and edits performed (NumEr, as recorded by the TERcom script), per task. Sentences that obtained extreme HER scores (HER 0 or HER>100) have been excluded from the analysis. Upper side: source tokens; lower side: pre-translated tokens.

positive and moderate for sentences in the PE corpus (r=.524, $p$<.001), while it is positive and weak for sentences in the REV corpus (r=.334, $p$<.001).

For the PE corpus, these results do not come as a surprise, since several studies in MT research have reported that the quality of the MT output degrades as source segment length increases (see for instance, Koehn and Knowles, 2017; Toral and Sánchez-Cartagena, 2017; Neishi and Yoshinaga, 2019). In the REV corpus, however, the weak correlation between sentence length and number of edits in human pre-translated sentences shows that longer sentences do not necessarily require more modifications compared to shorter ones. It remains to be proved whether, in longer sentences, these edits are applied to correct translators' errors or can rather be considered optional modifications.

Finally, as do Carmo (2021b) pointed out, we recall that NumEr values indicated by the TERcom script are an overestimation of the actual number of edits performed on the sentence, as they combine editing actions and number of edited words. Nonetheless, we can expect the script to display the same behaviour on both corpora, which would still allow for a meaningful comparison between the PE and the REV task.

### 6.5.2 Distribution of editing actions

We analysed the type of edits performed in PE and in REV, as recorded by the TERcom script (Figure 6.11). We recall that, while insertions, deletions and substitutions are recorded at word level (meaning that the insertion of two words is counted as two insertions), shifts

of contiguous words are always recorded as one edit. Therefore, to provide a more coherent analysis of edit types in our corpora, we decided to report the number of words shifted (WdSh in the TERcom script) [49], instead of simple shifts. In order to compare the data between the two tasks, we report the proportion of edit types on the total of edited words in each task.

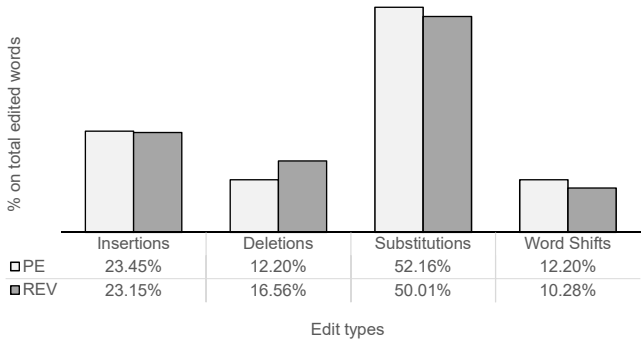



FIGURE 6.11 – Type of edits performed, per task (percentage on the total of words edited in each task).

We found that substitutions are the most frequent editing operation in both tasks (52% and 50% on the total number of words edited in PE and in REV, respectively), followed by insertions (23% in both tasks) and deletions (12% in PE and 17% in REV). Shifts are less frequent in REV (10%) than in PE assignments (12%).

Previous studies using TER had already reported this order of operations (Snover et al., 2006; Mutal et al., 2019; do Carmo et al., 2021), but with deletions as the second most common edit type, because of the inverted order of hypothesis and reference in the script (see Section 6.2.2). Shifts are usually the less frequent category; however, since it is common to report results for simple shifts, rather than words shifted, our findings are not completely comparable with those from previous studies. Indeed, in our PE corpus, we found the same percentage of words deleted and shifted.

A Chi-squared test of independence was used to test the difference in the distribution of editing actions between tasks, which was found statistically significant ($p<0.001$). This indicates that, although the percentage breakdown of editing actions appears similar between PE and REV, there is an association between the task and the type of editing action performed. Indeed, we observe a substantial deviation in the number of deletions in REV assignments which is higher than expected.

If we observe the distribution of edit types per HER score ranges (Figure 6.12), we notice that it remains consistent across the different HER categories, except for sentences that recorded a HER scores of 100 or above, in which – obviously – replacements, insertions, and even word shifts outperform deletions in both tasks.

49. do Carmo (2021b) has shown that shifts are difficult to identify for the TERcom script. In particular, the script tends to underestimate the exact number of words shifted. Nonetheless, since the other edit types are determined at word level, reporting word shifts can be considered more appropriate than reporting simple shifts.

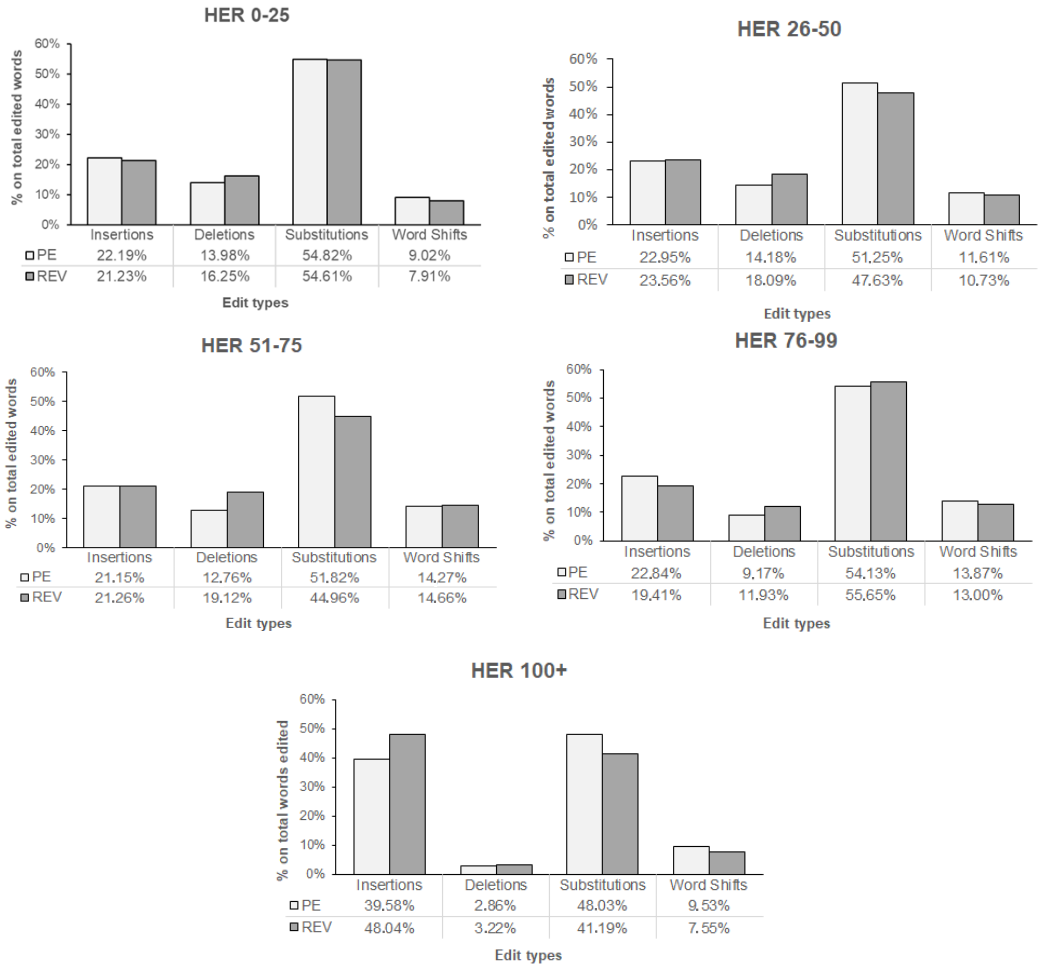



FIGURE 6.12 – Distribution of edit types per HER range in each task. Percentages are intended as number of edit types on the total number of words edited in each task.

Overall, we notice that the percentages of edit types are very similar in both tasks, with a slightly higher preponderance of deletions in REV compared to PE. This seems to mirror what previously observed for the average sentence length and number of tokens between pre-translations and final texts. Although there is a general tendency for expanding the number of words (as shown in the previous section), in human-translated texts it seems that revisers tend to delete more words compared to what post-editors do on the MT output. Indeed, if we look at sentences in which deletions outperform insertions, we find that they account for the 2.4% (n=987) of the sentences in the PE corpus, and for 25.54% (n=1,260) of the sentences in the REV corpus.

This result is also in line with previous studies on the product of revision and proofreading. For instance, Laflamme (2009) has found that deletions outperform insertions, while Parra Escartín and Goulet (2021) reported that a professional proofreader performed slightly more deletions during the proofreading process compared to non-professional post-editors during PE.

### 6.5.3 Parts of speech

We calculated the percentage of edited parts of speech (POS) categories on the total number of tokens modified in each task, by looking at pre-translated texts (Figure 6.13).

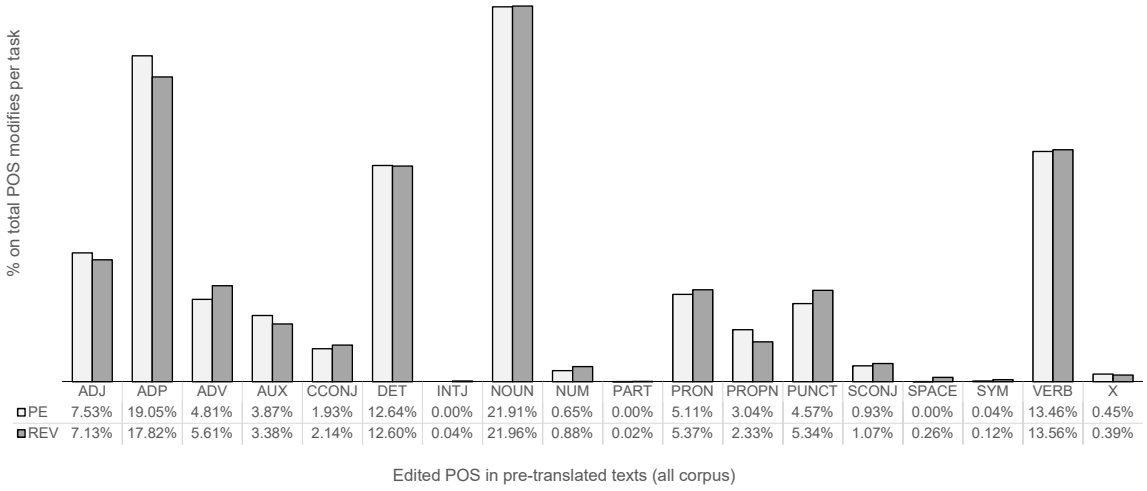



FIGURE 6.13 – Percentage of POS modified on the total number of POS modified (= deleted, replaced or moved) in pre-translated corpora, per task.

The most frequently modified POS categories in both tasks were nouns (21.91% for PE and 21.96% for REV) and prepositions (ADP; 19.05% for PE and 17.82% for REV), followed by verbs (13.46% for PE and 13.56% for REV) and determiners (12.64% for PE and 12.60% for REV). Overall, we observe that the distribution of edits on POS is almost identical between the two tasks. Slight differences can be found in the distribution of modifications on punctuation marks (4.57% in PE and 5.34% in REV), adverbs (4.81% in PE and 5.61% in REV) and proper nouns (3.04% in PE and 2.33% in REV). The only notable difference (above one percentage point) between the two tasks concerns the frequency of modifications to prepositions.

A Chi-square test of independence was used to test the difference in the distribution of POS affected by edits between tasks. The difference was found to be statistically significant ($p<0.05$), indicating an association between the task and the type of POS affected by edits.

When focusing on pre-translated texts falling in the HER category 0-25 (Figure 6.14), some notable (i.e., more than one percentage point) differences arise. For instance, REV assignments contain a higher percentage of edits on punctuation marks (6.17%) compared to PE assignments (4.79%), while in PE the modification of adjectives is more common (8.48%) than in REV (6.93%).

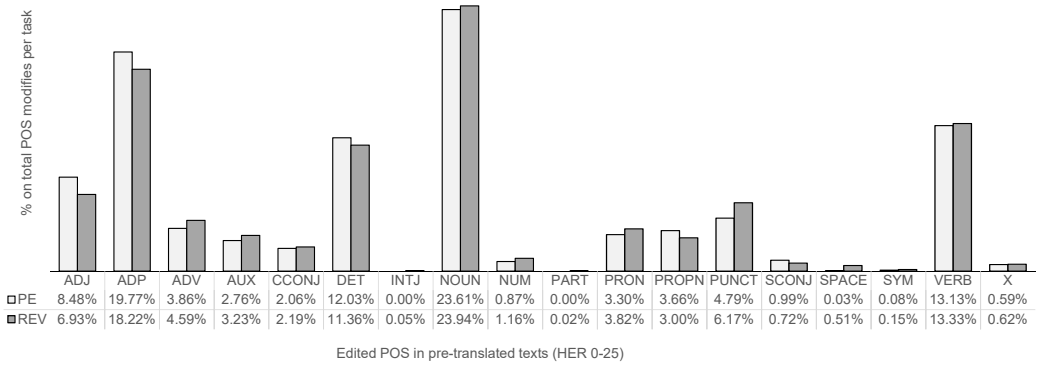



FIGURE 6.14 – Percentage of POS modified on the total number of POS modified (= deleted, replaced or moved) in pre-translated texts in the 0-25 HER range, per task.

Our results differ from those reported in previous studies. For example, professional revisers in Laflamme's study (2009) modify mainly verbs (24.3%), followed by nouns (18.8%), adverbs (13.8%) and prepositions (11%), while determiners lag behind (9.7%). In Goulet et al. (2017) the authors used a different taxonomy to identify POS categories and found that a professional proofreader modified mostly nouns, determiners and phrases. In Parra Escartín and Goulet (2021), phrases are the element most often modified during both PE and proofreading. When looking at single POS categories, nouns, verbs and determiners are the POS categories most frequently edited. These authors also report considerable differences in the POS categories modified by several non-professional post-editors and one professional proofreader.

As discussed in Section 6.4.4, our approach for identifying modified POS in pre-translated texts is limited to tokens that have been deleted, replaced, or moved during PE or REV. Unfortunately, this method does not account for inserted tokens, as they are only visible in the final texts and are not easily distinguishable from replaced tokens. The HER metric, as previously discussed, is also not suitable for identifying these inserted tokens. Indeed, the script used for HER does not differentiate between *logical insertions* aimed at replacing existing content and true insertions. To address this limitation, we manually annotated true insertions in final texts falling within the HER 1-25 range (as shown in Figure 6.15). This manual annotation process was necessary to accurately identify and analyse the inserted tokens that were not captured by our previous methods.

The analysis of true insertions in PE and REV segments falling in the HER 1-25 range

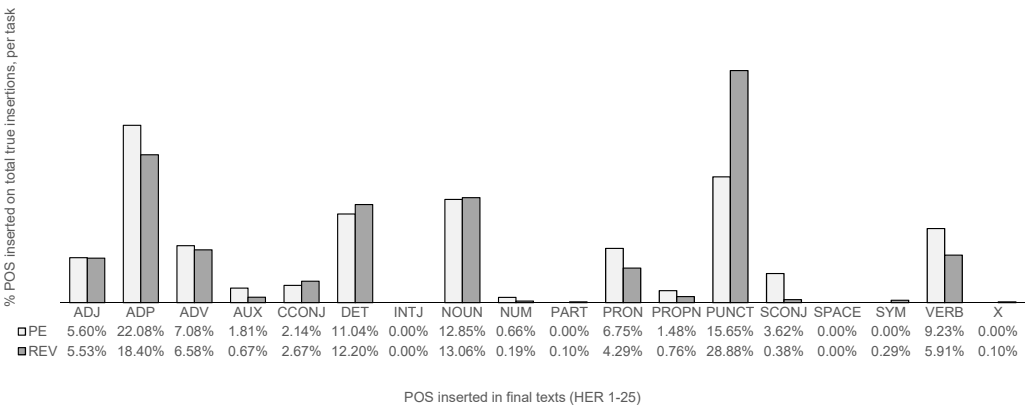



FIGURE 6.15 – Percentage of POS inserted on the total number of true insertions in the 1-25 HER range, per task.

reveals notable patterns. In both tasks, the most frequently inserted POS comprise prepositions, punctuation marks, nouns and determiners, collectively accounting for over 60% of the inserted tokens in PE assignments. This percentage increases to over 70% in REV assignments. Figure 6.15 also exhibits interesting differences between the tasks. Specifically, prepositions (22.08%), punctuation marks (15.65%) and nouns (12.85%) are the POS most commonly inserted in PE assignments. In contrast, punctuation marks (28.88%) rank highest among frequently inserted POS in REV assignments, followed by prepositions (18.40%) and nouns (13.06%). These findings suggest that the NMT system failed to generate appropriate prepositions, whereas human translators overlooked punctuation marks, leading revisers to add them for enhanced text readability.

In her study, Laflamme (2009, p. 120) examined the relationship between POS modified and the editing action performed by professional revisers. The findings of her study indicated that adverbs, verbs, and nouns were frequently inserted by revisers. Nevertheless, a direct comparison of these results with our own is not feasible, mainly due to two reasons. Firstly, Laflamme's study encompasses not only the insertion of lexical units, but also the insertion of characters within existing words. In contrast, our study specifically defines insertions as the addition of tokens, while the insertion of characters within existing tokens would be categorised as a substitution. Secondly, Laflamme's analysis focused solely on lexical modifications, omitting punctuation marks from the scope of her investigation.

To conclude, our analysis of POS modified in pre-translated texts revealed that PE and revisers focus on mainly on modifying the same parts of speech (i.e., nouns, prepositions, verbs and determiners), but some POS are typical of a particular task. Indeed, revisers tend to focus more on punctuation compared to post-editors, while the latter focus more on adjectives compared to the former. Punctuation is also the most frequent POS inserted in revision tasks, while post-editors insert most often prepositions.

### 6.5.4 Linguistic dimensions

The analysis of the linguistic dimensions involved in editing PE and REV assignments was carried out on a subset of sentences falling in the HER range 1-25. We recall that this subset consisted of 150 PE triples and 150 REV triples for each language, corresponding to a total of 436 edits for the language combination *de-fr* and 437 for *de-it*. In this section, we will first present aggregated results, then we will focus on the individual target language subsets. As the PE and the REV subset contained the same number of segments, but a different number of edits overall, we will report relative percentages, as already done in previous sections.

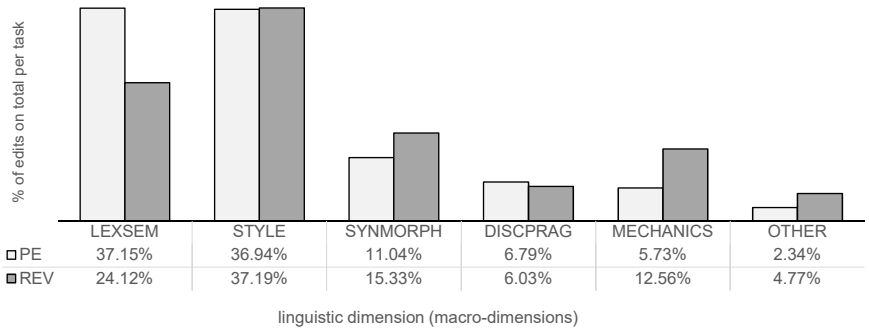



FIGURE 6.16 – Edit distribution by linguistic dimension involved (macro-dimension), per task.

Figure 6.16 illustrates that the majority of edits in the PE corpus belong to the category Lexis and semantics (37.15%), while the Style category follows closely behind (36.94%). Conversely, edits in the REV corpus come mostly from the Style category (37.19%), while lexico-semantic edits are less common (24.12%). The third most common category in both tasks is Syntax and morphology, accounting for 11.04% and for 15.33% of edits in PE and in REV, respectively. The next most frequent linguistic dimension for REV (12.56%) is Mechanics, i.e., edits on spelling and punctuation, which mirrors the results for edited POS (Section 6.5.3). Compared to REV, segments from PE assignments are more concerned with information restructuring (category Discourse and pragmatics, 6.79%). Finally, the residual category (Other) includes cases of edits that involve more than one linguistic dimension or whose linguistic dimension is unclear. Notably, the REV subset displays a slightly higher proportion of edits in this category (4.77%) compared to PE (2.34%).

When inspecting these results more closely, interesting differences arise between the two target language subsets (Figure 6.17). For instance, in the FR subset, stylistic edits represent the largest proportion of edits in both PE (42.61%) and REV (33.82%), whereas in the IT subset, stylistic edits are the largest linguistic dimension only among revisers (40.72%), while post-editors perform mainly lexico-semantic edits (40.66%). The third largest dimension for REV also varies between the two subsets: in the FR subset, we found edits related to Syntax and morphology (19.12%), while in the IT subset, we found edits related to spelling and punctuation (*Mechanics*, 12.89%). Finally, while in Figure 6.16 we identified a majority of REV edits falling under the Other category, we now understand that this is due to a preponderance of this category in the IT subset.

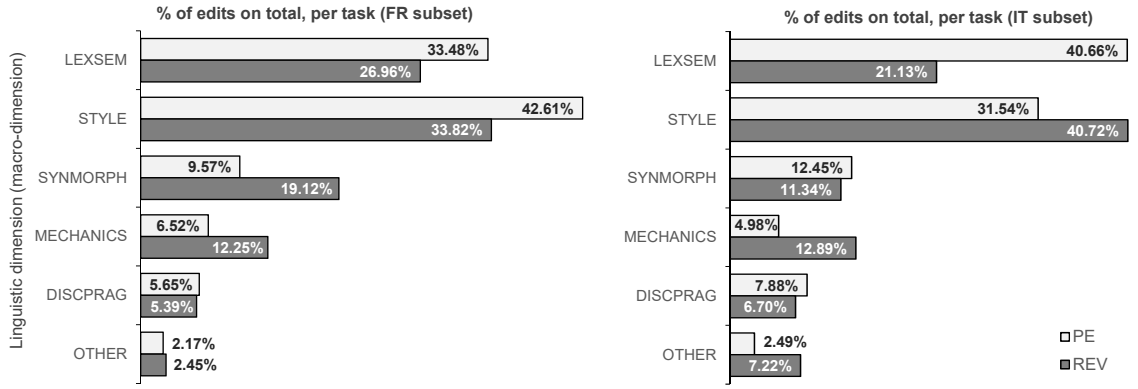



FIGURE 6.17 – Edit distribution by linguistic dimension involved (micro-dimension), per task and per target language.

Overall, our results partly confirm those of Parra Escartín and Goulet (2021), who report that the Style category is the most frequent linguistic dimension in both PE and proofreading. In their study, the Syntax dimension ranks second for both tasks, while the third most common linguistic dimension is Terminology for post-editors and Typography/orthography for the professional proofreader (which partly corresponds to our Mechanics dimension). The professional proofreader in the study by Goulet et al. (2017), too, performs mainly edits related to syntax, style and typography or punctuation, while only a tiny percentage of his/her modifications relate to semantics. However, as previously discussed, these studies deal with non-professional post-editors who post-edit into their L2; therefore, issues in grammar and syntax are more likely to emerge. Instead, Laflamme (2009) reports that professional revisers in her study perform mainly edits related to the dimension of semantics. Syntax and morphology (two separate categories) come second and third, respectively, while style only ranks fourth. As previously mentioned, this author does not take into account edits involving punctuation marks. Additionally, the author herself acknowledges that the six revisers in her study exhibit varying levels of experience, potentially influencing their capacity to detect and address semantic issues [50].

Let us now look at the distribution of edits across micro-dimensions (Figure 6.18). This fine-grained classification highlights some notable differences between PE and REV. First and foremost, the analysis of edits in PE assignments revealed that their primary focus is on addressing mistranslations, i.e., content distortions and lexical misselections (26.75% of the edits), suggesting that the NMT pre-translation still generates a substantial number of mistranslations compared to human pre-translated texts. It is noteworthy to mention that the customised NMT system was implemented for the first time during the Covid-19 pandemic, and therefore, it was not trained on Covid-19-related texts and terminology, leading to a higher occurrence of mistranslations related to this domain in the PE corpus. In contrast, the largest micro-dimension of edits in REV assignments involves providing synonyms or

50. "It seems that the more experienced revisers are, the more they detect and correct semantic issues [...]". « [Il] semble que plus les réviseurs sont expérimentés, plus ils détectent et corrigent des problèmes sémantiques [...] » (*ibid.*, p. 122, our translation).

rephrasing lexical content (20.10%). Nonetheless, revisers amend several content distortions and lexical misselections (14.82% of the edits in REV), and rephrasing is frequently applied by post-editors, too (15.71% of edits in PE).

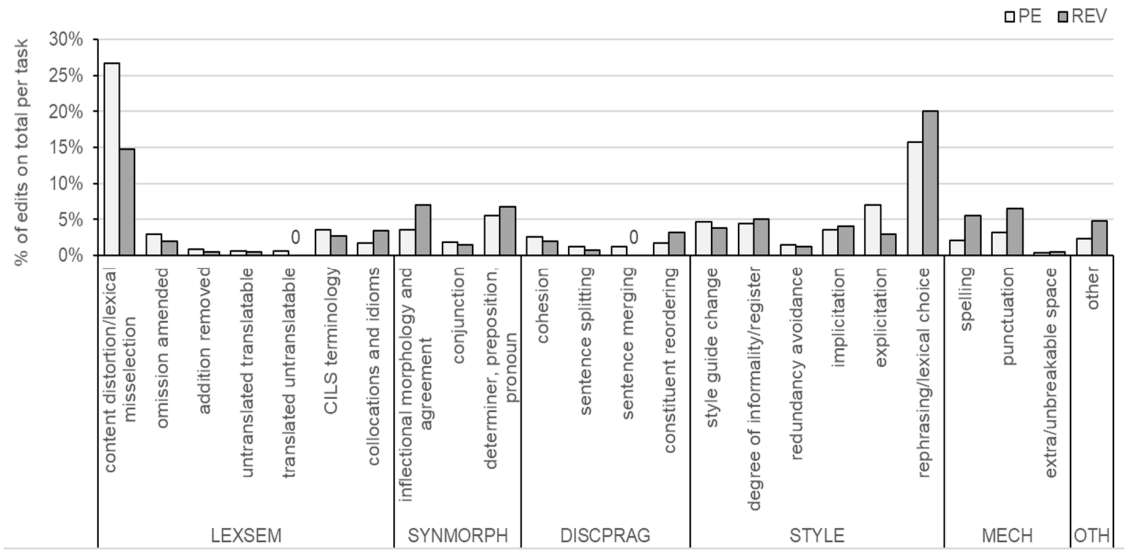



FIGURE 6.18 – Edit distribution by linguistic dimension involved (micro-dimension), per task.

Several other micro-dimensions exhibited notable differences between the two tasks, although the overall percentages of such edits were relatively lower. For instance, edits related to inflectional morphology and agreement are more common in REV (7.04%) than in PE (3.61%). The same holds true for edits related to determiners, prepositions and pronouns (6.78% in REV and 5.52% in PE), and to collocations and idioms (3.52% in REV and 1.70% in PE). More constituent reordering is also applied in REV (3.27%) than in PE (1.70%). Additionally, spelling and punctuation issues, as well as edits falling into the Other category, were more frequently observed in REV assignments. Instead, explicitation was found to be more common in PE (7.01%) than in REV (3.02%). The percentages of edits in all the other micro-dimensions are relatively similar between the two tasks. Interestingly, there were instances of merged sentences during PE assignments (1.27% of the edits), whereas no such examples were found in REV assignments included in our subset.

Let us take a closer look at the distribution of micro-dimensions in the two target languages considered (Figure 6.19). Analysing the results for the Italian subset, we observe a considerable difference in the percentage of edits performed to address content distortions and lexical misselections between PE (29.9%) and REV (11.3%). This discrepancy further highlights the disparity in the performance of the NMT engine used for Italian compared to the one used for the French subset. In contrast, the differences between PE and REV for the same category in the French subset are less pronounced.

Turning our attention to the French subset, we observe that the percentage of edits related to *Inflectional morphology and agreement* is much higher in REV (10.8%) compared to PE (3%). This should not be surprising, since NMT systems are less prone to commit grammar errors compared to previous generations of MT systems. In this case, it seems

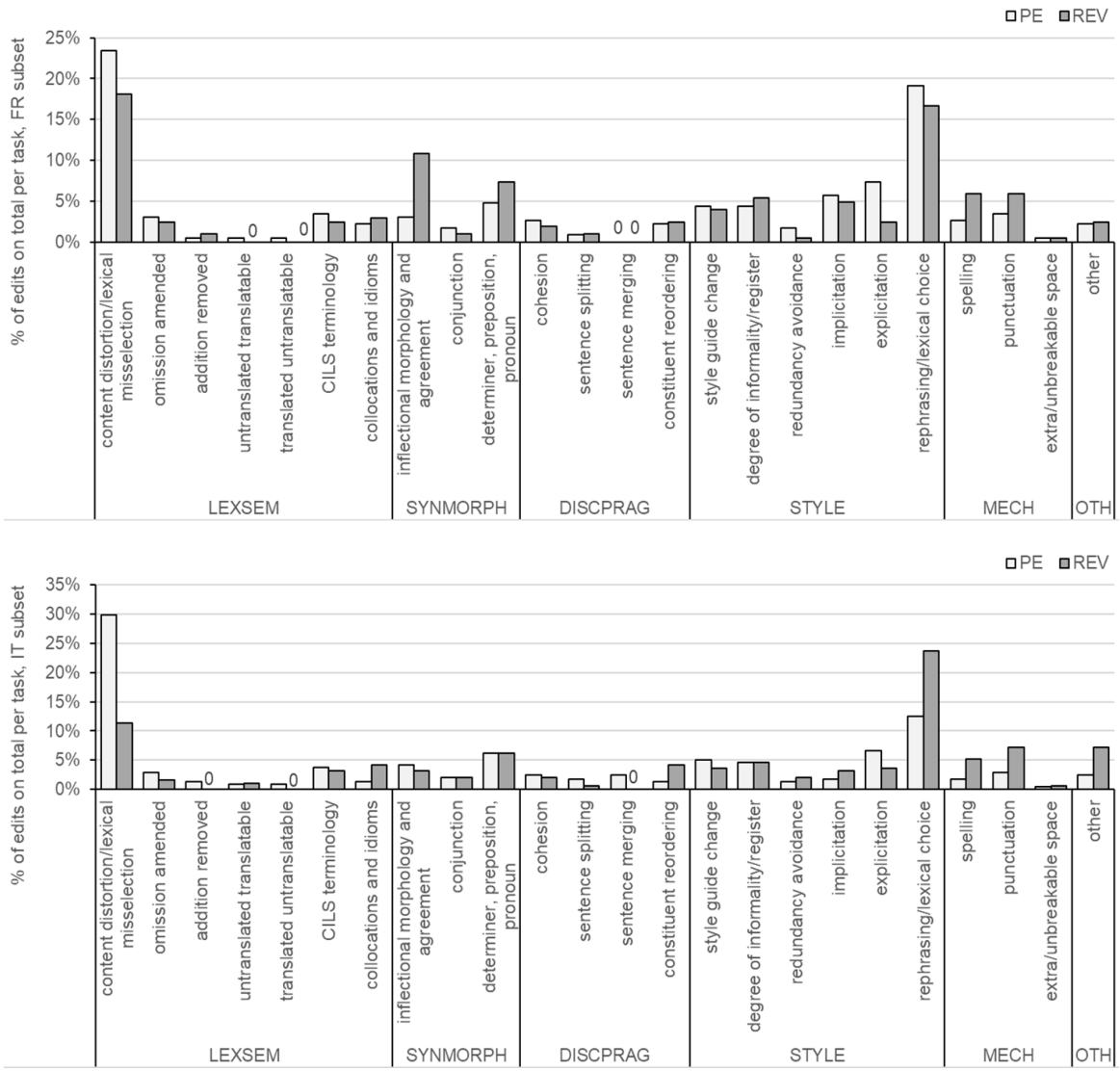



FIGURE 6.19 – Edit distribution by linguistic dimension involved (micro-dimension), per task and per target language (upper side: French subset; lower side: Italian subset).

that the NMT system for French is also likely to commit fewer grammar errors compared to humans. Nonetheless, it should be verified whether these edits correspond to actual errors or rather to stylistic preferences.

Similarly, in the micro-dimension *Determiner, preposition, and pronoun*, the percentage of edits is slightly higher in REV (7.4%) than in PE (4.8%). In the Italian subset, however, there is no substantial difference between PE and REV tasks for the micro-dimension *Inflectional morphology and agreement* (REV 3.1% and PE 4.2%), and both tasks exhibit the same percentage of edits falling in the micro-dimension *Determiner, preposition, and pronoun* (6.2%). Notable differences between target language subsets can be observed in the micro-dimension *Rephrasing and lexical choice*. In the Italian subset, a higher percentage of edits falls into this category for REV assignments (23.7%) compared to PE assignments (12.4%). Conversely, in the French subset, the opposite trend is observed, with a higher percentage of edits in PE (19.1%) compared to REV (16.7%), even if the difference between the two subsets is not as evident as in the Italian subset. In the Italian subset, we also observe a slight disparity in the percentage of edits made to amend omissions in PE assignments (2.9%) compared to REV assignments (1.5%). Notable differences between tasks are evident when considering the micro-dimension *Explicitation*. In both target language subsets, PE assignments involve a higher percentage of edits aimed at explicitating textual content. This distinction is slightly more prominent in the French subset, with 7.4% of edits in PE and 2.5% in REV, compared to the Italian subset, where the percentages are 6.6% in PE and 3.6% in REV. Furthermore, in the Italian subset, instances of *Implicitation* are more frequent in REV assignments (3.1%) than in PE assignments (1.7%). Conversely, in the French subset, PE assignments include slightly more edits (5.7%) involving the implicitation of textual content compared to REV assignments (4.9%).

These data can be compared with the findings of the study conducted by Macken, Prou, and Tezcan (2020), although the micro categories defined in their study may not correspond precisely to ours. The authors reported that instances of explicitation, implicitation, and sentence splitting were more frequent in REV compared to PE. They also found that structural changes were more prevalent in REV than in PE. Our results partially contradict these findings. In our dataset, we observed a higher occurrence of explicitation instances in PE assignments rather than in REV assignments. However, we did not observe a notable difference in implicitation instances between PE and REV. Additionally, we found a limited number of sentence splitting instances, with a similar distribution between PE and REV. Regarding the micro-dimension *constituent reordering*, which can be associated with Macken et al.'s *structural changes*, we observed a slightly higher frequency in REV than in PE, thus confirming Macken et al.'s results. However, it is important to note that the study by Macken, Prou, and Tezcan (2020) focused on a different language pair (English to Dutch) and a different textual genre (literary text) compared to our corpus of authentic texts in the "Communication" domain. It also captured the edits of only one post-editor and of one reviser working on the post-edited text.

### 6.5.5 Necessity

Figure 6.20 shows the results of the human evaluation of edits' necessity, conducted with three evaluators per target language.

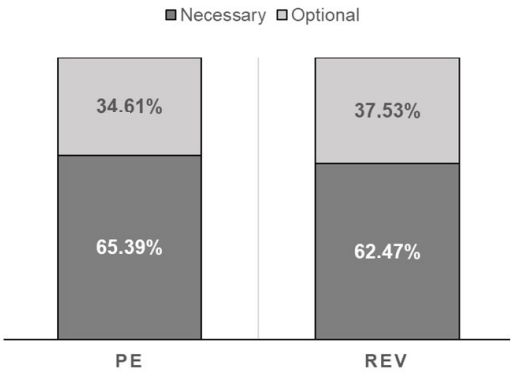



FIGURE 6.20 – Percentage of necessary and optional edits per task.

Overall, the results indicate that for both types of tasks the majority of edits are necessary rather than optional, suggesting that the pre-translations required significant improvement. However, the proportion of necessary edits is slightly higher for PE assignments (65.39%) than for revision assignments (62.47%). This was somehow expected, as we have seen in previous sections that PE segments, compared to REV ones, contain a higher number of edits belonging to the *Lexis and semantics* dimension – particularly those performed to amend content distortions or lexical misselections, which are commonly recognised as necessary. When examining the results per target language subset, we observe notable differences in the perception of edits' necessity between the Italian and French subsets (Figure 6.21).

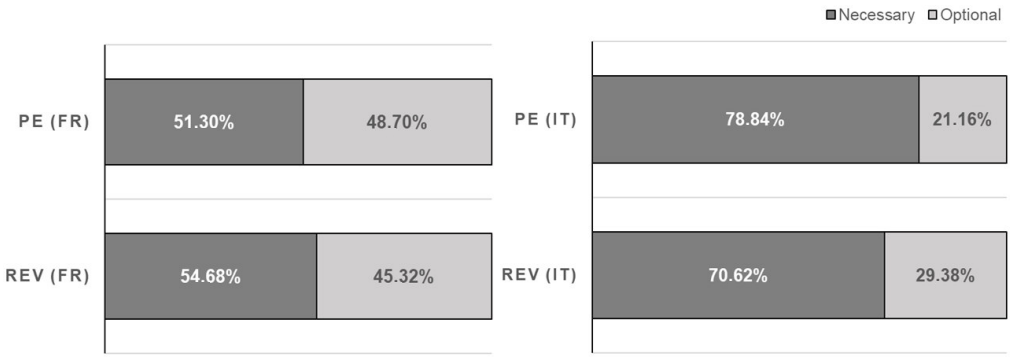



FIGURE 6.21 – Percentage of necessary and optional edits per task, in each target language subset.

In the Italian subset, a striking majority of edits are considered necessary, with 78.84% in PE assignments and 70.62% in REV assignments falling into this category. On the other hand, the French subset exhibits a lower percentage of necessary edits, with only slightly over half of the edits being classified as necessary (51.30% in PE and 54.68% in REV). This implies a difference in the distribution of "optional" edits between PE and REV assignments in the two subsets. French-speaking evaluators rated as optional a slightly higher number of edits in PE assignments (48.70%) compared to REV assignments (45.32%). Conversely,

Italian-speaking evaluators rated as optional a higher number of edits in REV assignments (29.86%) compared to PE (21.16%). Figure 6.22 reports the percentages of necessary and optional edits in PE and in REV, broken down per linguistic macro-dimension in which the edit has been previously classified.

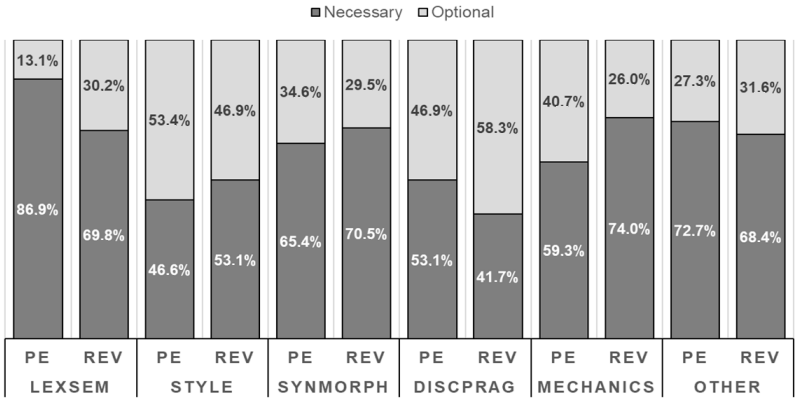



FIGURE 6.22 – Percentage of necessary and optional edits in PE and in REV, broken down per macro linguistic dimension affected.

The data reveal that in four out of six macro-dimensions, the majority of edits are classified as necessary in both PE and REV assignments. This is particularly evident in the *Lexis and semantics* dimension, where a substantial percentage of edits are deemed necessary, with a higher percentage in PE assignments (86.9%) compared to REV assignments (69.8%). In the *Syntax and morphology* dimension, the percentage of necessary edits is slightly higher for REV assignments (70.5%) than for PE ones (65.4%). The *Mechanics* dimension also displays a similar pattern, with a higher percentage of edits in REV assignments considered necessary (74%) compared to edits in PE assignments (59.3%). On the other hand, two macro-dimensions stand out for having a higher proportion of optional edits compared to necessary edits. Specifically, in the *Style* dimension, PE assignments show a greater percentage of optional edits (53.4%) than REV assignments (46.9%). Conversely, in the *Discourse and pragmatics* dimension, REV assignments have a higher percentage of optional edits (58.3%) compared to PE assignments (46.9%). Figure 6.23 details the results for individual target languages.

The findings reveal an interesting contrast between the Italian and French subsets in terms of the necessity of edits across different linguistic dimensions. In the Italian subset, the majority of edits are considered necessary in all linguistic dimensions. This trend holds true for both PE and REV tasks. On the other hand, the French subset shows a different pattern, with optional edits surpassing necessary edits in certain dimensions. Specifically, in the *Style* and *Discourse and pragmatics* dimensions, optional edits are more prevalent than necessary edits in both PE and REV tasks. Additionally, in the *Syntax and morphology* dimension, this pattern is observed only in PE assignments, while in the residual dimension (*Other*), it is observed only in REV assignments.

This difference between the Italian and French subsets has implications for the results of edits' necessity in both tasks, where the higher percentage of optional edits in the French

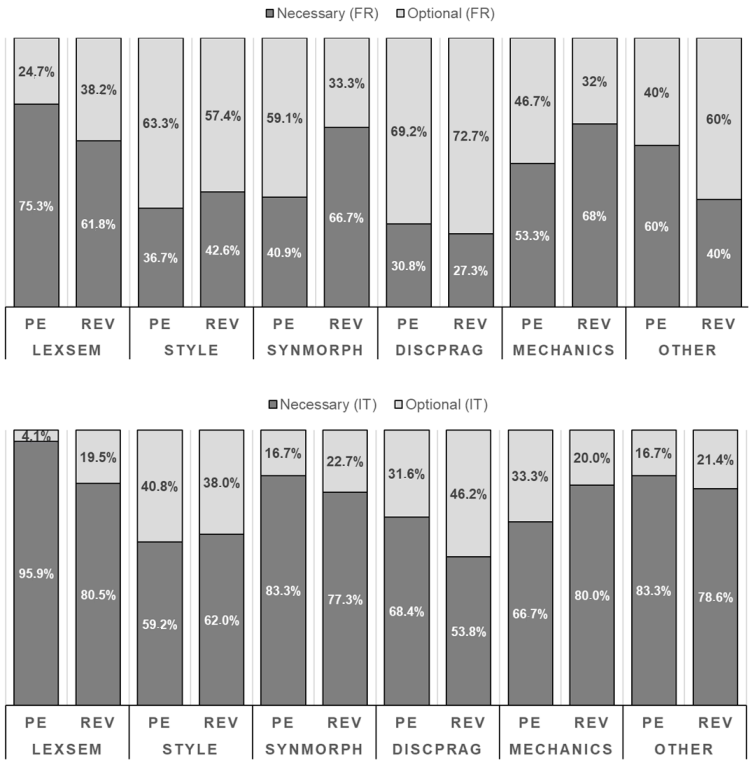



FIGURE 6.23 – Percentage of necessary and optional edits in PE and in REV, broken down per macro linguistic dimension affected and per target language (upper side: French subset; lower side: Italian subset).

subset mitigates the overall results. However, considering that stylistic edits are often considered optional (Nitzke and Gros, 2021) and given that the majority of edits in the REV subset primarily involve stylistic modifications, one would have expected a greater proportion of optional changes within the REV category for both target languages. Previous studies examining the necessity of edits performed by professional translators have yielded mixed results. Our findings partially align with those of de Almeida (2013), who found that the majority of edits in a PE task were considered essential. The author reported that essential changes primarily focused on language-related aspects such as grammar, with mistranslations as the second most frequently implemented change, followed by accuracy and style. Notably, changes related to lexical choices constituted a minimal percentage (0.2%) of all implemented edits. When considering preferential changes, language remained the most prevalent category, followed by lexical choice and style. However, direct comparisons between our findings and de Almeida's ones are complicated by the use of different taxonomies to categorise the edits. Our findings also align with those of Nitzke and Gros (2021), who observed that preferential edits in PE and REV tasks were primarily stylistic or lexical in nature. We identified optional edits in additional dimensions, such as *Discourse and pragmatics*, which have not been previously examined in similar studies.

In contrast to our findings, Parra Escartín et al. (2017) reported that a professional proofreader primarily implemented preferential changes when proofreading texts post-edited by non-professional translators. Bundgaard's study (2017a) also confirmed that professional

translators frequently make preferential changes during revision tasks, depending on the text type and on what is considered essential (e.g., the insertion of non-breaking spaces). Compared to these findings, we can conclude that the professionals in our study (at least Italian-speaking ones) professional translators in our sample demonstrated a successful ability to minimise the introduction of unnecessary optional modifications. While optional changes may potentially contribute to less time-efficient workflows and require additional editing effort from the linguists, they can play a pivotal role in enhancing the quality of final texts. Thus, it is important not to overlook their value.

### 6.5.6 Effectiveness

Table 6.12 shows the average effectiveness scores obtained by edits in the PE and the REV subsets. We recall that such score could range between -2, indicating that the edit degraded the pre-translation, and +2, indicating that the edit improved the pre-translation. A score of zero could indicate either a *truly preferential edit* or an *ineffective edit*.

| Task | | Effectiveness score (Mean) |
|---|---|---|
| **PE** | *de-fr* | 1.40 |
| | *de-it* | 1.72 |
| | ***all*** | **1.57** |
| **REV** | *de-fr* | 1.38 |
| | *de-it* | 1.55 |
| | ***all*** | **1.46** |

TABLE 6.12 – Mean effectiveness scores of edits in the PE and REV subsets.

We notice that both post-editors and revisers consistently improved the quality of the pre-translations, with effectiveness scores close to 2 for both subsets. This outcome aligns with expectations, considering the expertise of CILS linguists in enhancing the quality of pre-translated texts. However, it is noteworthy that the improvement was slightly higher in PE (1.57) compared to REV (1.46) assignments. A Mann-Whitney U test confirmed that the increase in edits' effectiveness in PE assignments, compared to REV assignments, is statistically significant ($p$=0.026). This finding suggests that the impact of edits is more noticeable in PE tasks, potentially due to the initially lower quality of the machine-translated output compared to human-translated texts.

The comparison of average scores obtained by the two target language subsets in each task further highlights the differences in edit effectiveness. Specifically, the Mann-Whitney U test conducted on the Italian subset reveals statistically significant differences in effectiveness scores ($p$=0.005), indicating that the impact of edits is more pronounced in assignments for Italian. No statistical significance is found between the effectiveness scores of the two tasks in the French subset ($p$=0.650). It is worth noting that the evaluation of edits in PE assignments yielded higher kappa values for the Italian subset ($\kappa$=0.409) compared to the French subset ($\kappa$=0.257). This implies that assessing the effectiveness of edits was easier and more consistent for Italian raters than for their French counterparts. This finding supports

previous analyses, indicating that the customised machine translation engine used for Italian provides lower translation quality compared to the one used for French.

| | Effectiveness score (Mean) | |
|---|---|---|
| | **PE** | **REV** |
| **Necessary** | 1.86 | 1.80 |
| **Optional** | 1.01 | 0.91 |

TABLE 6.13 – Mean effectiveness scores of edits in the PE and REV subsets, broken down per edits' necessity.

We examined the difference in effectiveness scores between necessary and optional edits (Table 6.13). The results reveal that the difference is statistically significant for both tasks, with $p$=0.000. This suggests that in both PE and REV tasks, necessary edits play a crucial role in enhancing the quality of the translation. On the other hand, optional edits have a less substantial impact, indicating that they may be less critical for achieving a satisfactory translation outcome. Additional Mann-Whitney U tests conducted on individual target language subsets further confirmed this observed difference. However, it is important to note that the mean effectiveness scores obtained by necessary and optional edits are not significantly different between the PE and REV tasks ($p$>0.05).

Let us now consider edits' effectiveness across linguistic dimensions. The hexagonal chart in Figure 6.24 displays the average effectiveness scores obtained by edits in the PE and REV subsets, across the six linguistic macro-dimensions previously defined. Data positioned closer to the external borders of the chart correspond to higher effectiveness scores, meaning that edits were more impactful in improving the quality of the pre-translated text. Conversely, data points closer to the centre of the chart indicate lower effectiveness scores, suggesting that the edits had less impact or were deemed less successful in enhancing the pre-translated text's quality.

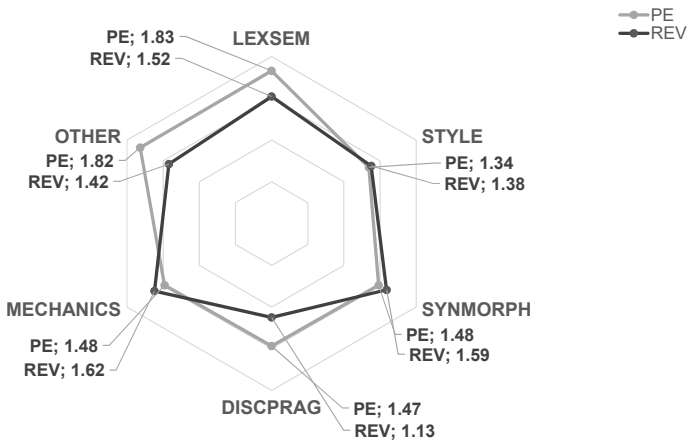



FIGURE 6.24 – Mean effectiveness scores of edits in PE and REV subsets, per linguistic dimension (macro). The proximity of data points to the borders indicates the magnitude of the effectiveness score.

In PE tasks, edits in the *Lexis and semantics* dimension achieve the highest mean effectiveness score (1.83), closely followed by edits classified under the *Other* dimension (1.82). Conversely, in REV tasks, edits in the *Mechanics* dimension obtain the highest mean effectiveness score (1.62), with edits in the *Syntax and morphology* dimension following closely behind (1.59). Overall, edits in PE assignments attain higher mean effectiveness scores compared to edits in REV assignments, except for the *Syntax and morphology* and the *Mechanics* dimensions. For the *Style* dimension, edits in REV assignments obtain a slightly higher mean effectiveness score (1.38 in REV versus 1.34 in PE). However, it is important to mention that Mann-Whitney U tests have shown statistically significant differences in mean effectiveness scores between PE and REV tasks only for the edits categorised under the *Lexis and semantics* dimension ($p<0.001$). Additionally, when examining optional and necessary lexico-semantic edits separately, we observed that a statistically significant difference between PE and REV is evident only for optional edits ($p=0.008$) and not for necessary edits ($p=0.075$). This suggests that while necessary edits in both PE and REV tasks improve the translation similarly across linguistic macro-dimensions, optional edits falling within the *Lexis and semantics* category significantly enhance machine-pre-translated texts compared to human-pre-translated texts. Furthermore, we compared mean effectiveness scores for edits in both PE and REV tasks across the various linguistic micro-dimensions. Among the linguistic micro-dimensions, statistically significant differences ($p<0.05$) in effectiveness scores between PE and REV were observed only for *Content distortion* and *Sentence splitting*, where edits in PE assignments demonstrated higher effectiveness scores in these specific micro-dimensions.

Finally, we focused on truly preferential edits, which are optional edits that ultimately had no effect on the final text (see Section 6.4.5). To identify these, we specifically considered cases where both raters agreed on the lack of effectiveness of the edit, resulting in a score of zero [51]. These edits accounted for 5.3% of the total number of edits in the PE dataset and 6% in the REV dataset (Table 6.14) and pertained mainly to the *Style* dimension.
While the occurrence of truly preferential edits is relatively low overall, it is noteworthy that the French subset exhibited nearly twice as many of these edits compared to the Italian subset. This observation suggests a potential tendency among French-speaking linguists to over-edit, irrespective of the task they are asked to perform. Since the profile of linguists in the two language teams is quite similar, this difference could potentially be attributed to the higher quality output provided by the MT system used for French. Consequently, French-speaking linguists may be more inclined to make truly preferential edits in an effort to imprint their unique touch on the text. However, further investigation is warranted to explore this phenomenon in greater depth.

51. We opted to exclude cases from our dataset where there was substantial disagreement between the two raters regarding the evaluation of edits. Specifically, we excluded instances where one rater considered an edit to be an improvement while the other rater perceived the same edit as degrading the pre-translation. It is worth noting that such cases were rare in our dataset. In the French subset, we identified 9 instances (5 in PE and 4 in REV), and in the Italian subset, we found 6 instances (5 in PE and 1 in REV) of edits that met this criterion.

| Task | | Truly preferential edits |
|---|---|---|
| **PE** | *de-fr* | 16 |
| | *de-it* | 9 |
| | ***all*** | **25 (5.3%)** |
| **REV** | *de-fr* | 16 |
| | *de-it* | 8 |
| | ***all*** | **24 (6%)** |

TABLE 6.14 – Number of truly preferential edits and respective percentage on the total number of edits in the subset, per task and per language combination.

## 6.6 Additional considerations

The findings presented in the previous sections have contributed to our understanding of the specific editing dynamics within PE and REV workflows at our CILS partner, highlighting their respective demands for intervention. Our study revealed that revisers made fewer modifications than post-editors during their assignments. Furthermore, these modifications tended to be narrower in scope, focusing primarily on refining existing translations with minimal intervention, particularly on stylistic and mechanical aspects. It is worth noting that our study targeted authentic texts from a professional workflow, where revisers worked on (pre)translations predominantly produced by professional and experienced translators. These translators, being skilled in their craft, often outperformed the customised NMT engine. Different results might have been obtained if we had included only revision assignments of inexperienced or non-professional translators, who are more likely to introduce semantic and terminological errors into their translations (as observed in the experiment detailed in Chapter 5).

Given that the NMT system still produces a significant number of mistranslations, the primary focus during the PE task often revolves around addressing semantic issues. However, the need to deliver a translation that meets the same quality standards as one produced in a traditional workflow (i.e., translation and revision) has prompted CILS linguists to perform a substantial number of edits targeting stylistic aspects during the PE tasks. Overall, despite these differences, we find that the disparities in the amount of editing performed by CILS revisers and post-editors are not substantial. Our analysis indicates that the wide majority of PE sentences are typically modified up to 50%, while REV sentences are modified up to 25%, suggesting a similarity in the overall magnitude of modifications required in both tasks. Furthermore, the parts of speech affected by edits in the two tasks are predominantly the same. These findings provide evidence of the advancements made by the NMT system, which have made the PE activity more akin to the revision task when compared to previous generations of MT systems (see Vasconcellos, 1987 and Wagner, 1985). Unfortunately, we were unable to verify this claim directly as the CILS partner did not use other MT systems for comparison.

The distribution of editing actions performed in PE and REV, as portrayed by the HER

metric, revealed minimal differences between the two tasks. Similar distributions of editing actions have been reported in previous studies using the TER metric. This suggests that, computationally speaking, the tasks of PE and REV may be almost indistinguishable. To explore this further, future studies could investigate the distribution of editing actions in PE and REV tasks across different production contexts and employing different metrics. Additionally, it would be interesting to manually identify logical edits and compare their distribution with that proposed by the HER metric, to ascertain whether the latter aligns with the reality.

Another aspect deserving further discussion is the evaluation of edits' necessity. The majority of edits were deemed necessary in both tasks, although differences were observed between language pairs. We contend that necessary edits are generally easier to assess compared to optional ones. Indeed, we found that the prevalence of necessary edits also corresponds to a higher level of inter-rater agreement. In our study, the three Italian-speaking raters exhibited a higher percentage of agreement compared to their French-speaking counterparts. Furthermore, for the Italian subset in particular, Light's kappa values were higher for PE edits than for REV edits, suggesting that in this target language judging the necessity of edits in PE assignments appears to be easier than in revision assignments.

Although less common, optional edits in our dataset apply to both PE and revision tasks. Mellinger and Shreve (2016) reported that preferential edits are also common when editing TM matches. In their study, participants made several preferential changes to TM exact matches (i.e., that did not require any modification), particularly at the lexical or syntactic level. The authors interpret this phenomenon as the result of participants having competing translation versions in mind, leading them to modify the text despite its correctness. As noted by Nitzke and Gros (2021, p. 22), a similar situation arises in both PE and revision tasks, where post-editors and revisers must confront their own translation ideas with a translation generated by a machine or another person. We agree with this observation and further propose that these types of edits may serve as a means for linguists to assert their ownership over the text and exert greater control over the final outcome.

Optional modifications hold a distinct status in PE and revision tasks. As highlighted by Nitzke and Gros (2021, p. 32), "some colleagues might feel patronised if their translations are revised according to personal preferences during the revision task". While the relationship with the translator holds significant importance in the revision task (see Riondel, 2021a), optional edits in PE tasks within CILS workflows are deemed less critical and are even welcomed, as long as they enhance the text without excessively sacrificing productivity. In the case of PE, we argue that suppressing personal preferences could undermine the satisfaction derived from the profession and perpetuate the negative reputation and perception of the task (Läubli and Orrego-Carmona, 2017; Nunes Vieira and Alonso, 2018; Sakamoto, 2019). In an era where the improved quality of MT engines may not necessarily make the task easier but potentially more enjoyable, it is important to acknowledge the value of personal preferences in PE tasks.

Finally, we would like to draw attention to an important lesson learned during the preparation of the human evaluation of edits' necessity with CILS linguists (as detailed in Section

6.4.5). The feedback provided by the linguists who were unable to complete the evaluation raised an important consideration. When linguists are asked to judge whether their colleagues' modifications are optional, there is a risk of triggering deep-seated concerns about job security and the effectiveness of one's own work. This fear can potentially introduce biases into the evaluation process, as linguists may feel inclined to label edits as "necessary" to avoid undervaluing their colleagues' work and, in turn, jeopardizing their own job.

To address this issue, it is worth exploring alternative wording choices that can minimise the negative implications associated with terms like "optional" or "preferential" which convey a sense that the work of professional translators is meaningless or disposable. One possibility is to use a binary distinction between "necessary" and "less necessary" instead. This approach focuses on the essentiality of the edits rather than emphasising their optional nature, which may help alleviate concerns about job security and mitigate potential bias. It is crucial to clarify and communicate the purpose of the evaluation process in advance to linguists involved in similar studies. By ensuring a transparent and supportive environment, researchers can help alleviate anxieties and foster a more accurate assessment of edits' necessity in future studies.

We acknowledge that the present study has several limitations that should be taken into consideration. Firstly, it only focuses on texts from a specific CILS, which may limit the generalisability of the findings to other contexts. Secondly, the qualitative analyses were conducted on a subset of data, representing less than 10% of the entire corpus, and specifically on a subset of lightly-modified sentences. This sampling approach may not fully capture the complexity and diversity of all types of edits in the corpus. Additionally, the qualitative evaluation relied on ratings from only three evaluators per target language, which introduces a level of subjectivity. While efforts were made to minimise this aspect, it is important to acknowledge that different raters could have reached different conclusions. Another limitation is that the edit annotation taxonomy proposed in this study was specifically tailored for the use with the CILS partner and focused primarily on lightly-modified sentences. However, the taxonomy can be adapted and modified to suit different contexts and to be applied with texts that require more extensive corrections.

The present study suggests several avenues for future research. Replicating the study with other corpora of authentic texts from different CILS could contribute to the generalisability of the findings. Additionally, analysing the editing of TM matches, which was not considered in this study, could allow for a comparison with another type of pre-translated text. In further studies, we would like to replicate the qualitative analyses using an expanded dataset including sentences that have undergone a moderate amount of corrections (HER score of up to 50). This approach would let us investigate the potential differences between lightly and moderately-modified sentences. Lastly, integrating measures of technical effort from Translation Process Research, such as keyloggers, could provide insights into the "invisible editing" that occurs during both PE and revision tasks. This would allow for a more comprehensive analysis that goes beyond the static comparison of the products.

## 6.7 Summary of the chapter

In this chapter, we presented a corpus-based study on editing in PE and revision. The term *editing* refers to all the changes (*edits*) performed on a pre-translated text, both from a quantitative and a qualitative perspective. Hence, the study aimed to investigate the quantity, type, and nature of modifications made to pre-translated texts during PE and revision, as well as to evaluate the necessity and effectiveness of these edits. The data used for the analysis were collected from a Swiss CILS in two language combinations, namely German to French and German to Italian. Two corpora were compiled: one consisting of authentic PE assignments (4,089 triples, i.e., source segment, pre-translation, final version) and another of authentic revision assignments (4,933 triples) completed by CILS linguists.

The data analysis methods included both quantitative and qualitative approaches, drawing from previous research in MTPE and Translation Studies. Six dependent variables were analysed to address our research question (**RQ4**: *When PE and revision are carried out under the same conditions, how do PE and revision compare in terms of modifications (edits) made to pre-translated texts?*). First, the amount of modifications made to machine-pre-translated and human-pre-translated texts, which has been measured using the HER metric (do Carmo, 2021b). Second, the distribution of editing actions performed during PE and revision, namely insertions, deletions, substitutions and word shifts, as recorded by the HER metric. Third, the parts of speech affected by the edits in both tasks. Fourth, the linguistic dimensions affected by the edits, using a customised taxonomy including six macro-dimensions and a total of 24 micro-dimensions. Our fifth dependent variable is the necessity of the edit, as assessed by three evaluators per target language. Finally, in the same human evaluation, we also analysed the effectiveness of each edit, namely whether the edit had improved, degraded or simply had no effect on the quality of the pre-translation.

While the quantitative analyses (first to third variable) involved analysing the whole corpora, the qualitative analyses (fourth to sixth variable) were calculated on a subset of data consisting of 300 triples in each target language subset, evenly distributed between PE and REV. The empirical findings provided interesting insights. Firstly, it was observed that CILS linguists performed more editing during PE tasks compared to revision tasks, as indicated by higher HER scores for PE assignments. There were differences between language pairs within each task, suggesting variations in the quality of the MT output. We also observed that a significant portion of sentences in the REV corpus underwent only slight modifications (HER scores up to 25), while major restructuring was less common. In the PE corpus, the majority of sentences underwent some more extensive changes (HER scores up to 50).

In terms of editing actions, substitutions emerged as the most frequent edit type in both PE and revision tasks, followed by insertions and deletions. Shifts were less common in the revision corpus compared to the PE corpus. Although the percentage breakdown of editing actions was similar between the two tasks, a statistically significant difference was found, indicating that the task has an influence on the distribution of editing actions. In particular, revision assignments include a higher percentage of deletions compared to PE assignments.

The analysis of edited parts of speech showed that nouns and prepositions were the most frequently modified categories in both PE and revision tasks. True insertions, manually annotated, revealed that prepositions, punctuation marks, nouns, and determiners were the most commonly inserted parts of speech in both tasks. While the distribution of edits on parts of speech appeared similar between the two tasks, statistical analysis highlighted significant differences, indicating an association between the task and the parts of speech affected by edits. Regarding linguistic dimensions, edits in the REV corpus were found to belong predominantly to the *Style* category, while lexico-semantic edits are less common. On the other hand, edits in the PE corpus were almost equally distributed between the *Lexis and semantics* and the *Style* category. When examining the results for different target language subsets, variations were again observed. In the French subset, stylistic edits constituted the largest proportion of edits in both PE and revision assignments. In the Italian subset, stylistic edits were the largest linguistic dimension among revisers, whereas post-editors focused mainly on lexico-semantic edits. Micro-dimension analysis revealed that the focus in PE assignments is on addressing mistranslations, while revision assignments mainly involve providing synonyms or rephrasing.

The majority of edits performed in both PE and revision tasks were deemed necessary, although revisers tended to make slightly more optional modifications than post-editors. There were notable differences between the Italian and French subsets, with a higher proportion of necessary edits in the Italian subset for both tasks. In contrast, in the French subset only slightly over half of the edits were classified as necessary. Edits in four out of six dimensions were predominantly classified as necessary, particularly in *Lexis and semantics*. Optional edits were more common in the *Style* and *Discourse and pragmatics* dimensions, with variations observed between the Italian and French subsets.

Both post-editors and revisers consistently improved the quality of the pre-translations, with slightly higher improvement observed in PE assignments. Statistical tests confirmed the significant impact of edits in PE tasks, probably due to the lower quality of the MT output compared to human-pre-translated texts. There were notable differences in effectiveness between the Italian and French subsets, with edits having a more pronounced impact in Italian assignments. We also observed that necessary edits were in general more effective, while optional edits had a lesser impact on the pre-translations.

# Chapter 7

# Conclusions

This thesis explored how PE and revision tasks compare in the professional context of Swiss corporate in-house language services. This concluding chapter includes an overview of our research journey, a synthesis of the main findings and their implications for pedagogy and professional practice. The chapter also highlights the contributions of our work to the field of Translation Studies and discusses the limitations of our overall approach. Lastly, it suggests possible avenues for further research on the topic.

## 7.1 An investigation of PE and revision tasks in corporate in-house language services

This thesis situated itself within the context of workplace empirical research, investigating two professional activities within a real-world professional context, namely the one of Swiss corporate in-house language services (CILS). In the introductory chapter (Chapter 1), we presented our object of study and motivated the choice of our topic. Our research was rooted in a desire to better understand the role of PE and revision in professional translation workflows, as well as their relationship from the perspective of CILS linguists. A review of the literature that focused on the two activities (in the same investigation) highlighted a number of research gaps. In particular, we noted a general lack of ecologically valid empirical studies that investigate the relationship between PE and revision tasks following the introduction of neural MT in professional workflows.

We therefore defined two main goals: first, gathering information on how PE and revision tasks are organised in CILS workflows, and – more in general – on the use of MT in Swiss CILS (Chapter 2). Second, investigating PE and revision tasks in CILS workflows, adopting a contrastive perspective. Goal 1 was achieved through a questionnaire sent to CILS' directors and project managers, which allowed us to delimit the context of our research and get to know better our cohort of participants. One important finding from this first study was that in the majority of surveyed CILS, PE is performed on fully pre-translated texts, typically as the last step in the workflow, meaning that post-edited texts remain most often unrevised.

To attain Goal 2, we employed a multimethod empirical approach, combining different research strategies and data collection methods to explore the similarities and differences between PE and revision in the context of Swiss CILS. We defined four key aspects which served as viewpoints to compare PE and revision, namely linguists' PE and revision practices (Chapter 3), satisfaction (Chapter 4), productivity (Chapter 5) and edits made to pre-translated texts (Chapter 6). Each aspect was detailed in a separate chapter and contextualised with a specific literature review focusing on how that aspect has been investigated in research on PE and revision, as well as in studies on both tasks (whenever such studies were available). Practices and satisfaction were investigated through a questionnaire sent to CILS linguists who carried out both PE and revision in their daily workflows. Productivity and edits made to pre-translated texts were explored in a specific context – the one of a CILS partner – through a field experiment and a product analysis, respectively. The field experiment was conducted with six CILS linguists and three language combinations, namely German into French, Italian and English. The product analysis involved collecting two corpora of authentic PE and revision assignments carried out at the CILS partner in two language combinations, namely German into French and Italian.
The main findings of these investigations are summarised in the next section.

### 7.1.1 Research questions and main findings

Goal 2 was achieved by providing answers to four specific research questions.

**RQ1**: *In the professional workflow of Swiss CILS, how do linguists' PE and revision practices differ?*

Results from the questionnaire on linguists' PE and revision practices revealed that the majority of CILS linguists claim to act differently depending on whether they work with human-translated or machine-translated texts. For instance, the majority of linguists follow specific parameters during revision, but such guidelines are not always available for PE tasks. Linguists systematically verify the correctness of terminology during PE tasks, but this happens more rarely during revision assignments. The different approach towards the two tasks is motivated by a lower trust in MT than in human colleagues. When it comes to reading strategies, however, most linguists reported using the same strategies for both revision and PE. These involved proceeding segment by segment, starting either with the source or the target segment, which emphasises the impact of text display within a CAT tool on PE and revision processes. The findings also uncovered that post-editors tend to employ consistently the same reading strategies, while revisers more often adapt them depending on text type, time constraints, and the translator who performed the translation.

When revising texts with different origins (i.e. human-translated or post-edited), the majority of respondents claimed not to change their strategies. Nonetheless, some respondents highlighted the need to pay higher attention to cohesion and terminology consistency when revising post-edited texts. Following the introduction of MT in the workflow, linguists did not seem to perceive an influence of MT on their usual revision strategies. It is worth

noting, however, that many respondents misunderstood this question and commented again on the strategies put in place to go over human-translated and machine-translated texts – thereby showing some linguistic bias towards viewing PE as the revision of MT output.

**RQ2**: *In the professional workflow of Swiss CILS, how do PE and revision compare in terms of task satisfaction for the linguist who performs these tasks?*

Satisfaction was investigated through various facets, including *self-efficacy, creativity, task complexity, stimulating nature of the task, control, ability utilisation, task identity, work-scheduling autonomy* and *task variety*, as well as *overall satisfaction* (*task enjoyment*). Translation tasks were added to the comparison of PE and revision to gather a more complete picture of linguists' satisfaction with their main job tasks.

Our findings revealed a hierarchy among these three tasks in terms of linguists' satisfaction. We observed a general tendency to value translation tasks more highly than revision tasks, and, in turn, revision tasks are valued more highly than PE tasks across almost all core concepts investigated. We found that PE is considered less fulfilling and slightly less demanding in terms of required skills and knowledge than revision. However, linguists also reported that PE, more than revision, enables them to create new content and exert control over the text's final quality.

Furthermore, linguists who received proper PE training reported higher satisfaction levels with PE tasks than those who did not receive such training. Conversely, revision training significantly enhances feelings of self-efficacy in revision tasks, but it is less effective in influencing overall satisfaction levels in the same tasks. Interestingly, the mode of using or displaying MT suggestions in the CAT tool does not seem to affect the levels of overall satisfaction with the PE task or the perception of the task as more or less creative. Contrary to our expectations, post-editors who work on fully-pre-translated texts reported a higher level of control over the text's final quality than those who display MT suggestions in a separate window.

According to our respondents, positive aspects of working with MT include productivity gains and suggestions to initiate the translation process. Some linguists even perceive bad MT quality as a positive aspect since it signals the need for human intervention. Negative aspects of working with MT include reduced creativity, lack of consistent terminology, and the fact that NMT does not take into account a broader context than the sentence.

Revising human translations is appreciated for its enriching nature, the opportunity to improve one's skills, and the valuable human interaction between revisers and translators. However, the latter also constitutes a source of issues for the reviser since the most-cited negative aspects of revising human translations include dealing with low-quality translations, unskilled translators, and possible conflicts during discussions with translators.

**RQ3**: *When PE and revision tasks are carried out under the same conditions, how does linguists' productivity compare in these tasks?*

Productivity was defined in terms of temporal effort and adequacy of the final product. Participants generally displayed a higher processing speed during revision tasks when contrasted with PE tasks. Nonetheless, a noteworthy trend emerged: the gap in processing speed between revision and PE tasks diminished when dealing with customised NMT output.

Quality assessments revealed that a higher percentage of sentences originating from MT and TM fuzzy matches needed further correction when compared to human pre-translated segments. This discrepancy was particularly pronounced when using SMT and exhibited marginal improvement with NMT. The reasons behind this trend remain open to interpretation. While a quality improvement of the NMT output is a possibility, external factors could also have contributed to the observed change.

In line with prior studies on productivity, our study highlighted individual differences among participants, both in terms of processing speed and the quality of the final translations. We also noted the propensity of some participants to overlook exact matches. This tendency might be attributed to their trust in TM resources and established industry practices that discourage modifications to this type of TM suggestion.

**RQ4**: *When PE and revision tasks are carried out under the same conditions, how do PE and revision compare in terms of edits made to pre-translated texts?*

Overall, we found that PE assignments required more editing compared to revision assignments. Furthermore, human-pre-translated sentences required modifications that were narrower in scope compared to sentences in the PE subcorpus, which underwent more extensive modifications. Differences between language pairs emerged within each subcorpus, suggesting, in the case of PE, a possible discrepancy in the quality provided by the customised NMT engines for the two target languages considered.

Regarding the distribution of editing actions in both tasks, we found substitutions to be the most frequent operation, followed by insertions and deletions. The difference in the distribution of editing actions between PE and revision tasks was found to be statistically significant, with the revision subcorpus exhibiting a higher number of deletions compared to the PE one. Parts of speech most often affected by edits in both tasks were nouns, prepositions, verbs and determiners, but we also observed a statistically significant difference in this distribution, with revision assignments containing a higher percentage of edits on punctuation than PE assignments.

In the subset of sentences considered for further analysis, we found that PE tasks predominantly involved addressing content distortions and lexical misselections, as well as providing synonyms or rephrasing lexical content (*Style* dimension). Edits performed during revision tasks belong mainly to the *Style* dimension, while lexico-semantic edits are less common.

The majority of edits in the subset for both PE and revision tasks were categorised as necessary, with a slightly higher proportion of such edits on post-edited sentences compared to revised ones. In two linguistic dimensions, optional edits outperform necessary ones: this applies to the *Style* dimension for post-edited sentences and to the *Discourse and pragmatics* dimension for revised sentences. Notable differences in the perception of edits' necessity were observed between target language subsets. In the Italian subset, a vast majority of edits were considered necessary, while the French subset had a higher proportion of optional edits.

Edits in both PE and revision tasks improved the quality of translations, with slightly higher effectiveness scores in PE assignments compared to revision ones. Italian-speaking evaluators found it easier to assess edit effectiveness than their French counterparts, potentially due to differences in NMT quality for the two engines. We observed that necessary edits were, in general, more effective, while optional edits had a lesser impact on the pre-translations. A small percentage of edits – referred to as "truly preferential" – had no impact on the final text's quality. French-speaking linguists exhibited a higher tendency to make these edits.

### 7.1.2 Implications for professional practice

Our findings have direct implications for professional practice, particularly for language service providers seeking to optimise their workflows and for linguists who wish to refine their PE and revision practices for increased efficiency.

First and foremost, we have shed light on how PE and revision tasks are structured within the workflows of different Swiss CILS, providing in-depth information on practical aspects that have previously received little attention (such as how MT is integrated into CAT tools). These insights can assist decision-makers at CILS who are still uncertain about whether or how to implement MTPE in their workflows. Additionally, they can benefit decision-makers at CILS that already use MT in optimising their processes and workflows, drawing from similar practices in related contexts.

Similarly, linguists can benefit from our findings to refine their approaches towards pre-translated texts. They can consider testing different reading strategies during revision and PE and reflect on their revision procedures when dealing with translated or post-edited texts.

Our findings have revealed that PE tasks cannot compare to revision tasks when it comes to the time required to complete these tasks. Furthermore, assessments of the final product suggested that omitting the revision step in PE workflows could pose risks. Therefore, the results of the productivity study should serve as an incentive for CILS to adapt project requirements and timelines to better match the actual demands of the PE and revision tasks. Language professionals can also derive benefits from closely monitoring their productivity during PE and revision tasks, as this data-driven approach can assist them in optimising their workflows and asking for adequate compensation.

Finally, our findings have demonstrated that PE training can significantly improve linguists' enjoyment of PE tasks. We strongly recommend that CILS decision-makers consider implementing regular, task-specific training sessions for their linguists to help them gain

confidence in their abilities to perform both PE and revision tasks effectively. It is crucial to equip individuals with the necessary skills and knowledge for each role, as our research has shown that PE and revision tasks involve distinct requirements. Additionally, we emphasise the importance of actively involving linguists in the MT implementation process, as this approach contributes to the successful integration of MT into the workflow.

By exploring linguists' perceptions of their job tasks, our research offered insights into factors that impact professionals' overall job satisfaction. This knowledge can guide in-house language services in creating more fulfilling work environments. In this regard, we agree with O'Brien and Duarte (2015, p. 92), who stated that in a project of MT integration in a language service,

> the place to start isn't on the MT, but the internal content, the tools and the people: the three main pillars we need to 'shape' in order to get acceptance for MT and effect whatever systems are needed.

.

### 7.1.3 Implications for translation pedagogy

As noted by Risku, Rogl, and Milošević (2020, p. 52):

> Workplace research enhances an academic institution's knowledge of current practices and trends and enables it to take them into account in their teaching programmes, in our case translator training. In this way, [academic institutions] are better equipped to produce well-educated, well-informed graduates who are able to deal with the challenges they will be confronted with in their future work contexts and thus to support industry in its recruitment processes.

Our research on PE and revision practices in Swiss CILS does precisely this, offering valuable insights for translation pedagogy. Our study highlighted the unique characteristics of a specific cohort of professional translators and the role of translation technology in their workflows. We gathered data on which MT systems and CAT tools are most frequently used, as well as the most common integration methods for MT within CAT tools. Since the majority of CILS linguists in our survey use batch pre-translation in PE assignments, instructors could propose PE exercises using this format. We note, however, that it is crucial to expose students to other integration methods, too, and encourage a critical reflection on their advantages and disadvantages. Our data also showed that PE texts are most often left unrevised. This emphasises the need to integrate self-revision exercises into PE training. This point is particularly relevant considering the persistence of errors in both PE and revision tasks, as our productivity study has shown.

Findings related to the modifications made to pre-translated texts have revealed the main tendencies of what revisers and post-editors most often (need to) look at in their respective tasks. This data directly informs pedagogy and could guide the development of targeted test suites to train students in identifying specific issues in PE and revision assignments.

Our evaluation of edits' necessity and effectiveness has shown that the latter was less visible in revision assignments. To address this, instructors can propose specific exercises that

encourage students to critically evaluate the effectiveness of edits. One simple activity could be to present students with before-and-after versions of revised texts (without telling them which is which) and ask them to identify the improved version, if any. If no improvement is visible, the effectiveness of that edit can be considered reduced.

Translation scholars have often recommended introducing PE and revision as two separate activities at a later stage in the translation curriculum (O'Brien, 2002; Guerberof and Moorkens, 2019; Mossop, 2020; Konttinen, Salmi, and Koponen, 2021), once students have gained some translation competence. While we support this staged approach, we believe it is important to maintain a complementary contrastive approach when examining PE and revision, to clearly stress both similarities and crucial differences between the two tasks. We recommend dedicating at least one session to contrastive analyses of PE and revision, in which instructors could introduce and discuss the key aspects explored in this thesis. One important aspect that would benefit from such a perspective is the choice of the specific reading strategy to employ. Our survey data suggested that linguists tend to use similar approaches when dealing with PE and revision assignments. Instructors can leverage this information by encouraging students to experiment with different strategies and then initiating a critical reflection on the implications of using each strategy. While empirical research on the most effective strategy for each task is limited, this type of pedagogical activity can empower students to develop their own informed reading strategies for each task.

Finally, our findings highlighted the role of MTPE training in fostering positive perceptions of the PE task. As MT integration into professional workflows continues to grow, graduates are increasingly likely to work with machine-pre-translated texts. Therefore, it is important to enhance their level of satisfaction towards PE. We contend that this process begins in their academic training. Instructors can play a pivotal role in shaping students' perceptions of the PE activity, dismantling the hierarchy that positions it below translation and revision (as our results revealed). This can be achieved by emphasising the positive aspects of working with MT and acknowledging it as a task requiring specific skills and knowledge. Furthermore, instructors should discuss the potential for overediting in PE tasks, which allows for greater flexibility in using MT (see Section 7.1.4). These points are instrumental in early-stage training to prepare future generations of translators who are not only satisfied with their profession but also equipped to tackle the evolving challenges of the language industry.

### 7.1.4 Why we should *not* refer to PE as *revision of MT output*

In light of what has been discussed so far, we strongly advocate for a reconsideration of how we conceptualise and refer to the task of post-editing MT output. It is crucial to avoid framing PE as the *revision of the MT output*. Such terminology can lead to misconceptions about the nature and expectations of the PE task. One potential risk is the tendency to humanise MT systems, which may result in either over-reliance on MT suggestions or a feeling of constraint when it comes to making substantial modifications to machine-pre-translated texts. Instead, we recommend reserving the term *revision* for its traditional meaning: a bilingual examination of a translated text, an activity in translation workflows

that involve at least two different linguists. Post-edited texts should certainly undergo a revision process. How the process of revising post-edited texts compares to revising human-translated texts remains an unanswered question that warrants further investigation.

The research conducted in this thesis underscores the need for a different approach to PE and revision tasks. While the benefits of revision tasks are evident and acknowledged by all linguists surveyed in our questionnaire, positive aspects of the PE task revolve primarily around enhancing productivity. Productivity improvements are crucial from a management perspective, but they might not be personally enriching for individual linguists. In particular, we posit that the emphasis in PE tasks should shift away from avoiding personal preferences and instead focus on the task as one that allows for greater creative freedom. Optional edits may increase editing effort, but we argue that they also contribute significantly to enhancing linguists' task satisfaction. Mellinger and Shreve (2016, p. 132) interpreted the tendency to over-edit TM suggestions as the result of a cognitive mismatch between a translation memory's candidate translation and the linguist's internal conception of what an optimal translation for that segment should be. This notion can be extended to MT suggestions as well. Post-editors are likely to implement true preferential changes when editing MT suggestions. While the revision task is inherently constrained by the involvement of another human [1], overediting MT suggestions should not be viewed as a negative practice. Instead, it has the potential to enhance productivity while also fostering positive attitudes toward PE tasks.

The introduction of MT into a professional translation workflow should lead to a positive change in terms of efficiency gains and provide valuable support to linguists in their daily tasks. However, realising these benefits becomes less likely if linguists view PE as nothing more than revising a text produced by an unskilled translator. Clarifying the place of MT in the workflow, providing linguists with proper PE training and customised MT engines, as well as organising regular team feedback sessions to discuss the challenges of working with this technology, could represent a way to add to PE the human interaction it lacks and that is much appreciated in revision tasks.

## 7.2 Additional contributions

First and foremost, our study provided empirical data on the relationship between PE and revision, shedding light on the intricate dynamics of these two crucial stages in the translation process. By examining a specific cohort of professional linguists, we offered insights into how these practices operate in a real-world setting. In addition to the findings detailed in the previous sections, we have contributed to the field of Translation Studies with four novel data collection instruments. These consist of:

— *two questionnaires for investigating PE and revision practices in a professional setting with a contrastive perspective.*

1. See Riondel, 2023 for further discussion of the relationship between translator and reviser in professional workflows.

These instruments allowed us to gather comprehensive data on PE and revision tasks in one professional context from two different points of view (the one of the management and the one of the linguist). The questionnaires can be used to describe how LSPs structure PE and revision practices in their workflows and how professional linguists interact with translation technology, with a level of detail seldom attained in previous surveys of a similar nature. They could be reused in different settings and also easily customised to encompass a diverse array of translation-related activities.

— *a questionnaire to investigate the satisfaction level of language professionals regarding translation-related activities.*
Although they were included as part of a broader questionnaire, questionnaire items on linguists' satisfaction can constitute a questionnaire on their own. By exploring linguists' perceptions regarding the activities they perform on the job, we offer insights into the factors that impact job satisfaction. Our questionnaire draws upon established instruments from the field of industrial-organisational psychology, as well as on previous satisfaction questionnaires from the field of Translation and Interpreting, thus ensuring robustness and reliability. Its adaptability makes it suitable for application in diverse professional contexts, enabling cross-category comparisons among language professionals.

— *a taxonomy to categorise edits in both human and machine-pre-translated texts.*
We proposed a novel approach to evaluating the nature and impact of edits. By making this taxonomy available to the research community, we facilitate more consistent and replicable assessments of modifications performed on pre-translated texts.

## 7.3 Limitations and perspectives for further research

Specific limitations inherent to each study, e.g. regarding the data collection methods employed, were discussed in individual chapters. In this section, we acknowledge the broader limitations of our approach and propose avenues for future research.

Firstly, it is essential to remember that we tackled the comparison between PE and revision from the perspective of a salaried linguist performing these tasks. When adopting the perspective of an end-user or the one of a freelance translator, different key aspects would deserve attention. For instance, pricing methods applied to PE and revision would likely be a crucial factor for these stakeholders.

Our study employed a multimethod approach primarily centred on quantitative methodologies. Future research endeavours could greatly benefit from the inclusion of qualitative research strategies. Conducting interviews or organising focus groups at the workplace would provide valuable insights into the workflow dynamics and complement the data we collected in our research. For example, a small percentage of respondents disagreed with the statement about PE increasing task variety. An interview could delve deeper into their perspectives. Did they feel that task variety remained the same, or rather, did they perceive that task variety has been reduced, with the introduction of MT as limiting their work solely to pre-translated texts?

While our investigation has primarily focused on the final product to investigate the process (Chapter 6), we acknowledge that other methods could better record the linguists' technical effort, revealing the extent of invisible editing that occurs during PE and revision. This may involve analysing keystroke logs – data that we collected but did not explore in this study. Similarly, an experimental study incorporating eye-tracking technology could validate the reading strategies reported by the linguists who responded to our questionnaire.

Our study focused on the relationship between PE and revision tasks within a specific cohort. This is the first study of its kind, and we advocate for more research into the unique characteristics of corporate in-house language services. A valuable direction for future research would be to quantify the presence of this cohort in other multilingual countries. Furthermore, replicating the productivity study in other CILS with different dynamics, such as those incorporating MT as an additional suggestion, could provide insights into how linguists' productivity is impacted by their method of working with MT.

Lastly, this thesis focused on machine and human-pre-translated texts, but it did not fully consider the role of translation memories. We framed translation as an activity that lets linguists start from a blank page. We intended to highlight the fact that, when comparing PE, revision and translation, the latter is the only activity that *allows for the possibility* to start from scratch. We acknowledge, however, that in CILS contexts, linguists leverage TMs extensively. Future research could quantify the extent to which linguists work with TM-pre-translated content in translation assignments and investigate satisfaction levels associated with mainly TM-pre-translated texts. This would complement our findings, offering valuable insights into how workflow variations affect linguists' satisfaction.

# Appendix A

# Survey invitation email

Bonjour,

Je m'appelle Sabrina Girletti et je suis doctorante à la Faculté de traduction et d'interprétation de l'Université de Genève. Dans le cadre de ma thèse, je m'intéresse à la relation entre les processus de post-édition de la traduction automatique neuronale et de révision de textes traduits, dans des contextes de traduction professionnelle.

Actuellement, je mène une enquête nationale auprès de services linguistiques internes aux entreprises qui travaillent avec (ou sans) la traduction automatique. L'objectif de cette étude est d'examiner comment la post-édition et la révision sont effectuées dans les services linguistiques internes des entreprises suisses. Les résultats mettront en lumière les pratiques actuelles et les problèmes à résoudre, offrant ainsi de nouvelles perspectives pour la formation des étudiants et des traducteurs expérimentés qui travaillent avec la traduction automatique.

L'enquête se compose de deux questionnaires :

— le premier (Q1) s'adresse aux responsables des services linguistiques ou aux gestionnaires de projets et a pour but de récolter des informations sur le service (p. ex. nombre d'employé(e)s, organisation des activités, etc.) et sur le système de traduction automatique utilisé ;

— le deuxième questionnaire (Q2) s'adresse aux linguistes internes (c'est-à-dire, les collaborateurs et collaboratrices du service linguistique qui s'occupent de traduction, révision ou post-édition) et porte sur leur pratique professionnelle, leurs opinions, ainsi que leur appréciation des activités exercées, notamment en ce qui concerne les tâches de post-édition et de révision.

Les questionnaires sont disponibles en allemand, anglais, français et italien, et accessibles via la plate-forme LimeSurvey. Voici les liens vers le premier questionnaire (Q1) : [LINKS] En cliquant sur le lien correspondant à la langue de votre choix, vous accéderez à la page de consentement informé, qui contient toutes les informations sur l'étude et sur le traitement des données. Une fois le premier questionnaire complété, et si vous travaillez déjà avec la

traduction automatique, je vous enverrai le lien vers le deuxième questionnaire (Q2) que vous pourrez partager avec les linguistes de votre équipe interne.
La participation à l'étude est volontaire et je serais très heureuse de vous présenter les résultats une fois qu'ils seront disponibles, à partir de février 2022.
Je vous remercie infiniment pour votre attention et je me tiens à disposition pour répondre à toute question ultérieure.

En vous souhaitant une excellente journée,
Meilleures salutations,
Sabrina Girletti

# Appendix B

# Consent form (survey-based study)

Dear respondent,
Thank you for your interest in our research project. Here below you will find further information regarding the main project and this questionnaire.
Please read all the information below before starting the survey.

**Main project title**: Empirical Investigation of Post-Editing of Neural Machine Translation Content and Traditional Translation Revision

**Researcher**: Sabrina Girletti (sabrina.girletti@unige.ch), PhD student, Faculty of Translation and Interpreting, University of Geneva

**Supervisor**: Pierrette Bouillon (pierrette.bouillon@unige.ch), Professor and Dean, Faculty of Translation and Interpreting, University of Geneva

## Goal

The aim of this study is to investigate how post-editing and revision are carried out in Swiss corporate in-house language services. The results will shed light on current practices and issues to address, providing new perspectives for the training of both students and experienced translators who work with machine translation. This questionnaire contains questions about the structure and workflow of your language service. (Q1) /or/ This questionnaire contains questions about your workflow, strategies and attitudes towards post-editing and revision. (Q2)

## Participation

This questionnaire is addressed to language services' directors or project managers. (Q1) /or/ This questionnaire is addressed to language service employees who translate, revise and post-edit texts. (Q2) Your participation in this survey is voluntary. You may refuse to take part in the research or exit the survey at any time without consequence.

## Duration

Depending on your answers, filling out this questionnaire may require between 5 and 20 minutes.

## Data collection

Data will be collected via Limesurvey, an online platform hosted in a protected server at the University of Geneva.

## Data collection period

Data will be collected between November 2021 and February 2022.

## Confidentiality

All the information you share with the Researcher will be kept strictly confidential.
Your personal and business data, such as your Company name, will be pseudonymised, i.e., replaced by a code. You will not be identifiable by your responses.
Only the Researcher and her Supervisor will have access to raw data (i.e., unprocessed survey responses). Your identity, answers and participation in this study will be known only to the Researcher and her Supervisor.
You can request the deletion of your personal data at any time until February 28th, 2022, after which your answers will be anonymised and it will no longer be possible to delete them.
At the end of the survey, you will be invited to participate in an additional interview by telephone, videoconference, or email. Please note that telephone and videoconference interviews will not be recorded, and emails will be deleted from the Researcher's inbox within 48 hours after receipt.
No names or identifying information will be included in any scientific publications or presentations based on these data.

## Data storage

Data will be stored:

— locally on the Researcher's computer, which is password-protected;

— in a protected server at the University of Geneva;

— on a password-protected USB key, stored in a locked cabinet at the University of Geneva.

Raw data, as defined in the previous section, will be stored on a password-protected USB key until completion of the study, and then deleted.
Anonymised and pseudonymised data will be stored indefinitely for scientific research purposes.

## Access to project results

The results will be available in aggregated form by the end of March 2022. No individual results will be made available, with no exceptions.
If you wish to access the aggregated results of this project, you can contact the Researcher by email (sabrina.girletti@unige.ch).
If you have any questions about this survey or the study in general, or if you would like a copy of this form, please do not hesitate to contact the Researcher by email (sabrina.girletti@unige.ch) or telephone +41(0) 22 379 86 85.
If you have questions about the ethical aspects of this study, please contact Professor Alexander Künzli, alexander.kuenzli@unige.ch, President of the Ethics Committee at the Faculty of Translation and Interpreting, University of Geneva.

Please tick the boxes to confirm the statements below:

- ☐ I have read and understood the information about the study.
- ☐ I participate in this study voluntarily.
- ☐ I understand that I can withdraw from the study at any time, without consequence.
- ☐ I understand that the anonymous data generated from this survey will be used in the Researcher's publications.

## Appendix C

# Questionnaire Q1

# Survey on post-editing and revision practices in Swiss corporate in-house language services Q1

Questions with an asterisk are mandatory.

## [A] Translation Service

### A01 - Company name *

### A02 - Select your language service's source and target languages *

Please choose **all** that apply:

- EN
- DE
- FR
- IT
- Other (please specify)

### A03 - Number of <u>in-house</u> linguists *

*Info: The term "linguist" comprises all language service employees who translate, revise or post-edit texts.*

### A04 - Number of in-house linguists <u>per target language</u>

### A05 - Average annual text volume (in words)

### A06 - Average annual text volume per language combination

*Info: Please give a percentage, e.g. DE>FR 50%, DE>IT 30%, DE>EN 15%, FR>DE 5%)*

### A07 - CAT-tool(s) currently in use *

### A08 - Do you use machine translation (MT) in your production workflows? *

- Yes
- No
- Not yet (we are currently testing MT)

### {If "No" at A08} A08.n01 - Are you considering introducing MT in your workflows in the next two years? *

- Yes
- No
- I don't know

*>>>exit questionnaire>>>*

### {If "Not yet" at A08} A08.ny01 - Who provides you with your current MT system (e.g. Textshuttle, Systran, DeepL, etc.)?

### {If "Not yet" at A08} A08.ny02 - Type of MT system currently in use *

- Customized (i.e. trained with your own translation memories)
- Generic (e.g. DeepL)
- I don’t know
- Other (e.g. more than one, please specify)

### {If "Not yet" at A08} A08.ny03 - Architecture of your MT system *

- Rule-based machine translation (RBMT)
- Statistical machine translation (SMT)
- Neural machine translation (NMT)
- I don’t know
- Other (or more than one, please specify)

### {If "Not yet" at A08} A08.ny04 - When will you start using MT in your production workflows (expected date)?

*>>>exit questionnaire>>>*

*{If "Yes" at A08, show next section}*

## [B] Workflow

### B09 - Percentage of in-house operations, on average (total must be 100%) *

Please choose all that apply and provide a comment:

- Translation
- Revision
- Other (please specify)

### B10 - Do all in-house linguists revise texts? *

- Yes, all in-house linguists revise texts
- No, only some in-house linguists revise texts

### {If "No" at B10} B10.n01 - For what reason?

### B11 - Do all in-house linguists post-edit texts? *

- Yes, all in-house linguists post-edit texts
- No, only some in-house linguists post-edit texts

### {If "No" at B11} B11.n01 - For what reason?

### B12 - Do you outsource some of your operations (e.g. translation, revision, etc.) *

- Yes
- No

### {If "Yes" at B12} B12.y01 - Which operations?

- Translation
- Revision
- Other (please specify)

### {If "Yes" at B12} B12.y02 - To whom do you outsource your operations?

- External freelancers
- Language service providers (agencies)
- Other (please specify)

## [C] Machine translation (MT) and post-editing (PE)

### C13 - Who provides you with your current MT system (e.g. Textshuttle, Systran, DeepL, etc.)?

### C14 - Type of MT system currently in use *

- Customized (i.e. trained with your own translation memories)
- Generic (e.g. DeepL)
- I don’t know
- Other (e.g. more than one, please specify)

### C15 - Architecture of your current MT system *

- Rule-based machine translation (RBMT)
- Statistical machine translation (SMT)
- Neural machine translation (NMT)
- I don’t know
- Other (or more than one, please specify)

### C16 - When did you start using your current MT system in the production workflow? *

### C17 - Did you assess the quality of MT or carry out a productivity test before integrating MT into the workflow? *

- Yes, we assessed MT quality
- Yes, we tested productivity gains
- Yes, we tested both
- No
- Other (please specify)

### {If "Yes" at C17} C17.y01 - Did you involve your in-house linguists in these assessments?

- Yes, all in-house linguists
- Yes, some in-house linguists
- No

### {If "Yes" at C17} C17.y02 - Did you consider linguists' feedback in your decision to implement MT?

- Yes
- No

### C18 - Is MT used on each project? *

- Yes
- No

### C19 - Who decides when to use MT? (e.g. PM, customer, etc.) *

### C20 - What are the criteria to apply MT? *

### C21 - Is MT integrated in your CAT-tool? *

- Yes
- No

### {If "No" at C21} C21.n01 - Please specify how you use your MT system (e.g. external interface) *

### {If "Yes" at C21} C21.y01 - What is your MT/CAT-tool integration scenario (regarding the display of suggestions)? *

- MT provided only if there are no translation memory matches or matches below a certain percentage
- MT always added as an additional suggestion, regardless of whether there are already translation memory matches
- Other (please specify)

### {If "Yes" at C21} C21.y02 - What is your MT/CAT-tool integration scenario (regarding the use of suggestions)? *

- Pre-translation of the entire text using MT and translation memory (i.e. one suggestion per segment, already in the editing zone)
- MT and translation memory suggestions in another window (i.e. the linguists can click on a suggestion to use it)
- Other (please specify)

### C22 - Can your in-house linguists change these parameters and choose their preferred integration scenario? *

- Yes
- No

### C23 - Did you implement different levels of post-editing (e.g. light/full PE)? *

- Yes
- No

*Info: As a reminder, light PE aims to obtain a comprehensible and factually accurate text, while full PE aims to provide quality similar or equal to human translation.*

### {If "Yes" at C23} C23.y01 - In which cases do you apply light PE?

### C24 - Do you have any PE guidelines that your in-house linguists can use for reference? *

- Yes, we drafted them internally
- Yes, we use existing guidelines (e.g. TAUS guidelines)
- No
- Other (please specify)

### C25 - Did you provide your in-house linguists with any post-editing training? *

- Yes
- No

## [D] Revision

### D26 - Which terms do you use in your company to refer to QA processes? Please provide a brief explanation for each of them.

### D27 - Are all translations revised? *

- Yes
- No

*Info: According to ISO standard 17100:2015, we intend revision as "the bilingual examination of source and target", carried out by "a person other than the translator".*

### {If "No" at D27} D27.n01 - If not all translations are revised, what are the selection criteria?

### D28 - Are texts always revised in their entirety?

- Yes
- No

Make a comment on your choice here:

### D29 - What is the status of revisers' corrections? *

- Corrections are sent back to the translator as suggestions
- Corrections are applied to the text by the reviser
- Other (please specify)

### D30 - Do you have any revision guidelines or parameters that your in-house linguists can use for reference? *

- Yes, we drafted them internally

- Yes, we use existing guidelines or parameters
- No
- Other (please specify)

*Info: Revision parameters are the types of errors a reviser checks for, e.g. accuracy, completeness etc.*

**D31 - Did you provide your in-house linguists with any revision training? ***

- Yes
- No

# [E] PE and revision

**E32 - Are post-edited texts also revised by another linguist? ***

- Yes, always
- Sometimes
- No

**{If "Sometimes" at E32} E32.sm01 - In which cases? ***

**{If "Yes, always" or "Sometimes" at E32} E32.ysm01 - When revising, do your in-house revisers know whether they are revising a human-translated text or a post-edited text? ***

- Yes, always
- Yes, sometimes
- No
- Other (please specify)

**Thank you very much for taking the time to fill this questionnaire.**

**If needed, would you be willing to provide more details about your answers by email or in a brief telephone interview?**

**If yes, please provide your contact details (email address and/or telephone number).**

*If you have any questions, please do not hesitate to contact the researcher by email (sabrina.girletti@unige.ch) or telephone +41 (0)22 379 86 85*

## Appendix D

# Questionnaire Q2

# Survey on post-editing and revision practices in Swiss corporate in-house language services Q2

Questions with an asterisk are mandatory.

## [A] Your profile

### A01 - Company name *

### A02 - Your mother tongue *

- DE
- EN
- FR
- IT
- Other (please specify)

### A03 - Your age *

- 18-29
- 30-39
- 40-49
- 50-59
- 60+

### A04 - Working in the translation field since (year) *

### A05 - Working at the corporate language service since (year) *

### A06 - Do you revise texts translated by other translators? *

- Yes, (almost) everyday
- Yes, sometimes
- Never

*Info: ISO standard 17100:2015 defines revision as the "bilingual examination of source and target", carried out by "a person other than the translator".*

### A07 - Do you use machine translation (MT) in your work? *

- Yes, (almost) everyday
- Yes, sometimes
- Never

# [B] Revision

*Info: ISO standard 17100:2015 defines revision as the "bilingual examination of source and target", carried out by "a person other than the translator".*

### B08 - Did you already have some professional revision experience before working at the company? *

- Yes
- No

### B09 - When revising, what is your main reading strategy? *

- I read the whole target text and refer to the source only if there are issues in target quality
- I read the source segment first, then the target segment
- I read the target segment first, then the source segment
- Other (please specify)

### B10 - Do you use a different reading strategy depending on the text or on other factors? *

- Yes
- No

Make a comment on your choice here:

### B11 - Do you follow any revision guidelines/parameters? *

- Yes
- No

*Info: Revision parameters are the types of errors a reviser checks for, e.g. accuracy, completeness etc.*

### B12 - When revising human-translated texts, do you check whether terminology is correct? *

- Always
- Often
- Sometimes
- Rarely
- Never

Make a comment on your choice here:

### B13 - Do you revise on paper? *

- Yes, (almost) always
- Yes, sometimes
- Never

### {If "Yes" at B13} B13.y01 - In which cases?

### B14 - Did you attend any revision training? *

- Yes
- No

# [C] Post-editing (PE)

*Info: ISO standard 18587:2017 defines post-editing as "[editing] and [correcting] machine translation output".*

### C15 - Did you already have some professional post-editing experience before working at the company? *

- Yes
- No

### C16 - When post-editing, what is your main reading strategy? *

- I read the whole target text and refer to the source only if there are issues in target quality
- I read the source segment first, then the target segment
- I read the target segment first, then the source segment
- Other (please specify)

### C17 - Do you use a different reading strategy depending on the text or on other factors? *

- Yes
- No

Make a comment on your choice here:

### C18 - Do you follow any post-editing guidelines? *

- Yes
- No

### C19 - When post-editing machine-translated texts, do you check whether terminology is correct? *

- Always
- Often
- Sometimes
- Rarely
- Never

Make a comment on your choice here:

### C20 - Do you post-edit on paper? *

- Yes, always
- Yes, sometimes
- Never

### {If "Yes" at C20} C20.y01 - In which cases?

**C21 - Did you attend any post-editing training? ***

- Yes
- No

**C22 - On average, how would you rate the quality produced by the machine translation system currently used at the company? ***

- Very good
- Good
- Neither good nor bad
- Bad
- Very bad

Make a comment on your choice here:

# [D] Post-editing, revision and overall strategies

**D23 - Do you use the same overall strategies when revising human-translated texts and post-editing machine-translated texts? ***

- Yes, I use the same strategies
- No, I use different strategies

Make a comment on your choice here:

**D24 - Do you use the same overall strategies when revising human-translated texts and revising texts that have been post-edited by another person? ***

- Yes, I use the same strategies
- No, I use different strategies

Make a comment on your choice here:

**D25 - After the introduction of MT in your workflow, did you change the way you revise texts?**

- Yes
- No

**{If "Yes" at D25} D25.y01 - If yes, how?**

# [E] Satisfaction

To what extent do you agree or disagree with the following statements?

**E26 - I enjoy... ***

| | Strongly disagree | Disagree | Neither disagree nor agree | Agree | Strongly agree | Cannot choose |
|---|---|---|---|---|---|---|
| **E26.01 ...translating (without using MT)** | | | | | | |
| **E26.02 ...revising (human) translations** | | | | | | |
| **E26.03 ...post-editing MT output** | | | | | | |

## E27 - I have the necessary skills and knowledge to carry out...*

| | Strongly disagree | Disagree | Neither disagree nor agree | Agree | Strongly agree | Cannot choose |
|---|---|---|---|---|---|---|
| **E27.01 ...translation jobs (without using MT)** | | | | | | |
| **E27.02 ...revision jobs** | | | | | | |
| **E27.03 ...post-editing jobs** | | | | | | |

*Info: ISO standard 17100:2015 defines revision as the "bilingual examination of source and target", carried out by "a person other than the translator".*

| | Strongly disagree | Disagree | Neither disagree nor agree | Agree | Strongly agree | Cannot choose |
|---|---|---|---|---|---|---|
| **E28 - The introduction of machine translation in my workflow has helped to make my daily work more varied and stimulating.*** | | | | | | |
| **E29 - I can choose the tasks I want to perform on a daily basis (e.g. ask to perform more translation, rather than revision).*** | | | | | | |

## E30 - Translation (without MT) is... *

| | Strongly disagree | Disagree | Neither disagree nor agree | Agree | Strongly agree | Cannot choose |
|---|---|---|---|---|---|---|
| **E30.01 ...a creative task** | | | | | | |
| **E30.02 ...stimulating** | | | | | | |
| **E30.03 ...complex (i.e. requires skills and knowledge)** | | | | | | |

## E31 - When translating without using MT... *

| | Strongly disagree | Disagree | Neither disagree nor agree | Agree | Strongly agree | Cannot choose |
|---|---|---|---|---|---|---|
| E31.01 ...I feel I have control over the text's final quality | | | | | | |
| E31.02 ...I make good use of my skills and knowledge | | | | | | |
| E31.03 ...I feel I am creating a new text | | | | | | |

## E32 - Revising human translations is… *

| | Strongly disagree | Disagree | Neither disagree nor agree | Agree | Strongly agree | Cannot choose |
|---|---|---|---|---|---|---|
| E32.01 ...a creative task | | | | | | |
| E32.02 ...stimulating | | | | | | |
| E32.03 ...complex (i.e. requires skills and knowledge) | | | | | | |

## E33 - When revising human translations... *

| | Strongly disagree | Disagree | Neither disagree nor agree | Agree | Strongly agree | Cannot choose |
|---|---|---|---|---|---|---|
| E33.01 ...I feel I have control over the text's final quality | | | | | | |
| E33.02 ...I make good use of my skills and knowledge | | | | | | |
| E33.03 ...I feel I am creating a new text | | | | | | |

## E34 - Post-editing MT output is… *

| | Strongly disagree | Disagree | Neither disagree nor agree | Agree | Strongly agree | Cannot choose |
|---|---|---|---|---|---|---|
| E34.01 ...a creative task | | | | | | |
| E34.02 ...stimulating | | | | | | |

| | Strongly disagree | Disagree | Neither disagree nor agree | Agree | Strongly agree | Cannot choose |
|---|---|---|---|---|---|---|
| **E34.03 ...complex (i.e. requires skills and knowledge)** | | | | | | |

**E35 - When post-editing MT output... ***

| | Strongly disagree | Disagree | Neither disagree nor agree | Agree | Strongly agree | Cannot choose |
|---|---|---|---|---|---|---|
| **E35.01 ...I feel I have control over the text's final quality** | | | | | | |
| **E35.02 ...I make good use of my skills and knowledge** | | | | | | |
| **E35.03 ...I feel I am creating a new text** | | | | | | |

**E36 - What do you appreciate about working with machine translation?**

**E37 - What do you not appreciate about working with machine translation?**

**E38 - What do you appreciate about revising human translations?**

**E39 - What do you not appreciate about revising human translations?**

**Thank you very much for taking the time to fill this questionnaire.**

**If needed, would you be willing to provide more details about your answers by email or in a brief telephone interview?**

**If yes, please provide your contact details (email address and/or telephone number).**

*If you have any questions, please do not hesitate to contact the researcher by email (sabrina.girletti@unige.ch) or telephone +41(0) 22 379 86 85*

## Appendix E

# Instructions for participants (productivity study)

Good morning,

The aim of this test is to compare revision with post-editing of translation memory and machine translation suggestions.

You will be working on two similar texts, each one of around 400 words, belonging to the project [NAME]. Some segments come from human translation, some others from machine translation or translation memory (fuzzy matches of 85% or above). The origin of the segment will not be shown (except for those coming from the translation memory).

In general, try to use as much of the translated segment as possible, and to make changes only when they are necessary to obtain a translation that is correct and fluent. However, if the suggested translation is incomprehensible or inadequate, you are free to erase the segment and translate it from scratch.

The final text shall be ready to be delivered to the client.

You will be working on your usual XTM interface and your screen will be recorded. You will have access to your usual resources. Important: the text has already been translated and it could be available online. Please do not look for the official translation.

Please let us know immediately if you experience connectivity issues or any other issues with the test environment.

Thank you for your valuable collaboration.

# Appendix F

# Consent form (productivity study)

## MT – Collaboration between Swiss Post and the FTI Productivity Test – Informed Consent Form

**Principal Investigators, Institution**: Sabrina Girletti and Pierrette Bouillon, University of Geneva.

Please complete the following:

Y/N – I have read the test instructions

Y/N – I understand the information provided

Y/N – I have had an opportunity to ask questions and discuss this test

Y/N – I have received satisfactory answers to all my questions

Y/N – I am aware that my screen will be recorded

I confirm that my involvement in the Productivity Test is voluntary.
I understand that I can withdraw from this Productivity Test at any time.
I am aware that under no circumstances will my personal data be published or revealed.
I understand that the data collected by the principal investigators during the Productivity Test may be presented in academic publications or conferences, but that no personal data will be identified.
I have read and understood the information in this form, and my questions and concerns have been answered by the principal investigators. Therefore, I consent to take part in this Productivity Test.

Participants' Signature:
Name in Block Capitals:
Date:

## Appendix G

# Instructions for evaluators (productivity study)

Dear [NAME],

Please find attached the Excel file for the evaluation.
The aim of this evaluation is to check whether there are any errors left in the translations. There are four translations and columns for comments. Please do not compare the translations with each other.

Read the source file and then one translation at a time. Mark in red the errors (if any) and use the "Comment" section to explain/correct it. Furthermore, if a translation seems dis-fluent, please let me know in the comment section.

Do not hesitate to contact me if you have any questions.

Thank you very much for your help.

Best regards,
Sabrina

# Appendix H

# Structure of the test subset (evaluation of edits)

| | *edits* | | *segments* | |
|---|---|---|---|---|
| **linguist** | **PE** | **REV** | **PE** | **REV** |
| L-FR001 | 42 | 9 | 31 | 9 |
| L-FR002 | 17 | 26 | 15 | 19 |
| L-FR003 | 18 | 26 | 13 | 21 |
| L-FR004 | 23 | 21 | 17 | 14 |
| L-FR005 | 52 | 42 | 29 | 28 |
| L-FR006 | 28 | 30 | 14 | 20 |
| L-FR008 | 12 | 12 | 7 | 10 |
| L-FR009 | 37 | 25 | 23 | 18 |
| L-FR012 | 1 | 4 | 1 | 3 |
| L-FR013 | 0 | 9 | 0 | 8 |
| **TOT** | **230** | **204** | **150** | **150** |

| | *edits* | | *segments* | |
|---|---|---|---|---|
| **linguist** | **PE** | **REV** | **PE** | **REV** |
| L-IT002 | 30 | 32 | 18 | 24 |
| L-IT003 | 50 | 35 | 32 | 30 |
| L-IT004 | 44 | 25 | 24 | 18 |
| L-IT006 | 29 | 16 | 16 | 9 |
| L-IT007 | 17 | 19 | 14 | 18 |
| L-IT008 | 48 | 41 | 31 | 31 |
| L-IT009 | 23 | 0 | 15 | 0 |
| L-IT011 | 0 | 7 | 0 | 5 |
| L-IT012 | 0 | 19 | 0 | 15 |
| **TOT** | **241** | **194** | **150** | **150** |

TABLE H.1 – Number of edits and segments in French (left side) and Italian (right side) subsets.

# Appendix I

# Instructions for evaluators (evaluation of edits)

Bonjour et merci d'avoir accepté de participer à cette évaluation.

Le but de l'évaluation est de déterminer si les modifications apportées pendant les tâches de révision et de post-édition (PE) sont en majorité nécessaires ou optionnelles. Vous allez évaluer une série de segments extraits de mandats de PE et de révision du client *Kommunikation*.
Voici un exemple de ce que vous trouverez dans la feuille *Evaluation* (I.1):

| ID | Target | Medium | Source | Pre-translation | Final version | Necessary or optional? | Comment |
|---|---|---|---|---|---|---|---|
| 20.03.1022_20 | Mitarbeiter | infocenter | Das Layout ist nach wie vor korrekt, das BAG hat lediglich die Hintergrundfarbe der Abbildungen geändert. | La mise en page reste *correcte,* l'OFSP a uniquement modifié la couleur d'arrière-plan des illustrations. | La mise en page reste *inchangée,* l'OFSP a modifié uniquement la couleur d'arrière-plan des illustrations. | | |

FIGURE I.1

La colonne A indique le numéro de segment et du mandat d'origine. Dans l'exemple ci-dessus: segment 20 du mandat 20.03.1022.
Les colonnes B et C (*Target* e *Medium*) contiennent des informations sur le mandat dont le segment a été extrait.
La colonne D contient le segment source en allemand.
La colonne E contient la prétraduction (humaine ou automatique).
La colonne F contient la version finale (post-éditée ou révisée).

La partie modifiée et sa correction apparaîssent en bleu dans les colonnes E et F, respectivement.
Chaque ligne contient une seule modification. Veuillez évaluer uniquement cette modification.
Si un segment contient plusieurs modifications, il est répété autant de fois qu'il y a de modifications dans le segment en question.
Si un mot qui apparaît en bleu dans la prétraduction n'a pas de correspondant dans sa version finale, il a été tout simplement éliminé.
À contrario, si rien n'est souligné dans la prétraduction et qu'un mot apparaît en bleu dans

la version finale, il a été ajouté dans celle-ci.
Évaluez la modification à l'aide de la liste déroulante dans la colonne G (cliquez sur la case G10 pour l'activer).
Astuce: après avoir évalué quelques segments à l'aide de la liste déroulante, vous pouvez accélerer le processus en cliquant N+Enter (pour *necessary*) ou O+Enter (pour *optional*).
Si, pour une raison spécifique, vous ne pouvez pas évaluer le segment, laissez un commentaire dans la colonne H.
Attention: il ne s'agit pas d'évaluer si la version modifiée est meilleure par rapport à la version initiale. Concentrez-vous plutôt sur la phrase prétraduite (colonne E) et demandez-vous s'il est nécessaire ou pas de modifier la partie en couleur.
Si vous avez la moindre question, n'hésitez pas à me contacter au 0223798685.
Merci beaucoup et bonne évaluation !

## Appendix J

# Approval letter from Ethical Review Board

UNIVERSITÉ
DE GENÈVE
FACULTÉ DE TRADUCTION
ET D'INTERPRÉTATION

**Prof. Alexander KÜNZLI**
Président de la Commission facultaire d'éthique
Ligne directe: 022 379 93 88
alexander.kuenzli@unige.ch

Mme Sabrina Girletti
Université de Genève
Faculté de traduction et d'interprétation
Sabrina.Girletti@etu.unige.ch

Genève, le 1er juillet 2021

N./Réf. 32

**Concerne** : demande d'approbation d'un projet de recherche par la Commission facultaire d'éthique

Chère Madame,

Par la présente, je confirme avoir bien reçu, le 17 juin dernier, votre demande d'approbation d'un projet de recherche intitulé « Survey on post-editing and revision practices in Swiss corporate in-house language services », ainsi que sa version remaniée du 29 juin 2021.

La Commission facultaire d'éthique a examiné votre demande à la lumière des normes pertinentes en la matière. Constatant que les modifications qu'elle avait demandées ont été apportées, elle considère que votre projet prévoit toutes les mesures nécessaires pour protéger les participants et pour garantir le respect des principes éthiques fondamentaux.

La Commission donne donc son approbation à la réalisation de votre projet du point de vue éthique.

Avec mes meilleures salutations,

Alexander Künzli

Prof. Alexander KÜNZLI

Président de la Commission
facultaire d'éthique

# Bibliography


AITI (2018). *Indagine sul mercato dei Traduttori e degli Interpreti 2018 - Associazione Italiana Traduttori e Interpreti (AITI)*.

Allen, Jeffrey (2003). "Post-editing". In: *Computers and Translation: A Translator's Guide*, pp. 297–317.

Allman, Spencer (2007). "Negotiating Translation Revision Assignments". In: *Translation and Negotiation. Proceedings of the 7th Annual Portsmouth Translation Conference*. Ed. by Ian Kemble, pp. 35–47.

Alvarez, Sergi, Antoni Oliver, and Toni Badia (2020). "Quantitative Analysis of Post-Editing Effort Indicators for NMT". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020*, pp. 411–420. ISBN: 9789893305898.

Álvarez-Vidal, Sergi, Antoni Oliver, and Toni Badia (2020). "Post-editing for Professional Translators : Cheer or Fear ?" In: *Revista Tradumàtica*, pp. 49–69.

— (2021). "What Do Post-Editors Correct? A Fine-Grained Analysis of SMT and NMT Errors". In: *Revista Tradumàtica* 19, pp. 131–147. ISSN: 15787559. DOI: 10.5565/rev/tradumatica.286.

Alves, Fabio et al. (2016). "Analysing the impact of interactive machine translation on post-editing effort". In: *New directions in empirical translation process research: exploring the CRITT TPR-DB*, pp. 77–94. ISSN: 3319203576.

Amabile, Teresa (2011). *Componential theory of creativity*. Harvard Business School Boston.

Amabile, Teresa M. et al. (2005). "Affect and creativity at work". In: *Administrative Science Quarterly* 50.3, pp. 367–403. ISSN: 00018392. DOI: 10.2189/asqu.2005.50.3.367.

Ananthakrishnan, R et al. (2007). "Some issues in automatic evaluation of english-hindi mt: more blues for bleu". In: *ICON*.

Aranberri, Nora (2017). "What do professional translators do when post-editing for the first time? first insight into the Spanish-Basque language pair". In: *Hermes (Denmark)* 56, pp. 89–110. ISSN: 19031785. DOI: 10.7146/hjlcb.v0i56.97235.

— (2020). "Can translationese features help users select an MT system for post-editing?" In: *Procesamiento del Lenguaje Natural* 64.0, pp. 93–100. ISSN: 19897553. DOI: 10.26342/2020-64-11.

Aranberri, Nora et al. (2014). "Comparison of post-editing productivity between professional translators and lay users". In: *Proceedings of the Third Workshop on Post-Editing Technology and Practice (WPTP-3)*, pp. 20–33.

Arthern, Peter J. (1987). "Four eyes are better than two". In: *Vital*. Vol. 1. 1. London: Aslib, pp. 14–26. DOI: 10.1038/sj.vital.vital040.

ATC and Nimdzi (2021). *ATC UK Language Services Industry Survey and Report*. Tech. rep.

Atkinson, David Peter (2012). "Freelance Translator Success and Psychological Skill : A Study of Translator Competence with Perspectives from Work Psychology". PhD thesis.

— (2014). "Developing psychological skill for the global language industry: An exploration of approaches to translator and interpreter training". In: *Translation Spaces* 3, pp. 1–24. ISSN: 2211-3711. DOI: 10.1075/ts.3.01atk.

Baird, Lloyd S. (1976). "Relationship of performance to satisfaction in stimulating and nonstimulating jobs." In: *Journal of Applied Psychology* 61.6, pp. 721–727. ISSN: 0021-9010. DOI: 10.1037//0021-9010.61.6.721.

Baker, Mona (1993). "Corpus Linguistics and Translation Studies: Implications and Applications." In: *Text and Technology. In Honour of John Sinclair*. Ed. by Mona Baker, Gill Francis, and Elena Tognini-Bonelli. Philadelphia/Amsterdam: John Benjamins Publishing Company.

— (1996). "Corpus-based Translation Studies: The Challenges that lie ahead". In: *Terminology, LSP and translation studies in language engineering in honour of Juan C. Sager*. Ed. by Harold Somers. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 175–186.

Bandura, Albert (1997). *Self-efficacy: The exercise of control*. New York.

Bar-Hillel, Yehoshua (1951). "The present state of research on mechanical translation". In: *American Documentation*, pp. 229–237.

Barnette, J. Jackson (2000). "Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems". In: *Educational and Psychological Measurement* 60.3, pp. 361–370. ISSN: 00131644. DOI: 10.1177/00131640021970592.

Baron, Reuben M. and David A. Kenny (1986). "The Moderator-Mediator Variable Distinction in Social Psychological Research. Conceptual, Strategic, and Statistical Considerations". In: *Journal of Personality and Social Psychology* 51.6, pp. 1173–1182. ISSN: 00223514. DOI: 10.1037/0022-3514.51.6.1173.

Bednárová-Gibová, Klaudia (2020). "Exploring the literary translator's work-related happiness: The case study of Slovakia". In: *Across Languages and Cultures* 21.1, pp. 67–87. ISSN: 15882519. DOI: 10.1556/084.2020.00004.

Bednárová-Gibová, Klaudia and Branislav Madoš (2019). "Investigating translators' work-related happiness: Slovak sworn and institutional translators as a case in point". In: *Meta* 64.1, pp. 215–241. ISSN: 00260452. DOI: 10.7202/1065335AR.

Bednárová-Gibová, Klaudia and Mária Majherová (2021). "A Comparative Study of Professional Wellbeing in Agency and Freelance Translators". In: *Cadernos de Tradução* 41.2, pp. 109–136. ISSN: 21757968. DOI: 10.5007/2175-7968.2021.e76280.

Beehr, Terry A. et al. (2006). "The nature of satisfaction with subordinates: Its predictors and importance to supervisors". In: *Journal of Applied Social Psychology* 36.6, pp. 1523–1547. ISSN: 00219029. DOI: 10.1111/J.0021-9029.2006.00070.X.

Bentivogli, Luisa et al. (2016). "Neural versus Phrase-Based Machine Translation Quality: a Case Study". In: *EMNLP-2016*. Vol. 2, pp. 257–267. arXiv: 1608.04631.

Bertaccini, Franco and Sara Di Nisio (2011). "Il traduttore e il revisore nei diversi ambiti professionali". In: *inTRAlinea*, pp. 1–9.

Bisiada, Mario (2017). "Universals of editing and translation". In: *Empirical Modelling of Translation and Interpreting*. Ed. by Silvia Hansen-Schirra, Oliver Czulo, and Sascha Hofmann. Berlin: Language Science Press, pp. 241–275.

— (2018a). "The editor's invisibility: Analysing editorial interventions in translation". In: *Target* 30.2, pp. 288–309. ISSN: 0924-1884. DOI: 10.1075/target.16116.bis.

— (2018b). "Translation and editing: a study of editorial treatment of nominalisations in draft translations". In: *Perspectives: Studies in Translatology* 26.1, pp. 24–38. ISSN: 17476623. DOI: 10.1080/0907676X.2017.1290121.

Black, Michelle L. et al. (2013). "Summer Research Training for Medical Students: Impact on Research Self-Efficacy". In: *Clinical and Translational Science* 6.6, pp. 487–489. ISSN: 1752-8062. DOI: 10.1111/CTS.12062.

Blain, Frédéric et al. (2011). "Qualitative Analysis of Post-Editing for High Quality Machine Translation". In: *Machine Translation Summit XIII*, pp. 164–171.

Blanz, Mathias (2017). "Employees' job satisfaction: A test of the job characteristics model among social work practitioners". In: *Journal of Evidence-Informed Social Work* 14.1, pp. 35–50. ISSN: 23761415. DOI: 10.1080/23761407.2017.1288187.

Bolaños-Medina, Alicia (2014). "Self-efficacy in translation". In: *Translation and Interpreting Studies* 9.2, pp. 197–218. ISSN: 1932-2798. DOI: 10.1075/tis.9.2.03bol.

— (2016). "Translation Psychology within the Framework of Translator Studies: New Research Perspectives and Pedagogical Implications". In: *From the Lab to the Classroom and Back Again: Perspectives on Translation and Interpreting Training*, pp. 59–99.

Bolaños-Medina, Alicia and Juan L. Núñez (2018). "A preliminary scale for assessing translators' self-efficacy". In: *Across Languages and Cultures* 19.1, pp. 53–78. ISSN: 15882519. DOI: 10.1556/084.2018.19.1.3.

Bonne, Linda and Michael Johnston (2016). "Students' beliefs about themselves as mathematics learners". In: *Thinking Skills and Creativity* 20, pp. 17–28. ISSN: 18711871. DOI: 10.1016/j.tsc.2016.02.001.

Bontempo, Karen and Jemina Napier (2011). "Evaluating emotional stability as a predictor of interpreter competence and aptitude for interpreting". In: *Interpreting* 13.1, pp. 85–105. DOI: 10.1075/intp.13.1.06bon.

Bordieu, Pierre (1986). "The forms of capital". In: *Handbook of Theory and research for the Sociology of Education*. Ed. by J. Richardson. Chap. 1, pp. 241–258.

Bouillon, Pierrette et al. (2018). "Integrating MT at Swiss Post's Language Service: preliminary results". In: *Proceedings of the 21st Annual Conference of the European Association*

*for Machine Translation*. Ed. by Juan Antonio Pérez-Ortiz et al., pp. 281–286. ISBN: 9788409019014.

Bowling, Nathan A., Kevin J. Eschleman, and Qiang Wang (2010). "A meta-analytic examination of the relationship between job satisfaction and subjective well-being". In: *Journal of Occupational and Organizational Psychology* 83.4, pp. 915–934. ISSN: 09631798. DOI: 10.1348/096317909X478557.

Bowling, Nathan A. and Valerie I. Sessa (2021). "Essentials of Job Attitudes and Other Workplace Psychological Constructs - An Introductory Chapter". In: *Essentials of Job Attitudes and Other Workplace Psychological Constructs*. Ed. by Valerie I. Sessa and Nathan A. Bowling. Routledge, pp. 3–12. ISBN: 9772081415. arXiv: arXiv:1011.1669v3.

Bowling, Nathan A., Stephen H. Wagner, and Terry A. Beehr (2018). "The Facet Satisfaction Scale : an Effective Affective Measure of Job Satisfaction Facets". In: *Journal of Business and Psychology* 33, pp. 383–403. DOI: 10.1007/s10869-017-9499-4.

Bowling, Nathan A. and Lucian Zelazny (2022). "Measuring General Job Satisfaction: Which Is More Construct Valid—Global Scales or Facet-Composite Scales?" In: *Journal of Business and Psychology* 37.1, pp. 91–105. ISSN: 1573353X. DOI: 10.1007/s10869-021-09739-2.

Brglez, Mojca and Špela Vintar (2022). "Lexical Diversity in Statistical and Neural Machine Translation". In: *Information* 13.2, p. 93. ISSN: 20782489. DOI: 10.3390/info13020093.

Briva-Iglesias, Vicent, Sharon O'Brien, and Benjamin R. Cowan (2023). "The impact of traditional and interactive post-editing on Machine Translation User Experience, quality, and productivity". In: *Translation, Cognition & Behavior*. DOI: 10.1075/tcb.00077.bri.

Brunette, Louise (1995). "Contribution à la pédagogie de la révision en pays bilingue: le cas du Canada". PhD thesis. Paris 3.

Brunette, Louise, Chantal Gagnon, and Jonathan Hine (2005). "The GREVIS Project: Revise or Court Calamity". In: *Across Languages and Cultures* 6.1, pp. 29–45. ISSN: 1585-1923. DOI: 10.1556/acr.6.2005.1.3.

Brunette, Louise and S O'Brien (2011). "Quelle ergonomie pour la pratique postéditrice des textes traduits ?" In: *ILCEA*. 14, pp. 1–9.

Bujacz, Aleksandra et al. (2016). "Why do we enjoy creative tasks? Results from a multigroup randomized controlled study". In: *Thinking Skills and Creativity* 19.October 2018, pp. 188–197. ISSN: 18711871. DOI: 10.1016/j.tsc.2015.11.002.

Bundgaard, Kristine (2017a). "(Post-)Editing – a Workplace Study of Translator-Computer Interaction at TextMinded Danmark A/S". PhD thesis.

— (2017b). "Translator attitudes towards translator-computer interaction - Findings from a workplace study". In: *Hermes (Denmark)* 56, pp. 125–144. ISSN: 19031785. DOI: 10.7146/hjlcb.v0i56.97228.

Bundgaard, Kristine, Tina Paulsen Christensen, and Anne Schjoldager (2016). "Translator-computer interaction in action — an observational process study of computer-aided translation". In: *Journal of Specialised Translation* 25.25, pp. 106–130. ISSN: 1740357X.

Burr, Renu and John Cordery (2001). "Self-management efficacy as a mediator of the relation between job design and employee motivation". In: *Human Performance* 14.1, pp. 27–44.

Cadwell, Patrick, Sharon O'Brien, and Carlos S.C. Teixeira (2017). "Resistance and accommodation: factors for the (non-) adoption of machine translation among professional translators". In: *Perspectives: Studies in Translatology* 26.3, pp. 301–321. ISSN: 17476623. DOI: 10.1080/0907676X.2017.1337210.

Cadwell, Patrick et al. (2016). "Human factors in machine translation and post-editing among institutional translators". In: *Translation Spaces* 5.2, pp. 222–243. ISSN: 2211-3711. DOI: 10.1075/ts.5.2.04cad.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn (2006). "Re-evaluating the role of BLEU in machine translation research". In: *EACL 2006 - 11th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 249–256. ISBN: 1932432590.

Canfora, Carmen and Angelika Ottmann (2020). "Risks in neural machine translation". In: *Translation Spaces* 9.1, pp. 58–77. ISSN: 2211-3711. DOI: 10.1075/ts.00021.can.

Caprara, Gian Vittorio et al. (2003). "Efficacy Beliefs as Determinants of Teachers' Job Satisfaction". In: *Journal of Educational Psychology* 95.4, pp. 821–832. ISSN: 00220663. DOI: 10.1037/0022-0663.95.4.821.

Caprara, Gian Vittorio et al. (2006). "Teachers' self-efficacy beliefs as determinants of job satisfaction and students' academic achievement: A study at the school level". In: *Journal of School Psychology* 44.6, pp. 473–490. ISSN: 00224405. DOI: 10.1016/j.jsp.2006.09.001.

Carl, Michael (2012). "Translog-II: A program for recording user activity data for empirical reading and writing research". In: *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, pp. 4108–4112. ISBN: 9782951740877.

Carl, Michael et al. (2011). "The Process of Post-Editing: a Pilot Study". In: *8th international NLPCS workshop*. ISSN: 0905-7269.

Casas, Sandra (2021). "Intégration de la post-édition dans un cours de traduction et révision". In: *Proceedings of the Affumt 2021 Conference: Former aux métiers de la traduction aujourd'hui et demain*, pp. 13–14.

Cascio, Wayne F. and Herman Aguinis (2005). *Applied Psychology in Human Resource Management*. 6th. Pearson Education.

Castilho, Sheila (2020). "On the Same Page? Comparing Inter-Annotator Agreement in Sentence and Document Level Human Machine Translation Evaluation". In: *Proceedings of the Fifth Conference on Machine Translation*, pp. 1150–1159.

Castilho, Sheila, Maja Popović, and Andy Way (2020). "On context span needed for machine translation evaluation". In.

Castilho, Sheila and Natália Resende (2022). "Post-Editese in Literary Translations". In: *Information* 13.2, p. 66.

Castilho, Sheila, Natália Resende, and Ruslan Mitkov (2019). "What Influences the Features of Post-Editese? A Preliminary Study". In: pp. 19–27. DOI: 10.26615/issn.2683-0078.2019_003.

Castilho, Sheila et al. (2017a). "A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators". In: *Proceedings of Machine Translation Summit XVI*. Nagoya, Japan.

Castilho, Sheila et al. (2017b). "Is Neural Machine Translation the New State of the Art?" In: *The Prague Bulleting of Mathematical Linguistics* 108, pp. 109–120. DOI: 10.1515/pralin-2017-0013.PBML.

Castilho, Sheila et al. (2018). "Evaluating MT for massive open online courses: A multifaceted comparison between PBSMT and NMT systems". In: *Machine Translation* August. ISSN: 0922-6567. DOI: 10.1007/s10590-018-9221-y.

CBTI-BKVT (2018). *Translation and Interpreting in Belgium: Market Survey Report 2018*. Tech. rep.

Chakhachiro, Raymond (2015). "Revision for quality". In: *Perspectives: Studies in Translatology* November. DOI: 10.1080/09076760508668994.

Chen, Gilad, Stanley M. Gully, and Dov Eden (2001). "Validation of a New General Self-Efficacy Scale". In: *Organizational Research Methods* 4.1, pp. 62–83. ISSN: 10944281. DOI: 10.1177/109442810141004.

Chen, Gilad, Stanley M Gully, and Robert N Kilcullen (2000). "Examination of Relationships Among Trait-Like Individual Differences, State-Like Individual Differences, and Learning Performance". In: *Journal of Applied Psychology* 85.6, pp. 835–847.

Chen, Zheng (2018). "Job satisfaction among frontline police officers in China: the role of demographic, work-related, organizational and social factors". In: *Psychology, Crime and Law* 24.9, pp. 895–914. ISSN: 14772744. DOI: 10.1080/1068316X.2018.1442835.

Chesterman, Andrew (2006). "Questions in the sociology of translation". In: pp. 9–27. DOI: 10.1075/BTL.68.03CHE.

— (2017). "The Name and Nature of Translator Studies". In: *HERMES - Journal of Language and Communication in Business* 22.42, pp. 13–22. ISSN: 0904-1699. DOI: 10.7146/hjlcb.v22i42.96844.

Choi, Jungwha (2007). "Study on Job Satisfaction and Directions for the Training of Conference Interpreters". In: *FORUM. Revue internationale d'interprétation et de traduction / International Journal of Interpretation and Translation* 5.2, pp. 23–38. ISSN: 1598-7647. DOI: 10.1075/forum.5.2.02cho.

Chyung, Seung Youn Yonnie, Julie R. Barkin, and Jennifer A. Shamsy (2018). "Evidence-Based Survey Design: The Use of Negatively Worded Items in Surveys". In: *Performance Improvement* 57.3, pp. 16–25. DOI: 10.1002/pfi.21749.

Cohen, Jacob (1960). "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1, pp. 37–46. ISSN: 15523888. DOI: 10.1177/001316446002000104.

Costa, Ângela et al. (2015). "A linguistically motivated taxonomy for Machine Translation error analysis". In: *Machine Translation* 29.2, pp. 127–161. ISSN: 15730573. DOI: 10.1007/s10590-015-9169-0.

Courtney, Jennifer and Mary Phelan (2019). "Translators' experiences of occupational stress and job satisfaction". In: *Translation and Interpreting* 11.1, pp. 100–113. ISSN: 18369324. DOI: 10.12807/ti.111201.2019.a06.

Crawford, Scott D., Mick P. Couper, and Mark J. Lamias (2001). "Web surveys: Perceptions of burden". In: *Social Science Computer Review* 19.2, pp. 146–162. ISSN: 08944393. DOI: 10.1177/089443930101900202.

Cressey, William W. (1987). "Review Article: Project for Integrated Development of English-Spanish Machine Translation". In: *Computers and Translation* 2.

Čulo, Oliver, Silvia Hansen-Schirra, and Jean Nitzke (2017). "Contrasting terminological variation in post-editing and human translation of texts from the technical and medical domain". In: *Empirical Translation Studies*. De Gruyter. DOI: 10.1515/9783110459586-007.

Čulo, Oliver and Jean Nitzke (2016). "Patterns of Terminological Variation in Post-editing and of Cognate Use in Machine Translation in Contrast to Human Translation". In: *Baltic Journal of Modern Computing* 4.2, SI, pp. 106–114. ISSN: 2255-8942.

Čulo, Oliver et al. (2014). "The Influence of Post-Editing on Translation Strategies". In: *Post-editing of Machine Translation : Processes and Applications*, pp. 200–218.

Cumbreño, Cristina and Nora Aranberri (2021). "What Do You Say? Comparison of Metrics for Post-editing Effort". In: *Explorations in Empirical Translation Process Research*. Springer, Cham, pp. 57–79. DOI: 10.1007/978-3-030-69777-8_3.

Daems, Joke, Orphée De Clercq, and Lieve Macken (2017). "Translationese and post-editese: How comparable is comparable quality?" In: *Linguistica Antverpiensia* 16, pp. 89–103. ISSN: 03042294.

Daems, Joke and Lieve Macken (2019). "Interactive adaptive SMT versus interactive adaptive NMT: a user experience evaluation". In: *Machine Translation*. ISSN: 15730573. DOI: 10.1007/s10590-019-09230-z.

— (2021). "Post-Editing Human Translations and Revising Machine Translations". In: *Translation Revision and Post-Editing*. Ed. by Maarit Koponen et al. Routledge, pp. 50–70. DOI: 10.4324/9781003096962-5.

Daems, Joke, Lieve Macken, and Sonia Vandepitte (2013). "Quality as the sum of its parts: a two-step approach for the identification of translation problems and translation quality assessment for HT and MT+PE". In: *MT Summit XIV Workshop on Post-editing Technology and Practice, Proceedings*, pp. 63–71.

Daems, Joke et al. (2017a). "Identifying the machine translation error types with the greatest impact on post-editing effort". In: *Frontiers in Psychology* 8.AUG, pp. 1–15. ISSN: 16641078. DOI: 10.3389/fpsyg.2017.01282.

— (2017b). "Translation methods and experience: A comparative analysis of human translation and post-editing with students and professional translators". In: *Meta* 62.2, pp. 245–270. ISSN: 00260452. DOI: 10.7202/1041023ar.

Dalal, Reeshad S. (2013). "Job Attitudes: Cognition and Affect". In: *Handbook of Psychology, Second Edition*. Ed. by Irving B. Weiner. Wiley & Sons. Chap. 14, pp. 341–366. DOI: 10.1002/9781118133880.hop212014.

Dam, Helle V and Karen Korning Zethsen (2016). ""I think it is a wonderful job" on the solidity of the translation profession". In: *The Journal of Specialised Translation* 25, pp. 174–187. ISSN: 1740357X.

Damerau, Fred J (1964). "A Technique for Computer Detection and Correction of Spelling Errors". In: *Communications of the ACM* 7.3, pp. 171–176.

de Almeida, Giselle (2013). "Translating the post-editor: an investigation of post-editing changes and correlations with professional experience across two Romance languages". PhD thesis, p. 350.

de Almeida, Giselle and Sharon O'Brien (2010). "Analysing Post-Editing Performance: Correlations with Years of Translation Experience". In: *Proceedings of the 14th annual conference of the European association for machine translation, St. Raphaël, France* May, pp. 27–28.

De Clercq, Orphée et al. (2021). "Uncovering machine translationese using corpus analysis techniques to distinguish between original and machine-translated French". In: *Translation Quarterly* 101, pp. 21–45.

De Simone, S, G Cicotto, and J Lampis (2016). "Occupational stress, job satisfaction and physical health in teachers". In: *Revue Europeenne de Psychologie Appliquee* 66.2, pp. 65–77. ISSN: 11629088. DOI: 10.1016/j.erap.2016.03.002.

de Sousa, Sheila C.M., Wilker Aziz, and Lucia Specia (2011). "Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles". In: *Recent Advances in Natural Language Processing, RANLP*. Hissar, Bulgaria, pp. 97–103.

de Sousa Mata, Ádala Nayana et al. (2021). "Training in communication skills for self-efficacy of health professionals: a systematic review". In: *Human Resources for Health* 19.1. ISSN: 14784491. DOI: 10.1186/s12960-021-00574-3.

Depraetere, Ilse, Nathalie De Sutter, and Arda Tezcan (2014). "Post-edited quality, post-editing behaviour and human evaluation: a case study". In: *Post-editing of Machine Translation*. Cambridge Scholars Publishing, pp. 78–109. ISBN: 1-4438-5476-X.

Dilig-Ruiz, Alison et al. (2018). "Job satisfaction among critical care nurses: A systematic review". In: *International Journal of Nursing Studies* 88, pp. 123–134. ISSN: 00207489. DOI: 10.1016/j.ijnurstu.2018.08.014.

Dillon, Sarah and Janet Fraser (2007). "Translators and TM : An investigation of translators ' perceptions of translation memory adoption". In: *Machine Translation*, pp. 67–79. DOI: 10.1007/s10590-006-9004-8.

do Carmo, Félix (2019). "Edit distances do not describe editing, but they can be useful for translation process research". In: pp. 1–2. DOI: 10.1145/363958.363994.

— (2020). "'Time is money' and the value of translation". In: *Translation Spaces* 9.1, pp. 35–57. ISSN: 2211-3711. DOI: 10.1075/ts.00020.car.

— (2021a). "Chasing the Complexity of Threads from the Translation Process that Hide in Translation Data". In: *Exploring the Implications of Complexity Thinking for Translation Studies*. London: Routledge, pp. 143–169. ISBN: 9781003105114. DOI: 10.4324/9781003105114-6.

— (2021b). "Editing Actions: A Missing Link Between Translation Process Research and Machine Translation Research". In: *Explorations in Empirical Translation Process Research*. Ed. by Michael Carl. Springer Nature, pp. 3–38. DOI: 10.1007/978-3-030-69777-8_1.

do Carmo, Félix and Joss Moorkens (2021). "Differentiating editing, post-editing and revision". In: *Translation Revision and Post-Editing. Industry Practices and Cognitive Processes*. Ed. by Maarit Koponen et al. Routledge. Chap. 2, pp. 35–49.

do Carmo, Félix et al. (2021). "A review of the state-of-the-art in automatic post-editing". In: *Machine Translation* 35.2, pp. 101–143. ISSN: 15730573. DOI: 10.1007/s10590-020-09252-y.

Dodd, Nancy G. and Daniel C. Ganster (1996). "The interactive effects of variety, autonomy, and feedback on attitudes and performance". In: *Journal of Organizational Behavior* 17.4, pp. 329–347. ISSN: 08943796. DOI: 10.1002/(SICI)1099-1379(199607)17:4<329::AID-JOB754>3.0.CO;2-B.

Doherty, Stephen (2016). "The impact of translation technologies on the process and product of translation". In: *International Journal of Communication* 10, pp. 947–969.

Dunnette, Marvin D., John P. Campbell, and Milton D. Hakel (1967). "Factors contributing to job satisfaction and job dissatisfaction in six occupational groups". In: *Organizational Behavior and Human Performance* 2.2, pp. 143–174. ISSN: 00305073. DOI: 10.1016/0030-5073(67)90027-X.

Eagly, Alice H. and Shelly Chaiken (1993). *The Psychology of Attitudes*. Harcourt Brace Jovanovich College Publishers. ISBN: 9780155000971.

Edmundson, Harold P and David G Hays (1958). "Research methodology for machine translation." In: *Mech. Transl. Comput. Linguistics* 5.1, pp. 8–15.

Egdom, Gys-Walt van and Mark Pluymaekers (2020). "Quality According to Language Service Providers. The Case of Post-Edited Machine Translation". In: *New Empirical Perspectives on Translation and Interpreting*. Ed. by Lore Vandevoorde, Joke Daems, and Bart Defrancq. New York: Routledge. Chap. 7, pp. 139–155. DOI: 10.4324/9780429030376-7.

Ehrensberger-Dow, Maureen and Gary Massey (2014). "Translators and machines : working together". In: pp. 1–9.

— (2020). "Translation workplace-based research". In: *The Routledge Handbook of Translation and Technology*. Ed. by Minako O'Hagan. Routledge, pp. 354–366.

Ehrensberger-Dow, Maureen and Sharon O'Brien (2015). "Ergonomics of the Translation Workplace: Potential for Cognitive Friction". In: *Translation Spaces* 4.1, pp. 98–118. ISSN: 2211-3711. DOI: 10.1075/ts.4.1.05ehr.

Ehrensberger-Dow, Maureen et al. (2016). "An International Survey of the Ergonomics of Professional Translation". In: *ILCEA. Revue de l'Institut des langues et cultures d'Europe, Amérique, Afrique, Asie et Australie* 27.27. ISSN: 1639-6073. DOI: 10.4000/ilcea.4004.

ELIA et al. (2019). "Expectations and concerns of the European Language Industry 2019". In.

ELIS (2023). *European Language Industry Survey (ELIS) 2023*. Tech. rep.

Elming, Jakob, Laura Winther Balling, and Michael Carl (2014). "Investigating User Behaviour in Post-editing and Translation using the CASMACAT Workbench". In: *Post-editing of Machine Translation: Processes and Applications*. Ed. by Lucia O'Brien, Sharon,

Balling, Laura Winther, Carl, Michael, Simard, Michel, Specia. Cambridge Scholars Publishing. Chap. 7, pp. 147–169. ISBN: 144385476x, 9781443854764.

EMT Expert Group (2017). *European Masters' in Translation Competence Framework*. Tech. rep., pp. 1–12.

Ewen, Robert B., Patricia C. Smith, and Charles L. Hulin (1966). "An Empirical Test of the Herzberg Two-Factor Theory". In: *Journal of Applied Psychology* 50.6, pp. 544–550. ISSN: 00219010. DOI: 10.1037/h0024042.

Farrell, Michael (2018). "Machine Translation Markers in Post-Edited Machine Translation Output". In: *Proceedings of the 40th Conference Translating and the Computer*, pp. 50–59.

— (2022). "Do translators use machine translation and if so, how? Results of a survey held among professional translator". In: *Proceedings of the 44th Conference Translating and the Computer*. Tradulex.

Farrús, Mireia et al. (2010). "Linguistic-based evaluation criteria to identify statistical machine translation errors". In: *EAMT 2010 - 14th Annual Conference of the European Association for Machine Translation* May.

Federici, Roger A and Einar M Skaalvik (2012). "Principal self-efficacy: Relations with burnout, job satisfaction and motivation to quit". In: *Social Psychology of Education* 15.3, pp. 295–320. ISSN: 13812890. DOI: 10.1007/s11218-012-9183-5.

Federico, M, Alessandro Cattelan, and Marco Trombetti (2012). "Measuring user productivity in machine translation enhanced computer assisted translation". In: *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers*.

Fiederer, Rebecca and Sharon O'Brien (2009). "Quality and Machine Translation: A realistic objective?" In: *The Journal of Specialised Translation* 11.11, pp. 52–74. ISSN: 1740357X.

Finn, C. P. (2001). "Autonomy: An important component for nurses' job satisfaction". In: *International Journal of Nursing Studies* 38.3, pp. 349–357. ISSN: 00207489. DOI: 10.1016/S0020-7489(00)00065-1.

Fischer, Lukas and Samuel Läubli (2020). "What's the Difference between Professional Human and Machine Translation? A Blind Multi-language Study on Domain-specific MT". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020*, pp. 215–224. ISBN: 9789893305898. arXiv: 2006.04781.

Fisher, Cynthia D. (1980). "On the Dubious Wisdom of Expecting Job Satisfaction to Correlate with Performance". In: *Academy of Management Review* 5.4, pp. 607–612. ISSN: 0363-7425. DOI: 10.5465/amr.1980.4288967.

— (2003). "Why do lay people believe that satisfaction and performance are correlated? Possible sources of a commonsense theory". In: *Journal of Organizational Behavior* 24.6, pp. 753–777. DOI: 10.1002/JOB.219.

Fisher, Gwenith G., Russell A. Matthews, and Alyssa Mitchell Gibbons (2016). "Developing and investigating the use of single-item measures in organizational research". In: *Journal of Occupational Health Psychology* 21.1, pp. 3–23. ISSN: 19391307. DOI: 10.1037/a0039139.

Flanagan, Marian and Tina Paulsen Christensen (2014). *Testing post-editing guidelines: How translation trainees interpret them and how to tailor them for translator training purposes*. DOI: 10.1080/1750399X.2014.936111.

Flournoy, Raymond and Christine Duran (2009). "Machine translation and document localization at Adobe: From pilot to production". In: pp. 425–428.

Forsgren, Susanne, Berit Forsman, and Eric D. Carlström (2009). "Working with Manchester triage - Job satisfaction in nursing". In: *International Emergency Nursing* 17.4, pp. 226–232. ISSN: 1755599X. DOI: 10.1016/j.ienj.2009.03.008.

Fried, Yitzhak and Gerald R. Ferris (1987). "The Validity of the Job Characteristics Model: a Review and Meta-Analysis". In: *Personnel Psychology* 40.2, pp. 287–322. ISSN: 17446570. DOI: 10.1111/j.1744-6570.1987.tb00605.x.

García, Ignacio (2006). "Translators on translation memories: a blessing or a curse?" In: *Translation Technology and its Teaching (with much mention of localization)* July 2002, pp. 97–105.

Garcia, Ignacio (2011). "Translating by post-editing: Is it the way forward?" In: *Machine Translation* 25.3, pp. 217–237. ISSN: 09226567. DOI: 10.1007/s10590-011-9115-8.

García, Ignacio (2012). "A brief history of post-edting and of research on post-editing". In: *Revista Anglo Saxonica* III.3, pp. 291–310. ISSN: 0873-0628.

Gaspari, Federico (2022). "Expanding the reach of corpus-based translation studies". In: *Extending the Scope of Corpus-Based Translation Studies*. Bloomsbury Publishing, pp. 42–63.

Gaspari, Federico, Hala Almaghout, and Stephen Doherty (2015). "A survey of machine translation competences: Insights for translation technology educators and practitioners". In: *Perspectives: Studies in Translatology* 23.3, pp. 333–358. ISSN: 17476623. DOI: 10.1080/0907676X.2014.979842.

Ginovart Cid, Clara (2021). "The need for practice in the acquisition of the post-editing skill-set: lessons learned from the industry". PhD thesis.

Ginovart Cid, Clara and Carme Colominas (2021). "The MT Post-Editing Skill Set". In: *Translation Revision and Post-Editing*. Ed. by Maarit Koponen et al. Routledge, pp. 226–246. DOI: 10.4324/9781003096962-17.

Girletti, Sabrina (2022). "Working with Pre-translated Texts: Preliminary Findings from a Survey on Post-editing and Revision Practices in Swiss Corporate In-house Language Services". In: *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pp. 269–278.

Girletti, Sabrina et al. (2019). "Preferences of end-users for raw and post-edited NMT in a business environment". In: *Proceedings of the 41st Conference Translating and the Computer*. AsLing, The International Association for Advancement in Language Technology, pp. 47–59.

Gough, Joanna (2011). "An empirical study of professional translators' attitudes, use and awareness of Web 2.0 technologies, and implications for the adoption of emerging technologies and trends". In: *Linguistica Antverpiensia, New Series – Themes in Translation Studies* 10, pp. 195–217. ISSN: 2295-5739.

Goulet, Marie-Josée et al. (2017). "La traduction automatique comme outil d'aide à la rédaction scientifique en anglais langue seconde : résultats d'une étude exploratoire sur la qualité linguistique". In: *ASp* 72, pp. 5–28. ISSN: 1246-8185. DOI: 10.4000/asp.5045.

Granger, Sylviane and Marie Aude Lefer (2020). "The Multilingual Student Translation corpus: a resource for translation teaching and research". In: *Language Resources and Evaluation* 54.4, pp. 1183–1199. ISSN: 15728412. DOI: 10.1007/s10579-020-09485-6.

— (2021). *TAS - Translation-oriented Annotation System manual*.

— (2022). "Corpus-based translation and interpreting studies". In: *Extending the scope of corpus-based translation studies*. Bloomsbury Publishing, pp. 13–41.

Green, Spence, Jeffrey Heer, and Christopher D Manning (2013). "The efficacy of human post-editing for language translation". In: *Conference on Human Factors in Computing Systems - Proceedings*, pp. 439–448. ISBN: 9781450318990. DOI: 10.1145/2470654.2470718.

Grether, Thorana, Julia F. Sowislo, and Bettina S. Wiese (2018). "Top-down or bottom-up? Prospective relations between general and domain-specific self-efficacy beliefs during a work-family transition". In: *Personality and Individual Differences* 121.August 2017, pp. 131–139. ISSN: 01918869. DOI: 10.1016/j.paid.2017.09.021.

Griffin, Ricky W., Ann Welsh, and Gregory Moorhead (1981). "Perceived Task Characteristics and Employee Performance : A Literature Review". In: *The Academy of Management Review* 6.4, pp. 655–664. DOI: 10.5465/AMR.1981.4285726.

Groves, Declan and Dag Schmidtke (2009). "Identification and Analysis of Post-Editing Patterns for MT". In: *Proceedings of MT Summit XII*, pp. 429–436.

Guerberof, Ana and Joss Moorkens (2019). "Machine translation and post-editing training as part of a master's programme". In: *The Journal of Specialised Translation* 31, pp. 217–238.

Guerberof Arenas, Ana (2012). "Productivity and quality in the post-editing of outputs from translation memories and machine translation". PhD thesis. Universitat Rovira I Virgili.

— (2013). "What do professional translators think about post-editing?" In: *The Journal of Specialised Translation* 19, pp. 75–95.

— (2014a). "Correlations between productivity and quality when post-editing in a professional context". In: *Machine Translation* 28.3-4, pp. 165–186. ISSN: 09226567. DOI: 10.1007/s10590-014-9155-y.

— (2014b). "The Role of Professional Experience in Post-editing". In: *Post-editing of Machine Translation*, pp. 51–76.

Guerberof Arenas, Ana and Antonio Toral (2022). "Creativity in translation: machine translation as a constraint for literary texts". In: *Translation Spaces* March, pp. 1–29. arXiv: 2204.05655v1.

Hackman, Richard J. and Edward E. Lawler (1971). "Employee reactions to job characteristics". In: *Journal of Applied Psychology* 55.3, pp. 259–286. ISSN: 00219010. DOI: 10.1037/h0031152.

Hackman, Richard J. and Greg R. Oldham (1974). *The Job Diagnostic Survey: An Instrument for the Diagnosis of Jobs and the Evaluation of Job Redesign Projects*. Tech. rep. May, p. 75.

— (1975). "Development of the Job Diagnostic Survey". In: *Journal of Applied Psychology* 60.2, pp. 159–170. ISSN: 00219010. DOI: 10.1037/h0076546.

— (1976). "Motivation through the Design of Work: Test of a Theory". In: *Organizational Behavior and Human Performance* 16.170, pp. 250–279. ISSN: 0030-5073. arXiv: arXiv:1011.1669v3.

— (1980). *Work redesign*. Vol. 2779. Reading, Mass.: Addison-Wesley. ISBN: 0201027798.

Hagan, John and Fiona Kay (2007). "Even lawyers get the blues: Gender, depression, and job satisfaction in legal practice". In: *Law and Society Review* 41.1, pp. 51–78. ISSN: 00239216. DOI: 10.1111/j.1540-5893.2007.00291.x.

Hagemann, Susanne (2019). "Directionality in translation and revision teaching: a case study of an A–B teacher working with B–A students". In: *The Interpreter and Translator Trainer* 13.1, pp. 86–101. ISSN: 1750-399X. DOI: 10.1080/1750399X.2018.1540742.

Haro-Soler, María del Mar (2017). "Teaching practices and translation students' self-efficacy: A qualitative study of teachers'perceptions". In: *Current Trends in Translation Teaching and Learning E* 4, pp. 198–228.

Haro-Soler, María del Mar (2019). "Vicarious learning in the translation classroom: how can it influence students' self-efficacy beliefs?" In: *English Studies at NBU* 5.1, pp. 92–113. ISSN: 2367-5705. DOI: 10.33919/esnbu.19.1.5.

Hassan, Hany et al. (2018). "Achieving Human Parity on Automatic Chinese to English News Translation". In: arXiv: 1803.05567.

Heafield, Kenneth (2011). "KenLM: Faster and smaller language model queries". In: *Proceedings of the sixth workshop on statistical machine translation*, pp. 187–197.

Heneman, Herbert G. and Donald P. Schwab (1985). "Pay Satisfaction: its Multidimensional Nature and Measurement". In: *International Journal of Psychology* 20.1, pp. 129–141. ISSN: 1464066X. DOI: 10.1080/00207598508247727.

Hermans, Johan and José Lambert (1998). "From Translation Markets to Language Management". In: *Target. International Journal of Translation Studies* 10.1, pp. 113–132. DOI: 10.1075/TARGET.10.1.06HER.

Hernández Morin, Katell (2009a). "Pratiques et perceptions de la révision en France". In: *Traduire* 221, pp. 58–78. ISSN: 0395-773X. DOI: 10.4000/traduire.368.

— (2009b). "La révision comme clé de la gestion de la qualité des traductions en contexte professionnel". PhD thesis.

Herzberg, Frederick, Bernard Mausner, and Barbara Bloch Snyderman (1959). *The Motivation to Work*. 2nd. John Wiley.

Ho, Hsin-Ning (2010). "The relationship between levels of expertise, task difficulty, perceived self-efficacy, and mental effort investment in task performance". PhD thesis, p. 153. ISBN: 9781109771961.

Hoerger, Michael (2010). "Participant Dropout as a Function of Survey Length in Internet-Mediated University Studies : Implications in Psychological Research". In: *Cyberpsychology, Behavior, and Social Networking* 13.6, pp. 697–701.

Hofmann, Nadira and Maryse Lepan (2017). "MT in real-world practice: Challenges and solutions at Swiss Federal Railways". In: *20th Annual Conference of the European Association for Machine Translation, EAMT 2017*, pp. 36–40.

Hora, Gabriela Pereira Rangel, Rodolfo Ribas Júnior, and Marcos Aguiar De Souza (2018). "State of the art of job satisfaction measures: A systematic review". In: *Trends in Psychology* 26.2, pp. 987–1002. ISSN: 23581883. DOI: 10.9788/TP2018.2-16En.

Horguelin, Paul A. and Louise Brunette (1998). *Pratique de la révision*. Brossard. Quebec.

Horváth, Péter I. (2009). *Revision competence*.

House, Robert J. and Lawrence A. Wigdor (1967). "Herzberg's Dual-Factor Theory of Job Satisfaction and Motivation: A Review of the Evidence and a Criticism". In: *Personnel Psychology* 20.4, pp. 369–390. ISSN: 17446570. DOI: 10.1111/j.1744-6570.1967.tb02440.x.

Hu, Ke and Patrick Cadwell (2016). "A Comparative Study of Post-editing Guidelines". In: *Baltic J. Modern Computing* 4.2, pp. 346–353. ISSN: 2255-8942. DOI: 10.13140/RG.2.1.2253.1446.

Huang, Jin (2016). "Working Styles of Student Translators in Revision and Post-editing: an Empirical-Experimental Study with Eye-tracking, Keylogging and Cue-based Retrospection". PhD thesis. Durham University.

— (2018). "Working styles of student translators in self-revision, other-revision and post-editing". In: *Eye tracking and Multidisciplinary Studies on Translation*. Ed. by Callum Walker and Federico M Federici. John Benjamins Publishing Company. Chap. 8, pp. 145–184. DOI: 10.1075/btl.143.08hua.

Hubley, Anita M and Bruno D Zumbo (2013). "Psychometric characteristics of assessment procedures: An overview". In: ISSN: 1433812290.

Hubscher-Davidson, Séverine (2016). "Trait emotional intelligence and translation: A study of professional translators". In: *Target. International Journal of Translation Studies* 28.1, pp. 132–157. ISSN: 1569-9986. DOI: 10.1075/target.28.1.06hub.

— (2018). *Translation and emotion: a psychological perspective*. New York and London: Routledge, pp. 1–235. ISBN: 9780333227794. DOI: 10.4324/9781315720388.

Hulin, Charles L. and Patricia C. Smith (1967). "An Empirical Investigation of Two Implications of the Two-Factor Theory of Job Satisfaction". In: *Journal of Applied Psychology* 51.5, pp. 396–402. ISSN: 00219010. DOI: 10.1037/h0025094.

Humphrey, Stephen E., Jennifer D. Nahrgang, and Frederick P. Morgeson (2007). "Integrating Motivational, Social, and Contextual Work Design Features: A Meta-Analytic Summary and Theoretical Extension of the Work Design Literature". In: *Journal of Applied Psychology* 92.5, pp. 1332–1356. ISSN: 00219010. DOI: 10.1037/0021-9010.92.5.1332.

Humphrys, Peter and Gordon O'Brien (1986). "The relationship between skill utilization, professional orientation and job satisfaction for pharmacists". In: *Journal of Occupational Psychology* 59.4, pp. 315–326.

Hunter, Albert and John D. Brewer (2015). "Designing multimethod research". In: *The Oxford Handbook of Multimethod and Mixed Methods Research Inquiry*. Ed. by Sharlene Nagy Hesse-Biber and R. Burke Johnson. Oxford University Press.

Hutchins, John (2003). "ALPAC: the (in)famous report". In: *Readings in machine translation* 14, pp. 131–135.

Hutchins, John W. (2010). "Machine translation : a concise history". In: *Journal of Translation Studies. Special Issue: The teaching of computer-aided translation*. 13.1 & 2, pp. 29–70.

Hutchins, William John (1986). *Machine translation: past, present, future*. Ellis Horwood Chichester.

Hvelplund, Kristian Tangsgaard (2014). "Eye tracking and the translation process: Reflections on the analysis and interpretation of eye-tracking data". In.

Inghilleri, Moira (2005). "The sociology of Bourdieu and the construction of the 'object' in translation and interpreting studies". In: *Translator* 11.2, pp. 125–145. ISSN: 17570409. DOI: 10.1080/13556509.2005.10799195.

International Organization for Standardization (2015). *Translation services-Requirements for translation services ISO 17100:2015 (E)*. Tech. rep. Geneva.

Ipsen, A. Helene and Helle V. Dam (2016). "Translation Revision: Correlating revision procedure and error detection". In: *Hermes (Denmark)* 0.55, pp. 143–156. ISSN: 19031785. DOI: 10.7146/hjlcb.v0i55.24612.

Ironson, G. H. et al. (1989). "Construction of a Job in General Scale: A Comparison of Global, Composite, and Specific Measures". In: *Journal of Applied Psychology* 74.2, pp. 193–200. DOI: 10.1037/0021-9010.74.2.193.

ISO (2017). *ISO 18587:2017 (fr)*.

Jääskeläinen, Riitta (2012). "Translation psychology". In: *Handbook of translation studies* 3, pp. 191–197.

Jakobsen, Arnt Lykke (2019). "Moving Translation, Revision, and Post-editing Boundaries". English. In: *Moving Boundaries in Translation Studies*. Ed. by Helle V. Dam, Matilde Nisbeth Brøgger, and Karen Korning Zethsen. Routledge, pp. 64–80. ISBN: 9781138563650.

JDI (2009). *Job Descriptive Index (abridged form)*. Tech. rep. Bowling Green State University.

Jia, Yanfang, Michael Carl, and Xiangling Wang (2019a). "How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study". In: *Journal of Specialised Translation* 31, pp. 60–86. ISSN: 1740357X.

— (2019b). "Post-editing neural machine translation versus phrase-based machine translation for English–Chinese". In: *Machine Translation* 33.1-2, pp. 9–29.

Jiménez Ivars, Amparo, Daniel Pinazo Catalayudb, and Marta Ruiz i Forés (2014). "Self-efficacy and language proficiency in interpreter trainees". In: *Interpreter and Translator Trainer* 8.2, pp. 167–182. ISSN: 17570417. DOI: 10.1080/1750399X.2014.908552.

Judge, Timothy A. and Joyce E. Bono (2001). "Relationship of core self-evaluations traits - Self-esteem, generalized self-efficacy, locus of control, and emotional stability - With job

satisfaction and job performance: A meta-analysis". In: *Journal of Applied Psychology* 86.1, pp. 80–92. ISSN: 00219010. DOI: 10.1037/0021-9010.86.1.80.

Judge, Timothy A, Joyce E Bono, and Edwin A Locke (2000). "Personality and job satisfaction: The mediating role of job characteristics." In: *Journal of applied psychology* 85.2, p. 237.

Judge, Timothy A., Amir Erez, and Joyce E. Bono (1998). "The power of being positive: The relation between positive self-concept and job performance". In: *Human Performance* 11.2-3, pp. 167–187. ISSN: 08959285. DOI: 10.1207/s15327043hup1102&3_4.

Judge, Timothy A. and John D. Kammeyer-Mueller (2012). "Job attitudes". In: *Annual Review of Psychology* 63, pp. 341–367. ISSN: 00664308. DOI: 10.1146/annurev-psych-120710-100511.

Judge, Timothy A. and Ryan Klinger (2008). "Job Satisfaction: Subjective Well-Being at Work". In: *The Science of Subjective Well-Being*. Ed. by Michael Eid and Randy J. Larsen. The Guilfors Press. Chap. 19, pp. 393–413. ISBN: 9781593856526.

Judge, Timothy A., Shuxia (Carrie) Zhang, and David R. Glerum (2021). "Job satisfaction". In: *Essentials of Job Attitudes and Other Workplace Psychological Constructs*. Ed. by Valerie I. Sessa and Nathan A. Bowling. Routledge. Chap. 9, pp. 207–241. DOI: 10.47760/ijcsmc.2020.v09i09.003.

Judge, Timothy A. et al. (2001). "The Job Satisfaction-Job Performance Relationship: A Qualitative and Quantitative Review". In: *Psychological Bulletin* 127.3, pp. 376–407.

Judge, Timothy A et al. (2017). "Job Attitudes, Job Satisfaction, and Job Affect: A Century of Continuity and of Change". In: *Journal of Applied Psychology* 102.3, pp. 356–374. ISSN: 0021-9010. DOI: 10.1037/apl0000181.supp.

Jurgensen, Clifford E. (1978). "Job preferences (What makes a job good or bad?)" In: *Journal of Applied Psychology* 63.3, pp. 267–276. ISSN: 0021-9010. DOI: 10.1037//0021-9010.63.3.267.

Katan, David (2009). "Translation Theory and Professional Practice: A Global Survey of the Great Divide". In: *Hermes* 42.42, pp. 111–153. ISSN: 09041699. DOI: 10.7146/HJLCB.V22I42.96849.

— (2011). "Occupation or profession: A survey of the translators' world". In: *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association* 4.2, pp. 187–209. DOI: 10.1075/bct.32.06kat.

Kato-Nitta, Naoko and Tadahiko Maeda (2013). "The job satisfaction-job performance relationship for creative tasks: An empirical investigation of the role of attitude and behavior in job performance among scientists". In: *Japanese Journal of Administrative Science* 26.3, pp. 201–214. ISSN: 0914-5206. DOI: 10.5651/jaas.26.201.

Kenny, Dorothy (2022). "Human and machine translation". In: *Machine translation for everyone: Empowering users in the age of artificial intelligence*. Berlin: Language Science Press, pp. 23–50.

Kim, Hyosu and Dennis Kao (2014). "A meta-analysis of turnover intention predictors among U.S. child welfare workers". In: *Children and Youth Services Review* 47.P3, pp. 214–223. ISSN: 01907409. DOI: 10.1016/J.CHILDYOUTH.2014.09.015.

Kim, Jay S (1980). "Relationships of Personality to Perceptual and Behavioral Responses in Stimulating and Nonstimulating Tasks". In: *Academy of Management Journal* 23.2, pp. 307–319. ISSN: 0001-4273. DOI: 10.5465/255433.

King, Patrick (2020). "Small and medium-sized enterprise (SME) translation service provider as technology user. Translation in New Zealand". In: *The Routledge Handbook of Translation and Technology*. Ed. by Minako O'hagan. Routledge. Chap. 9, pp. 148–165.

Kinnunen, Tuija and Kaisa Koskinen (2010). *Translators' agency*. Tampere University Press.

Klassen, Robert M. and Ming Ming Chiu (2010). "Effects on Teachers' Self-Efficacy and Job Satisfaction: Teacher Gender, Years of Experience, and Job Stress". In: *Journal of Educational Psychology* 102.3, pp. 741–756. ISSN: 00220663. DOI: 10.1037/a0019237.

Klein, Guillaume et al. (2017). "OpenNMT: Open-source toolkit for neural machine translation". In: *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations*, pp. 67–72. ISBN: 9781945626715. DOI: 10.18653/v1/P17-4012. arXiv: 1709.03815.

Klubička, Filip, Antonio Toral, and Víctor M Sánchez-Cartagena (2017). "Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation". In: *The Prague Bulletin of Mathematical Linguistics*, pp. 121–132. ISSN: 1804-0462. DOI: 10.1515/pralin-2017-0014. arXiv: 1701.02901.

Knowles, Rebecca and Philipp Koehn (2016). "Neural Interactive Translation Prediction". In: *Proceedings - AMTA 2016: 12th Conference of the Association for Machine Translation in the Americas*. Vol. 1, pp. 107–120. ISBN: 9780000000002.

Knowles, Rebecca, Marina Sanchez-Torron, and Philipp Koehn (2019). "A user study of neural interactive translation prediction". In: *Machine Translation* 33.1-2, pp. 135–154. ISSN: 15730573. DOI: 10.1007/s10590-019-09235-8.

Koby, Geoffrey S (2001). "Editor's introduction – Post-editing of machine translation output: Who, what, why, and how (much)". In: *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. The Kent State University Press, pp. 1–24.

Koehn, Philipp (2009). "A process study of computer-aided translation". In: *Machine Translation* 23.4, pp. 241–263. ISSN: 09226567. DOI: 10.1007/s10590-010-9076-3.

Koehn, Philipp and Ulrich Germann (2014). "The impact of machine translation quality on human post-editing". In: *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, pp. 38–46.

Koehn, Philipp and Rebecca Knowles (2017). "Six Challenges for Neural Machine Translation". In: *Proceedings of the First Workshop on Neural Machine Translation. Association for Computational Linguistics (ACL)*. Vancouver, Canada, pp. 28–39. DOI: 10.18653/v1/w17-3204. arXiv: 1706.03872.

Koehn, Philipp et al. (2007). "Moses: Open Source Toolkit for Statistical Machine Translation". In: *Proceedings of the 45th Annual Meeting of the ACL 2007 - Demo and Poster Sessions*, pp. 177–180. ISBN: 0736-587X. DOI: 10.3115/1557769.1557821.

Konttinen, Kalle (2021). "A self-efficacy scale for measuring student progress in translation company simulations". In: *Across Languages and Cultures* 22.1, pp. 64–81. ISSN: 15882519. DOI: 10.1556/084.2021.00004.

Konttinen, Kalle (2022). "Monitoring student self-efficacy in a translation company simulation. A multilevel model approach". In: *trans-kom*, pp. 1–25. ISSN: 1867-4844.

Konttinen, Kalle, Leena Salmi, and Maarit Koponen (2021). "Revision and post-editing competences in translator education". In: *Translation Revision and Post-editing: Industry Practices and Cognitive Processes*. Ed. by Maarit Koponen et al. Routledge, pp. 187–202. ISBN: 9781000201512. DOI: 10.4324/9781003096962-15.

Koponen, Maarit (2012). "Comparing human perceptions of post-editing effort with post-editing operations". In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pp. 181–190.

— (2016). "Is machine translation post-editing worth the effort? A survey of research into post-editing and effort". In: *Journal of Specialised Translation* 25, pp. 131–148. ISSN: 1740357X.

Koponen, Maarit and Leena Salmi (2015). "On the correctness of machine translation: A machine translation post-editing task". In: *Journal of Specialised Translation* 23.23, pp. 118–136. ISSN: 1740357X.

— (2017). "Post-editing quality: Analysing the correctness and necessity of post-editor corrections". In: *Linguistica Antverpiensia* 16, pp. 137–148. ISSN: 03042294.

Koponen, Maarit, Leena Salmi, and Markku Nikulin (2019). "A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output". In: *Machine Translation*, pp. 61–90. ISSN: 15730573. DOI: 10.1007/s10590-019-09228-7.

Koponen, Maarit et al. (2012). "Post-editing time as a measure of cognitive effort". In: *AMTA Workshop on Post-editing Technology and Practice*. Vol. 47. 3, pp. 11–20.

Koponen, Maarit et al., eds. (2021a). *Translation Revision and Post-editing: Industry Practices and Cognitive Processes*. London and New York: Routledge, p. 283. DOI: 10.4324/9781003096962.

— (2021b). "Introduction". In: *Translation Revision and Post-Editing. Industry Practices and Cognitive Processes*. Ed. by Maarit Koponen et al. Routledge. ISBN: 9788578110796.

Korpal, Paweł (2021). "Stress experienced by Polish sworn translators and interpreters". In: *Perspectives: Studies in Translation Theory and Practice* 29.4, pp. 554–571. ISSN: 17476623. DOI: 10.1080/0907676X.2021.1889004.

Krings, Hans P. (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Ed. by Geoffrey Koby. Kent, Ohio & London: The Kent State University Press, p. 558.

Kruger, Haidee (2017). "The effects of editorial intervention. Implications for studies of the features of translated language". In: *Empirical Translation Studies*, pp. 113–156. DOI: 10.1515/9783110459586-005.

Künzli, Alexander (2005). "What Principles Guide Translation Revision? A Combined Product and Process Study". In: *Translation Norms: What is 'normal' in the translation profession? Proceedings of the 4th Translation Conference, University of Portsmouth , November 2004*. Ed. by I. Kemble. Portsmouth, pp. 31–44.

— (2006). "Teaching and Learning Translation Revision: Some Suggestions based on Evidence from a Think-Aloud Protocolo Study". In: *Current trends in translation teaching and learning*. Ed. by Mike Garant. Helsinki: Helsinki University, pp. 9–24.

— (2007). "Translation revision: a study of the performance of ten professional translators revising a legal text". In: *Doubts and directions. Selected contributions from the 4th EST Congress, Lisbon, 2004*. Ed. by Y. Gambier, M. Shlesinger, and R. Stolze. Amsterdam: Benjamins, pp. 115–126.

— (2009). "Qualität in der Übersetzungsrevision–eine empirische Studie". In: *Translation zwischen Text und Welt: Translationswissenschaft als historische Disziplin zwischen Moderne und Zukunft*, pp. 291–303.

— (2014). "Die Übersetzungsrevision – Begriffsklärungen, Forschungsstand, Forschungsdesiderate". In: *trans-kom*.

Kuznik, Anna and Joan Miquel Verd (2017). "Investigating Real Work Situations in Translation Agencies. Work Content and Its Components". In: *HERMES - Journal of Language and Communication in Business* 23.44, pp. 25–43. ISSN: 0904-1699. DOI: 10.7146/hjlcb.v23i44.97263.

Lacruz, Isabel, Michael Denkowski, and Alon Lavie (2014). "Cognitive Demand and Cognitive Effort in Post-Editing". In: *Proceedings of the AMTA 2014 Workshop on Post-Editing Technology and Practice*, pp. 73–84.

Lacruz, Isabel and Gregory M. Shreve (2014). "Pauses and Cognitive Effort". In: *Post-editing of Machine Translation: Processes and Applications*. Ed. by Sharon O'Brian et al. Newcastle upon Tyne: Cambridge Scholars Publishing. Chap. 11, pp. 246–272.

Laflamme, Caroline (2009). "Les modifications lexicales apportées par les réviseurs professionnels dans leur tâche de révision: du problème à la solution". PhD thesis.

Lagoudaki, Elina (2008). "The Value of Machine Translation for the Professional Translator". In: *8th AMTA Conference* October, pp. 21–25.

Lambert, Eric G. et al. (2018). "Job Stress, Job Involvement, Job Satisfaction, and Organizational Commitment and Their Associations with Job Burnout Among Indian Police Officers: a Research Note". In: *Journal of Police and Criminal Psychology* 33.2, pp. 85–99. ISSN: 19366469. DOI: 10.1007/s11896-017-9236-y.

Landis, J. Richard and Gary G. Koch (1977). "The Measurement of Observer Agreement for Categorical Data". In: *Biometrics* 33.1, p. 159. ISSN: 0006341X. DOI: 10.2307/2529310.

Läubli, Samuel and David Orrego-Carmona (2017). "When Google Translate is better than Some Human Colleagues , those People are no longer Colleagues". In: *Proceedings of the 39th Conference Translating and the Computer*. November, pp. 59–69.

Läubli, Samuel, Rico Sennrich, and Martin Volk (2018). "Has machine translation achieved human parity? A Case for Document-level Evaluation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pp. 4791–4796. ISBN: 9781948087841. DOI: 10.18653/v1/d18-1512. arXiv: 1808.07048.

Läubli, Samuel et al. (2013). "Assessing post-editing efficiency in a realistic translation environment". In: *MT Summit XIV Workshop on Post-editing Technology and Practice* September, pp. 83–91. DOI: 10.5167/uzh-80891.

Läubli, Samuel et al. (2019). "Post-editing Productivity with Neural Machine Translation: An Empirical Assessment of Speed and Quality in the Banking and Finance Domain". In: *Proceedings of Machine Translation Summit XVII: Research Track*. Dublin, Ireland, pp. 267–272. arXiv: 1906.01685.

Läubli, Samuel et al. (2020a). "A Set of Recommendations for Assessing Human – Machine Parity in Language Translation". In: *Journal of Artificial Intelligence Research* 67, pp. 653–672.

Läubli, Samuel et al. (2020b). *The Impact of Text Presentation on Translator Performance*. Tech. rep. arXiv: 2011.05978v1.

Laurian, Anne-Marie (1984). "Machine Translation: What type of post-editing on what type of documents for what type of users". In: *Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeeting on Association for Computational Linguistics*, pp. 236–238. ISBN: 9788578110796. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

Lazar, Jonathan, Jinjuan Heidi Jh Feng, and Harry Hochheiser (2017). *Research methods in human-computer interaction*. Morgan Kaufmann. ISBN: 9780470723371. DOI: 10.1002/asi.21187. arXiv: arXiv:1011.1669v3.

LeBlanc, Matthieu (2013). "Translators on translation memory (TM). Results of an ethnographic study in three translation services and agencies". In: *The International Journal for Translation and Interpreting Research* 5.2, pp. 1–13. ISSN: 18369324. DOI: 105202.2013.a01.

— (2017). "'I can't get no satisfaction!' Should we blame translation technologies or shifting business practices?" In: *Human issues in translation technology*, pp. 45–62.

Lee, Hyang (2006). "Révision: Définitions et paramètres". In: *Meta* 51.2, pp. 410–419. DOI: 10.7202/013265AR.

Lee, Jieun (2017). "Professional interpreters' job satisfaction and relevant factors". In: *Translation and Interpreting Studies* 12.3, pp. 427–448. ISSN: 1932-2798. DOI: 10.1075/tis.12.3.04lee.

Lee, Katherine et al. (2018). "Hallucinations in Neural Machine Translation". In: *Conference on Neural Information Processing Systems (NeurIPS 2018)*. Montreal, Canada, pp. 1–18.

Lee, Sang Bin (2014). "An Interpreting Self-Efficacy (ISE) scale for undergraduate students majoring in consecutive interpreting: Construction and preliminary validation". In: *Interpreter and Translator Trainer* 8.2, pp. 183–203. ISSN: 17570417. DOI: 10.1080/1750399X.2014.929372.

Lefer, Marie-Aude (2020). "Parallel Corpora". In: *A Practical Handbook of Corpus Linguistics*. Springer, pp. 257–282. DOI: 10.1007/978-3-030-46216-1_12.

Lefer, Marie-Aude, Justine Piette, and Romane Bodart (2022). *MTPEAS manual - Machine Translation Post-Editing Annotation System*.

Leijten, Mariëlle and Luuk Van Waes (2013). "Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes". In: *Written Communication* 30.3,

pp. 358–392. ISSN: 07410883. DOI: 10.1177/0741088313491692/ASSET/IMAGES/LARGE/10.1177_0741088313491692-FIG2.JPEG.

Lepold, Angelika et al. (2018). "The Efficient Measurement of Job Satisfaction: Facet-Items versus Facet Scales". In: *International Journal of Environmental Research and Public Health* 15.7. DOI: 10.3390/IJERPH15071362.

Levenshtein, Vladimir (1966). "Binary codes capable of correcting deletions, insertions, and reversals". In: *Soviet Physics Doklady* 10.8, pp. 707 –710.

Li, Ming, Bo Wen, and Chih Wei Hsieh (2021). "Understanding the role reward types play in linking public service motivation to task satisfaction: evidence from an experiment in China". In: *International Public Management Journal* 25.2, pp. 300–319. ISSN: 10967494. DOI: 10.1080/10967494.2021.1963891.

Liang, Yaqing, Yanzhi Li, and Zhonggang Sang (2023). "A Study on Peer Mediation in Dynamic Assessment of Translation Revision Competence". In: *Language Assessment Quarterly* 20.1, pp. 108–126.

Light, Richard J. (1971). "Measures of response agreement for qualitative data: Some generalizations and alternatives". In: *Psychological Bulletin* 76.5, pp. 365–377. ISSN: 00332909. DOI: 10.1037/h0031643.

Liu, Christy (2013). "A quantitative enquiry into the translator's job-related happiness: Does visibility correlate with happiness?" In: *Across Languages and Cultures* 14.1, pp. 123–147. ISSN: 15851923. DOI: 10.1556/Acr.14.2013.1.6.

Liu, Jing (2021). "Correlating self-efficacy with self-assessment in an undergraduate interpreting classroom: How accurate can students be?" In: *Porta Linguarum* 36, pp. 9–25. ISSN: 16977467. DOI: 10.30827/PORTALIN.V0I36.13897.

Liu, Peng and Zhizhong Li (2012). "Task complexity : A review and conceptualization framework". In: *International Journal of Industrial Ergonomics* 42.6, pp. 553–568. ISSN: 0169-8141. DOI: 10.1016/j.ergon.2012.09.001.

Locke, Edwin A. (1969). "What is job satisfaction?" In: *Organizational Behavior and Human Performance* 4.4, pp. 309–336. ISSN: 00305073. DOI: 10.1016/0030-5073(69)90013-0.

Loher, Brian T. et al. (1985). "A Meta-Analysis of the Relation of Job Characteristics to Job Satisfaction". In: *Journal of Applied Psychology* 70.2, pp. 280–289. ISSN: 00219010. DOI: 10.1037/0021-9010.70.2.280.

Lommel, Arle (2018). "Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies". In: Springer, Cham, pp. 109–127. DOI: 10.1007/978-3-319-91241-7_6.

Loo, Robert (2002). "A caveat on using single-item versus multiple-item scales". In: *Journal of Managerial Psychology* 17.1, pp. 68–75. DOI: 10.1108/02683940210415933.

Loock, Rudy (2018). "Traduction automatique et usage linguistique : Une analyse de traductions anglais-français réunies en corpus". In: *Meta* 63.3, pp. 786–806. ISSN: 00260452. DOI: 10.7202/1060173ar.

Lu, Hong, Yang Zhao, and Alison While (2012). "Job satisfaction among hospital nurses: A literature review". In: *International Journal of Nursing Studies* 94, pp. 21–31. DOI: 10.1016/j.ijnurstu.2019.01.011.

Luo, Jinru and Dechao Li (2022). "Universals in machine translation? A corpus-based study of Chinese-English translations by WeChat Translate". In: *International Journal of Corpus Linguistics* 27.1, pp. 31–58. ISSN: 1384-6655. DOI: 10.1075/ijcl.19127.luo.

Macken, Lieve, Daniel Prou, and Arda Tezcan (2020). "Quantifying the Effect of Machine Translation in a High-Quality Human Translation Production Process". In: *Informatics* 7.2. DOI: 10.3390/informatics7020012.

Macken, Lieve et al. (2022). "Literary translation as a three-stage process: machine translation, post-editing and revision". In: *EAMT 2022 - Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, pp. 101–110. ISBN: 9789464597622.

Macnamara, Brooke (2012). "Interpreter Cognitive Aptitudes". In: *Journal of Interpretation* 19.1, p. 1. ISSN: 0882-7893.

Malmkjær, Kirsten (2020). *Translation and creativity*, pp. 1–130. ISBN: 9781315648958. DOI: 10.4324/9781315648958.

Mark, Melvin M. (2015). "Mixed and multimethods in predominantly quantitative studies, especially experiments and quasi-experiments". In: *The Oxford Handbook of Multimethod and Mixed Methods Research Inquiry*. Ed. by Sharlene Nagy Hesse-Biber and R. Burke Johnson. Oxford University Press.

Markovic, Milan and Gabriele Plickert (2018). "Attorneys' career dissatisfaction in the new normal". In: *International Journal of the Legal Profession* 25.2, pp. 147–173. ISSN: 14699257. DOI: 10.1080/09695958.2018.1456437.

Marley, Robert and Lee Kersting (2020). "Task satisfaction: the effect of relative performance information versus outcome feedback". In: *International Journal of Productivity and Performance Management* 70.8, pp. 2281–2293. ISSN: 17410401. DOI: 10.1108/IJPPM-08-2019-0378.

Marshman, Elizabeth (2012). "In the Driver's Seat: Perceptions of Control as Indicators of Language Professionals' Satisfaction with Technologies in the Workplace". In: *Translating and the Computer 34*, pp. 1–12.

— (2014). "Taking control: Language professionals and their perception of control when using language technologies". In: *Meta* 59.2, pp. 380–405. ISSN: 00260452. DOI: 10.7202/1027481ar.

Martikainen, Hanna and Natalie Kübler (2016). "Ergonomie cognitive de la post-édition de traduction automatique : enjeux pour la qualité des traductions". In: *Http://Ilcea.Revues.Org* 27. ISSN: 2101-0609.

Martikainen, Liisa, Petri Karkkola, and Matti Kuittinen (2018). "Encountering Change : Job Satisfaction of Sign Language Interpreters in Finland". In: *International Journal of Interpreter Education* 10.2, pp. 43–57.

Martin, Tim (2007). "Managing risks and resources: a down-to-earth view of revision". In: *Journal of Specialised Translation* 8, p. 7.

Maruf, Sameen, Fahimeh Saleh, and Gholamreza Haffari (2021). "A survey on document-level neural machine translation: Methods and evaluation". In: *ACM Computing Surveys (CSUR)* 54.2, pp. 1–36.

Mashhady, Habibollah, Moslem Fatollahi, and Maryam Pourgalavi (2015). "Self-efficacy and Prediction of Note-taking Inclination among Undergraduate Translation Students". In: *Theory and Practice in Language Studies* 5.11, p. 2366. ISSN: 1799-2591. DOI: 10.17507/tpls.0511.22.

Massardo, Isabella et al. (2016). *MT Post-Editing Guidelines*. Tech. rep. TAUS.

Mc Donough Dolmaya, Julie (2015). "Revision history: Translation trends in Wikipedia". In: *Translation Studies* 8.1, pp. 16–34. ISSN: 17512921. DOI: 10.1080/14781700.2014.943279.

McElhaney, Terrence and Muriel Vasconcellos (1988). "The Translator and the Postediting Experience". In: *Technology as Translation Strategy*. Ed. by Muriel Vasconcellos. American T, pp. 140–148. DOI: 10.4324/9781315679624-16.

McEnery, Tony and Andrew Hardie (2012). *Corpus linguistics: Method, Theory and Practice*. Cambridge University Press. ISBN: 978-0-521-83851-1.

Mellinger, Christopher D. and Gregory M. Shreve (2016). "Match evaluation and over-editing in a translation memory environment". In: pp. 131–148. ISBN: 9789027266347. DOI: 10.1075/btl.128.07mel.

Mellinger, Christopher D et al. (2017). *Quantitative Methods in Translation and Interpreting Studies*. ISBN: 9781138124950.

Mesa-Lao, Bartolomé (2014). "Gaze Behaviour On Source Texts: An Exploratory Study Comparing Translation and Post-Editing". In: *Post-editing of Machine Translation: Processes and Applications*, pp. 219–245.

Mitchell, Linda, Johann Roturier, and Sharon O'Brien (2013). "Community-based post-editing of machine-translated content: monolingual vs. bilingual". In: *Proceedings of the 2nd Workshop on Post-editing Technology and Practice*.

Miyoshi, Akiko (2012). "The stability and causal effects of task-specific and generalized self-efficacy in college". In: *Japanese Psychological Research* 54.2, pp. 150–158. ISSN: 00215368. DOI: 10.1111/J.1468-5884.2011.00481.X.

Moe, Marija Zlatnar, Tamara Mikolic Juznic, and Tanja Zigon (2021). "Who determines the final version? The roles of translators, language revisers and editors in the publishing of a literary translation". In: *Across Languages and Cultures* 22.1, pp. 14–44. ISSN: 15882519. DOI: 10.1556/084.2021.00002.

Mohammadi Dehcheshmeh, Maryam (2017). "Machine Translation and Translation Memory Systems: An Ethnographic Study of Translators' Satisfaction". PhD thesis. University of Ottawa.

Monahan, John and Jeffrey Swanson (2019). "Lawyers at the Peak of Their Careers: A 30-Year Longitudinal Study of Job and Life Satisfaction". In: *Journal of Empirical Legal Studies* 16.1, pp. 4–25. ISSN: 17401461. DOI: 10.1111/jels.12207.

Moorkens, Joss (2018). "What to expect from Neural Machine Translation: a practical in-class translation evaluation exercise". In: *The Interpreter and Translator Trainer* 12.4, pp. 1–13. ISSN: 1750-399X. DOI: 10.1080/1750399X.2018.1501639.

Moorkens, Joss (2020). "Comparative satisfaction among freelance and directly-employed Irish-language translators". In: *Translation and Interpreting* 12.1, pp. 55–73. ISSN: 18369324. DOI: 10.12807/TI.112201.2020.A04.

Moorkens, Joss and Sharon O'Brien (2015). "Post-Editing Evaluations : Trade-offs between Novice and Professional Participants". In: *The 18th Annual Conference of the European Association for Machine Translation (EAMT 2015)*, pp. 75–81.

Moorkens, Joss and Marta Rocchi (2021). "Ethics in the translation industry". In: *The Routledge Handbook of Translation and Ethics* January, pp. 320–337. DOI: 10.4324/9781003127970-24.

Moorkens, Joss and Ryoko Sasamoto (2017). "Productivity and Lexical Pragmatic Features in a Contemporary CAT Environment : An Exploratory Study in English to Japanese". In: *Hermes* 56, pp. 111–123.

Moorkens, Joss et al. (2015). "Correlations of perceived post-editing effort with measurements of actual effort". In: *Machine Translation* 29.3-4, pp. 267–284. ISSN: 15730573. DOI: 10.1007/s10590-015-9175-2.

Moorkens, Joss et al. (2018). *Translation Quality Assessment. From Principles to Practice.* ISBN: 9783319912400.

Moran, John, David Lewis, and Christian Saam (2014). "Analysis of Post-editing Data: A Productivity Field Test using an Instrumented CAT Tool John". In: *Post-editing of Machine Translation: Processes and Applications*. Ed. by Sharon O'Brien et al. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 126–146. ISBN: 2013206534.

Morgeson, Frederick P., Michael T. Brannick, and Edward L. Levine (2019). *Job and Work Analysis: Methods, Research, and Applications for Human Resource Management*. 3rd. SAGE Publications. ISBN: 9781544329529. DOI: 10.4135/9781071872536.

Morgeson, Frederick P., Kelly Delaney-Klinger, and Monica A. Hemingway (2005). "The importance of job autonomy, cognitive ability, and job-related skill for predicting role breadth and job performance". In: *Journal of Applied Psychology* 90.2, pp. 399–406. ISSN: 00219010. DOI: 10.1037/0021-9010.90.2.399.

Morgeson, Frederick P. and Stephen E. Humphrey (2006). "The Work Design Questionnaire (WDQ): Developing and validating a comprehensive measure for assessing job design and the nature of work". In: *Journal of Applied Psychology* 91.6, pp. 1321–1339. ISSN: 00219010. DOI: 10.1037/0021-9010.91.6.1321.

Morrison, David et al. (2005). "Job design, opportunities for skill utilization, and intrinsic job satisfaction". In: *European journal of work and organizational psychology* 14.1, pp. 59–79.

Mossop, Brian (2020). *Revising and Editing for Translators*. Routledge. ISBN: 9788578110796. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

— (2023). "Reading order during bilingual quality checking of translations: An issue in search of studies". In: *Ampersand* 11, p. 100138.

Muñoz Martín, Ricardo (2014). "Situating translation expertise. A review with a sketch of a construct". In: *The Development of Translation Competence : Theories and Methodologies from Psycholinguistics and Cognitive Science*. Ed. by John W. Schwieter and Aline Ferreira. Cambridge Scholars Publishing, pp. 2–56. ISBN: 9781443854504.

Mutal, Jonathan et al. (2019). "Differences between SMT and NMT Output – a Translators' Point of View". In: *The Second Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, pp. 75–81. DOI: 10.26615/issn.2683-0078.2019_009.

Nagy, Mark S. (2002). "Using a single-item approach to measure facet job satisfaction". In: *Journal of Occupational and Organizational Psychology* 75.1, pp. 77–86. ISSN: 09631798. DOI: 10.1348/096317902167658.

National Research Council (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". In.

Neishi, Masato and Naoki Yoshinaga (2019). "On the relation between position information and sentence length in neural machine translation". In: *CoNLL 2019 - 23rd Conference on Computational Natural Language Learning, Proceedings of the Conference*, pp. 328–338. ISBN: 9781950737727. DOI: 10.18653/v1/k19-1031.

Nitzke, Jean (2016). "Monolingual post-editing: An exploratory study on research behaviour and target text quality". In: *Eyetracking and Applied Linguistics (Translation and Multilingual Natural Language Processing 2)*. ISBN: 978-3-944675-98-5.

Nitzke, Jean and Anne-kathrin Gros (2021). "Preferential Changes in Revision and Post-editing". In: *Translation Revision and Post-Editing. Industry Practices and Cognitive Processes*. Ed. by Maarit Koponen et al. Routledge.

Nitzke, Jean, Silvia Hansen-Schirra, and Carmen Canfora (2019). "Risk management and post-editing competence". In: *The Journal of Specialised Translation* 31, pp. 239–259.

Nørgaard, Birgitte et al. (2012). "Communication skills training increases self-efficacy of health care professionals". In: *Journal of Continuing Education in the Health Professions* 32.2, pp. 90–97. ISSN: 08941912. DOI: 10.1002/chp.21131.

Nunes Vieira, Lucas (2018). "Automation anxiety and translators". In: *Translation Studies* 13.1, pp. 1–21. ISSN: 17512921. DOI: 10.1080/14781700.2018.1543613.

— (2020a). "Machine translation in the news". In: *Translation Spaces* 9.1, pp. 98–122. ISSN: 2211-3711. DOI: 10.1075/ts.00023.nun.

— (2020b). "Post-editing of machine translation". In: *Routledge Handbook of Translation and Technology*. Ed. by Minako O'hagan. Routledge. Chap. 19, pp. 319–332.

Nunes Vieira, Lucas and Elisa Alonso (2018). *The use of machine translation in human translation workflows: Practices, perceptions and knowledge exchange*. Tech. rep. Institute of Translation and Interpreting.

Nunziatini, Mara (2019). "Machine translation in the financial services industry: A case study". In: *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pp. 57–63.

Oates, Briony J. (2005a). "Questionnaires". In: *Researching Information Systems and Computing*.

Oates, Briony J. (2005b). "Surveys". In: *Researching Information Systems and Computing.*

O'Brien, Gordon (1982a). "Evaluation of the job characteristics theory of work attitudes and performance". In: *Australian Journal of Psychology* 34.3, pp. 383–401.

O'Brien, Gordon E (1982b). "The relative contribution of perceived skill-utilization and other perceived job attributes to the prediction of job satisfaction: A cross-validation study". In: *Human Relations* 35.3, pp. 219–237.

O'Brien, Morgan and Ana Duarte (2015). "Machine translation for enterprise technical communications – a journey of discovery". In: *Proceedings of Machine Translation Summit XV: User Track*. Vol. 2. Miami, pp. 92–105.

O'Brien, Sharon (2002). "Teaching Post-editing : A Proposal for Course Content". In: *Proceedings of the 6th EAMT Workshop, Teaching Machine Translation* November, pp. 99–106.

O'Brien, Sharon (2006a). "Methodologies for measuring the correlations between post-editing effort and machine translatability". In: *Machine Translation* 19.1, pp. 37–58. ISSN: 09226567. DOI: 10.1007/s10590-005-2467-1.

— (2006b). "Pauses as Indicators of Cognitive Effort in Post-Editing Machine Translation". In: *Across Languages and Cultures* 7.1, pp. 1–21.

O'Brien, Sharon (2009). "Eye tracking in translation process research: methodological challenges and solutions". In: *Methodology, technology and innovation in translation process research* 38, pp. 251–266.

O'Brien, Sharon (2011). "Towards predicting post-editing productivity". In: *Machine Translation* 25.3, pp. 197–215. ISSN: 09226567. DOI: 10.1007/s10590-011-9096-7.

O'Brien, Sharon; (2012). "Translation as Human-Computer Interaction". In: *Translation Spaces* 1.1, pp. 101–122.

O'Brien, Sharon (2020). "Translation, human–computer interaction and cognition". In: *The Routledge Handbook of Translation and Cognition*. Ed. by Fabio Alves and Arnt Lykke Jakobsen. London: Routledge. Chap. 21, pp. 376–388. ISBN: 9781315178127.

O'Brien, Sharon, Michel Simard, and Marie-Josée Goulet (2018). "Machine translation and self-post-editing for academic writing support: Quality explorations". In: *Translation quality assessment: From principles to practice*, pp. 237–262. ISSN: 3319912402.

O'Brien, Sharon et al. (2017). "Irritating CAT tool features that matter to translators". In: *Hermes (Denmark)* 56, pp. 145–162. ISSN: 19031785.

Oldham, Greg R and Yitzhak Fried (2016). "Job design research and theory: Past, present and future". In: *Organizational Behavior and Human Decision Processes* 136, pp. 20–35. ISSN: 07495978. DOI: 10.1016/j.obhdp.2016.05.002.

Ortiz-Martínez, Daniel et al. (2016). "Integrating Online and Active Learning in a Computer-Assisted Translation Workbench". In: *New Frontiers in Translation Studies*. Springer Science and Business Media Deutschland GmbH, pp. 57–76. DOI: 10.1007/978-3-319-20358-4_3.

Oshagbemi, Titus (1999). "Overall job satisfaction: How good are single versus multiple-item measures?" In: *Journal of Managerial Psychology* 14.5, pp. 388–403. ISSN: 02683946. DOI: 10.1108/02683949910277148.

O'Brien, Sharon (2022). "How to deal with errors in machine translation: Postediting". In: *Machine translation for everyone: Empowering users in the age of artificial intelligence*. Berlin: Language Science Press, pp. 105–120.

Papineni, Kishore et al. (2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, pp. 311–318. DOI: 10.3115/1073083.1073135.

Parra Escartín, C. et al. (2017). "Machine Translation as an Academic Writing Aid for Medical Practitioners". In: *MT Summit XVI, vol.1: Research Track* 1, pp. 254–267.

Parra Escartín, Carla and Manuel Arcedillo (2015a). "Living on the edge: productivity gain thresholds in machine translation evaluation metrics". In: *Proceedings of the 4th Workshop on Post-editing Technology and Practice*.

— (2015b). "Machine translation evaluation made fuzzier: A study on post-editing productivity and evaluation metrics in commercial settings". In: *Proceedings of Machine Translation Summit XV: Papers*.

Parra Escartín, Carla and Marie-Josée Goulet (2021). "When the Post-Editor is not a Translator". In: *Translation Revision and Post-Editing*. Ed. by Maarit Koponen et al. Routledge, pp. 89–106. DOI: 10.4324/9781003096962-8.

Parra Galiano, Silvia (2005). "La revisión de traducciones en traductología: aproximación a la práctica de la revisión en el ámbito profesional mediante el estudio de casos y propuestas de investigación". PhD thesis. ISBN: 8433834215.

— (2015). "El conocimiento experto (pericia) en la revisión de traducciones: clave en la gestión y propuestas de investigación". In: *Traducimos desde el sur: VI Congreso Internacional de la Asociación Ibérica de Estudios de Traducción e Interpretación* January 2015, pp. 587–603.

Pereira, Ariana López (2019). "Traducción automática neuronal y traducción automática estadística: Percepción y productividad". In: *Revista Tradumatica* 17.17, pp. 1–19. ISSN: 15787559. DOI: 10.5565/rev/tradumatica.235.

Pérez Macías, Lorena (2020). "What Do Translators Think About Post-Editing ? A Mixed-Methods Study Of Translators ' Fears , Worries And Preferences On Machine Translation Post-Editing". In: *Revista Tradumàtica*, pp. 0–3.

Pérez-Ortiz, Juan Antonio, Mikel L Forcada, and Felipe Sánchez-Martínez (2022). "How neural machine translation works". In: *Machine translation for everyone: Empowering users in the age of artificial intelligence*. Berlin: Language Science Press, pp. 141–164.

Peris, Álvaro, Miguel Domingo, and Francisco Casacuberta (2017). "Interactive neural machine translation". In: *Computer Speech & Language* 45, pp. 201–220.

Piecychna, Beata (2019). "Polish sworn translators' attitudes towards the job they perform - Results obtained from a job satisfaction survey". In: *Studies in Logic, Grammar and Rhetoric* 58.1, pp. 125–154. ISSN: 0860150X. DOI: 10.2478/slgr-2019-0020.

Pielmeier, Helene and Arle Lommel (2019). *Machine Translation Use at LSPs*.

Pielmeier, Hélène and Paul O'Mara (2020). *The State of the Linguist Supply Chain. Translators and Interpreters in 2020*. Tech. rep.

Plitt, Mirko and François Masselot (2010). "A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context". In: *The Prague Bulletin of Mathematical Linguistics* 93, pp. 7–16. ISSN: 1804-0462. DOI: 10.2478/v10108-010-0010-x.

Poibeau, Thierry (2017). *Machine translation*. MIT Press.

Pontrandolfo, Gianluca (2017). "La revisión de traducciones jurídicas y la evaluación de su calidad en el ámbito profesional". In: *Translation and Translanguaging in Multilingual Contexts* 3.1, pp. 114–144. ISSN: 2352-1805. DOI: 10.1075/ttmc.3.1.08pon.

Popič, Damjan (2014). "Revising translation revision in Slovenia". In: *New Horizons in Translation Research and Education 2*. Ed. by Tamara Mikolič Južnič, Kaisa Koskinen, and Nike Pokorn. Joensuu: University of Eastern Finland, pp. 72–89.

Popović, Maja (2011). "Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output". In: *The Prague Bulletin of Mathematical Linguistics* 96.1. ISSN: 0032-6585. DOI: 10.2478/v10108-011-0011-4.

Popović, Maja (2018). "Error classification and analysis for machine translation quality assessment". In: *Translation quality assessment: From principles to practice*, pp. 129–158.

Popović, Maja et al. (2014). "Relations between different types of post-editing operations, cognitive effort and temporal effort". In: *The Seventeenth Annual Conference of the European Association for Machine Translation (EAMT 14)*, pp. 191–198.

Porro Rodríguez, Victoria, Lucía Morado Vázquez, and Pierrette Bouillon (2017). "Study on the use of machine translation and post-editing in Swiss-based language service providers". In: *Parallèles* 29.2, pp. 19–35. DOI: 10.17462/para.2017.02.02.

Prioux, René and Michel Rochard (2007). "Economie de la révision dans une organisation internationale: le cas de l'OCDE". In: *Journal of Specialised Translation* 6.8, pp. 21–41. ISSN: 1740357X.

Quintana, Rocío Caro and Sheila Castilho (2022). "A review of the Integration of Machine Translation in CAT tools". In: *New Trends in Translation and Technology 2022*, pp. 214–221.

Ramírez-Sánchez, Gema (2022). "Custom machine translation". In: *Machine translation for everyone: Empowering users in the age of artificial intelligence*. Berlin: Language Science Press, pp. 165–186.

Rasmussen, Kirsten and Anne Schjoldager (2011). "Revising Translations A Survey of Revision Policies in Danish Translation Companies". In: *The Journal of Specialised Translation* 15, pp. 87–120. ISSN: 1740357X.

Rega, Lorenza (1999). "Alcune considerazioni sul problema della revisione nell'ambito della traduzione". In: *Rivista internazionale di tecnica della traduzione / International Journal of Translation* 4.

Rensburg, Alta van (2017). "Developing assessment instruments: The effect of a reviser's profile on the quality of the revision product". In: *Linguistica Antverpiensia, New Series – Themes in Translation Studies* 16, pp. 71–88. ISSN: 22955739. DOI: 10.52034/lanstts.v16i0.433.

Rico, Celia and Diana González Pastor (2022). "The role of machine translation in translation education: A thematic analysis of translator educators' beliefs". In: *Translation and Interpreting* 14.1, pp. 177–197. ISSN: 18369324. DOI: 10.12807/TI.114201.2022.A010.

Riondel, Aurélien (2021a). "Two approaches to quality in institutional settings. Comparison of the revision policies of an intergovernmental organisation and the Swiss Confederation". In: *MikaEL* 14, pp. 82–96.

— (2021b). "Revision as an emerging subfield in Translation Studies". In: *Perspectives*, pp. 1–8. ISSN: 0907-676X. DOI: 10.1080/0907676x.2021.2005457.

— (2022). "Source First or Target First ? Insight into the Order of Reading in Revision Using In-depth Interviews". In: *HERMES - Journal of Language and Communication in Business* 62, pp. 73–86.

— (2023). "Le grand écart de la révision. Étude par entretien des politiques de révision et des relations traducteurs réviseurs". PhD thesis. University of Geneva.

Risku, Hanna, Regina Rogl, and Jelena Milošević (2020). "Researching workplaces". In: *The Bloomsbury companion to language industry studies*. Bloomsbury London, pp. 37–62.

Robbins, Stephen P. and Timothy A. Judge (2022). *Essentials of organizational behavior*. Fifteenth. Pearson. ISBN: 9781292406664.

Robert, Isabelle (2008). "Translation revision procedures: An explorative study". In: *Translation and Its others: selected papers of the CETRA Seminar in translation Studies 2007*. Pp. 1–25.

— (2012). "La révision en traduction : les procédures de révision et leur impact sur le produit et le processus de révision". PhD thesis, p. 343. ISBN: 9789057283703.

— (2013). "Translation revision: does the revision procedure matter?" In: *Tracks and Treks in Translation Studies*. Ed. by Catherine Way et al., pp. 87–102. ISBN: 9027271429.

— (2016). "Should Revision Trainees Think Aloud while Revising Somebody Else's Translation? Insights from an Empirical Study with Professionals". In: *Meta LXI* 2.

Robert, Isabelle, Iris Schrijver, and Jim J. Ureel (2022). "Measuring translation revision competence and post-editing competence in translation trainees: methodological issues". In: *Perspectives: Studies in Translation Theory and Practice*, pp. 1–15. ISSN: 17476623. DOI: 10.1080/0907676X.2022.2030377.

Robert, Isabelle, Jim J. J. Ureel, and Iris Schrijver (2022). "Translation, translation revision and post-editing competence models: Where are we now?" In: *The Human Translator in the 2020s*, pp. 44–59. ISBN: 9781003223344. DOI: 10.4324/9781003223344-4.

Robert, Isabelle and Luuk Van Waes (2014). "Selecting a translation revision procedure: do common sense and statistics agree?" In: *Perspectives: Studies in Translatology* 22.3, pp. 304–320. ISSN: 0907676X. DOI: 10.1080/0907676X.2013.871047.

Robert, Isabelle et al. (2017). "Conceptualising translation revision competence: A pilot study on the 'tools and research' subcompetence". In: 28, pp. 293–316.

Robert, Isabelle S. (2018). "La recherche en révision. Portrait bibliométrique, questions de recherche et méthodologies". In: *Paralleles* 30.2, pp. 129–152. DOI: 10.17462/para.2018.02.07.

Robert, Isabelle S, Aline Remael, and Jim J.J. Ureel (2017). "Towards a model of translation revision competence". In: *Interpreter and Translator Trainer* 11.1, pp. 1–19. ISSN: 17570417. DOI: 10.1080/1750399X.2016.1198183.

Robert, Isabelle S, Iris Schrijver, and Jim JJ Ureel (2023). "Comparing L2 translation, translation revision, and post-editing competences in translation trainees: An exploratory study into Dutch–French translation". In: *Babel* 69.1, pp. 99–128.

Robin, Edina (2014a). "Explicitation and Implicitation in Revised Translations". In: *Complex Visibles Out There. Proceedings of the 2014 Olomouc Linguistic Colloquium*. Ed. by L. Veselovská and M. Janebová. Olomouc: Palacký University, pp. 559–574.

— (2014b). "Fordítási Univerzálék a Lektorált Szövegekben [Translation Universals in Revised Texts]". PhD thesis. Eötvös Loránd University.

— (2018). "A Classification of Revisional Modifications". In: *Latest Trends in Hungarian Translation Studies*. Ed. by Ildikó Horváth. Budapest: OFFI-ELTE, pp. 155–163. ISBN: 9786150013701.

— (2019). "The Modification of Translation Universals in Revised Texts". In: *Getting translated*. Ed. by Anikó Sohár, Ildikó Limpár, and Dalma Galambos. Budapest: Tinta Könyvkiadó, pp. 123–146.

Rodríguez-Castro, Mónica (2011). "Elements of Task, Job and Professional Satisfaction in the Language Industry: An Empirical Model". PhD thesis.

— (2015). "Conceptual construct and empirical validation of a multifaceted instrument for translator satisfaction". In: *Translation and Interpreting* 7.2, pp. 30–50. ISSN: 18369324. DOI: 10.12807/ti.107202.2015.a03.

— (2016). "Intrinsic and Extrinsic Sources of Translator Satisfaction: an Empirical Study". In: *Entreculturas. Revista de traducción y comunicación intercultural* 8.7-8, pp. 195–229. DOI: 10.24310/entreculturasertci.vi7-8.11334.

— (2019). "Critical distinctions between expert and novice translators: Task and professional satisfaction". In: *Hermes (Denmark)* 59, pp. 123–138. ISSN: 19031785. DOI: 10.7146/hjlcb.v59i1.117021.

Rossi, Caroline and Jean Pierre Chevrot (2019). "Uses and perceptions of machine translation at the European Commission". In: *Journal of Specialised Translation* 31, pp. 177–200. ISSN: 1740357X.

Rothwell, Andrew et al. (2023). *Translation Tools and Technologies*. Taylor & Francis.

Rubenstein, Alex L. et al. (2018). "Surveying the forest: A meta-analysis, moderator investigation, and future-oriented discussion of the antecedents of voluntary employee turnover". In: *Personnel Psychology* 71.1, pp. 23–65. ISSN: 17446570. DOI: 10.1111/peps.12226.

Ruokonen, Minna, Jannika Lassus, and Taru Virtanen (2020). "' I fulfil my place among the humankind , in the universe ' Finnish translators ' job satisfaction in three empirical studies". In: *MikaEl* 13, pp. 109–123.

Ruokonen, Minna and Jukka Mäkisalo (2018). "Middling-status profession, high-status work: Finnish translators status perceptions in the light of their backgrounds, working conditions and job satisfaction". In: *Translation and Interpreting* 10.1, pp. 1–17. ISSN: 18369324. DOI: 10.12807/ti.110201.2018.a01.

Rutherford, Brian et al. (2009). "The role of the seven dimensions of job satisfaction in salesperson's attitudes and behaviors". In: *Journal of Business Research* 62.11, pp. 1146–1151. ISSN: 01482963. DOI: 10.1016/j.jbusres.2008.10.019.

Saari, Lise M. and Timothy A. Judge (2004). "Employee attitudes and job satisfaction". In: *Human Resource Management* 43.4, pp. 395–407. ISSN: 00904848. DOI: 10.1002/hrm.20032.

Sakamoto, Akiko (2019). "Why do many translators resist post-editing? A sociological analysis using Bourdieu's concepts". In: *Journal of Specialised Translation* 31, pp. 201–216. ISSN: 1740357X.

Salancik, Gerald R. and Jeffrey Pfeffer (1978). "A Social Information Processing Approach to Job Attitudes and Task Design". In: *Administrative Science Quarterly* 23.2, pp. 224–253.

Saldanha, Gabriela and Sharon O'Brien (2013). *Research methodologies in translation studies*. Routledge. ISBN: 9781909485006. DOI: 10.1080/0907676X.2015.1109906.

Salmi, Leena (2021). "A good servant but a bad master: Finnish Translators' Perceptions on Translation Technology". In: *Revista Tradumatica* 19, pp. 112–130. ISSN: 15787559. DOI: 10.5565/rev/tradumatica.287.

Sánchez-Gijón, Pilar, Joss Moorkens, and Andy Way (2019). "Post-editing neural machine translation versus translation memory segments". In: *Machine Translation* 33.1-2, pp. 31–59. ISSN: 15730573. DOI: 10.1007/s10590-019-09232-x.

Sánchez-Torron, Marina, Koehn, and Phillipp (2016). "Machine Translation Quality and Post-Editor Productivity". In: *Proceedings of AMTA 2016*, pp. 16–26.

Sanchis-Trilles, Germán et al. (2014). "Interactive translation prediction versus conventional post-editing in practice: A study with the CasMaCat workbench". In: *Machine Translation* 28.3-4, pp. 217–235. ISSN: 09226567. DOI: 10.1007/s10590-014-9157-9.

Saragih, Susanti (2011). "The Effects of Job Autonomy on Work Outcomes:" in: *International Research Journal of Business Studies* 4.3, pp. 203–215.

Savoldi, Beatrice et al. (2021). "Gender Bias in Machine Translation". In: *Transactions of the Association for Computational Linguistics* 9, pp. 845–874. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00401.

Scansani, Randy and Lamis Mhedhbi (2020). "How do LSPs compute MT discounts? Presenting a company's pipeline and its use". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*. Lisboa, pp. 393–401.

Scarpello, Vida and John P. Campbell (1983). "Job Satisfaction: are all the parts there?" In: *Personnel Psychology* 36.3, pp. 577–600. ISSN: 17446570. DOI: 10.1111/j.1744-6570.1983.tb02236.x.

Schaefer, Falko, Joeri Van De Walle, and Joachim Van den Bogaert (2014). "Moses SMT as an Aid to Translators in the Production Process". In: *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT)*, pp. 89–92.

Schaeffer, Moritz et al. (2019). "Eye-tracking revision processes of translation students and professional translators". In: *Perspectives: Studies in Translation Theory and Practice* 27.4, pp. 589–603. ISSN: 17476623. DOI: 10.1080/0907676X.2019.1597138.

Schäfer, Falko (2003). "MT post-editing: How to shed light on the "unknown task " - Experices made at SAP". In: *EAMT-CLAW 03; joint conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop: controlled language translation*, pp. 133–140.

Schaumberg, Rebecca L and Francis J Flynn (2017). "Supplemental Material for Clarifying the Link Between Job Satisfaction and Absenteeism: The Role of Guilt Proneness". In: *Journal of Applied Psychology* 102.6, pp. 982–992. ISSN: 0021-9010. DOI: 10.1037/apl0000208.supp.

Schjoldager, Anne, Kirsten Wølch Rasmussen, and Christa Thomsen (2008). "Précis-writing, Revision and Editing: Piloting the European Master in Translation". In: *Meta* 53.4, pp. 798–813. ISSN: 0026-0452. DOI: 10.7202/019648ar.

Schnierer, Madeleine (2021). "Revision and Quality Standards". In: *Translation Revision and Post-Editing*. 2011, pp. 109–130. DOI: 10.4324/9781003096962-10.

Schwartz, Lane (2014). "Monolingual post-editing by a domain expert is highly effective for translation triage". In: *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*. Vancouver, Canada: Association for Machine Translation in the Americas, pp. 34–44. URL: https://aclanthology.org/2014.amta-wptp.3.

Schwarzer, Ralf and Matthias Jerusalem (1995). "The General Self-Efficacy Scale (GSE)". In: *Measures in health psychology: A user's portfolio. Causal and control beliefs*. Ed. by John Weinman, Stephen Wright, and Marie Johnston. Windsor, England: NFER-NELSON, pp. 35–37.

Schyns, Birgit and Gernot von Collani (2002). "A new occupational self-efficacy scale and its relation to personality constructs and organizational variables". In: *European Journal of Work and Organizational Psychology* 11.2, pp. 219–241. ISSN: 1359-432X. DOI: 10.1080/13594320244000148.

Schyns, Birgit and Marcel A. Croon (2006). "A model of task demands, social structure, and leader-member exchange and their relationship to job satisfaction". In: *International Journal of Human Resource Management* 17.4, pp. 602–615. ISSN: 09585192. DOI: 10.1080/09585190600581378.

Scocchera, Giovanna (2015). "La revisione della traduzione editoriale dall'inglese all'italiano tra ricerca accademica, professione e formazione: stato dell'arte e prospettive future". PhD thesis, pp. 1–315.

— (2017). "Translation Revision as Rereading : Different Aspects of the Translator's and Reviser's Approach to the Revision Process". In: *Mémoires du livre / Studies in Book Culture* 9.1, pp. 1–21.

— (2020). "The competent reviser: a short-term empirical study on revision teaching and revision competence acquisition". In: *Interpreter and Translator Trainer* 14.1, pp. 19–37. ISSN: 17570417. DOI: 10.1080/1750399X.2019.1639245.

Selinger, Jessica (2020). *Quo vadis traduzione automatica ?*

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2015). "Neural machine translation of rare words with subword units". In: *arXiv preprint arXiv:1508.07909*.

Serbina, Tatiana, Mario Bisiada, and Stella Neumann (2021). "Linguistic profiles of translation manuscripts and edited translations". In: *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pp. 34–45.

Setton, Robin and Alice Guo Liangliang (2009). "Attitudes to role, status and professional identity in interpreters and translators with Chinese in Shanghai and Taipei". In: *Translation and Interpreting Studies* 4.2, pp. 210–238. ISSN: 1932-2798. DOI: 10.1075/tis.4.2.05set.

SFT, Société française des traducteurs (2022). *Survey report on professional practices in translation [Rapport de l'enquête 2022 sur les pratiques professionnelles en traduction]*. Tech. rep.

Sherer, Mark et al. (1982). "The Self-Efficacy Scale: Construction and Validation". In: *Psychological Reports* 51, pp. 663–671.

Shreve, Gregory, Erik Angelone, and Isabel Lacruz (2014). "Efficacy of screen recording in the other-revision of translations: episodic memory and event models". In: *MonTI. Monografías de Traducción e Interpretación* Special Issue - Minding Translation, pp. 225–245. ISSN: 19899335. DOI: 10.6035/MonTI.2014.ne1.7.

Siefer, Katharina, Timo Leuders, and Andreas Obersteiner (2021). "Which Task Characteristics Do Students Rely on When They Evaluate Their Abilities to Solve Linear Function Tasks? – A Task-Specific Assessment of Self-Efficacy". In: *Frontiers in Psychology* 12.March, pp. 1–14. ISSN: 16641078. DOI: 10.3389/fpsyg.2021.596901.

Silva, Roberto (2014). "Integrating Post-Editing MT in a Professional Translation Workflow". In: *Post-editing of Machine Translation*. Ed. by Sharon; O'Brien et al. Newcastle upon Tyne: Cambridge Scholars Publishing. Chap. 2, pp. 24–50.

Sims, Henry P., Andrew D. Szilagyi, and Robert T. Keller (1976). "The Measurement of Job Characteristics". In: *The Academy of Management Journal* 19.2, pp. 195–212.

Skaalvik, Einar M. and Sidsel Skaalvik (2014). "Teacher Self-efficacy and Perceived Autonomy: Relations with Teacher Engagement, Job Satisfaction, and Emotional Exhaustion". In: *Psychological Reports* 114.1, pp. 68–77. ISSN: 00332941. DOI: 10.2466/14.02.PR0.114k14w0.

Skadiņš, Raivis et al. (2014). "Application of Machine Translation in Localization into Low-Resourced Languages". In: *Proceedings of the 17th Annual Conference of the European Association for Machine Translation EAMT2014*, pp. 209–216.

Smith, Patricia C., Lorne M. Kendall, and Charles L. Hulin (1969). *The Measurement of Satisfaction in Work and Retirement*. Chicago, p. 186.

Smith, Sara Anne et al. (2006). "If at First You Don't Succeed: Effects of Failure on General and Task-Specific Self-Efficacy and Performance." In: *North American journal of psychology* 8.1, pp. 171–182.

Snover, Matthew et al. (2006). "A study of translation edit rate with targeted human annotation". In: *AMTA 2006 - Proceedings of the 7th Conference of the Association for Machine Translation of the Americas: Visions for the Future of Machine Translation*, pp. 223–231.

Snover, Matthew et al. (2009). "TER-Plus : paraphrase , semantic , and alignment enhancements to Translation Edit Rate". In: *Machine Translation* 23, pp. 117–127. DOI: 10.1007/s10590-009-9062-9.

Solum, Kristina (2018). "The tacit influence of the copy-editor in literary translation". In: *Perspectives: Studies in Translation Theory and Practice* 26.4, pp. 543–559. ISSN: 17476623. DOI: 10.1080/0907676X.2018.1458888.

Somers, Harold (1997). "A practical approach to using machine translation software: 'Post-editing' the source text". In: *Translator* 3.2, pp. 193–212. ISSN: 17570409. DOI: 10.1080/13556509.1997.10798998.

Sorvali, Irma (1998). "The Translator as a Creative Being with special regard to the translation of literature and LSP". In: *Babel. Revue internationale de la traduction / International Journal of Translation* 44.3, pp. 234–243. ISSN: 0521-9744. DOI: 10.1075/babel.44.3.04sor.

Spector, Paul E. (1985). "Measurement of Human Service Staff Satisfaction: Development of the Job Satisfaction Survey". In: *American Journal of Community Psychology* 13.6, pp. 693–713. ISSN: 00910562. DOI: 10.1007/BF00929796.

Spector, Paul E (2022). *Job satisfaction : from assessment to intervention*. Routledge. ISBN: 9781003250616.

Stanton, Jeffrey M et al. (2001). "Development of a compact measure of job satisfaction: the abridged Job Descriptive Index". In: *Educational and Psychological Measurement* 61.6, pp. 1104–1122.

Taber, Tom D and George M Alliger (1995). "A task-level assessment of job satisfaction". In: *Journal of Organizational Behavior* 16.2, pp. 101–121. ISSN: 10991379. DOI: 10.1002/job.4030160202.

Tatsumi, Midori (2010). "Post-Editing Machine Translated Text in A Commercial Setting: Observation and Statistical Analysis". PhD thesis.

Teixeira, Carlos (2014a). "The Impact of Metadata on Translator Performance: How Translators Work With Translation Memories and Machine Translation". PhD thesis, p. 225. DOI: 10.13140/RG.2.1.2190.2887.

Teixeira, Carlos S C (2014b). "Perceived vs. measured performance in the post-editing of suggestions from machine translation and translation memories". In: *Proceedings of the Third workshop on post-editing technology and practice (WPTP-3)*, pp. 45–59.

Temizöz, Özlem (2016). "Postediting machine translation output: subject-matter experts versus professional translators". In: *Perspectives: Studies in Translatology* 24.4, pp. 646–665. ISSN: 17476623. DOI: 10.1080/0907676X.2015.1119862.

— (2017). "Translator Post-Editing and Subject-Matter Expert Revision versus Subject-Translator Post-Editing and Subject-Matter Expert Revision versus Subject-Matter Expert Post-Editing and Translator Revision". In: *Journal of Translator Education and Translation Studies*, pp. 3–21.

Temnikova, Irina (2010). "Cognitive evaluation approach for a controlled language post-editing experiment". In: *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, pp. 3485–3490. ISBN: 2951740867.

Terryn, Ayla Rigouts et al. (2017). "Conceptualizing translation revision competence: A pilot study on the acquisition of the knowledge about revision and strategic subcompetences". In: *Across Languages and Cultures* 18.1, pp. 1–27. ISSN: 1585-1923. DOI: 10.1556/084.2017.18.1.1.

Tezcan, Arda, Véronique Hoste, and Lieve Macken (2018). "SCATE Taxonomy and Corpus of Machine Translation Error". In: *Trends in E-Toold and Resources for Translators and Interpreters*. Ed. by Gloria Corpas Pastor and Isabel Durán-Muñoz. Brill Rodo. Leiden. Chap. 10, pp. 219–248.

Tian, Amy Wei, Ying Wang, and Terence Chia (2018). "Put my skills to use? Understanding the joint effect of job security and skill utilization on job satisfaction between skilled migrants and Australian born workers in Australia". In: *Social Indicators Research* 139, pp. 259–275.

Toral, Antonio (2019). "Post-editese: an Exacerbated Translationese". In: *arXiv*. ISSN: 23318422. arXiv: 1907.00900.

Toral, Antonio and Víctor M Sánchez-Cartagena (2017). "A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Vol. 1. Valencia, Spain: Association for Computational Linguistics (ACL), pp. 1063–1073. ISBN: 9781510838604. DOI: 10.1515/pralin-2017-0014. arXiv: 1701.02901.

Toral, Antonio, Martijn Wieling, and Andy Way (2018). "Post-editing effort of a novel with statistical and neural machine translation". In: *Frontiers in Digital Humanities* 5, p. 9.

Toral, Antonio et al. (2018). "Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation". DOI: 10.18653/v1/w18-6312. arXiv: arXiv:1808.10432v1.

Torkzadeh, Gholamreza and Thomas P. Van Dyke (2002). "Effects of training on Internet self-efficacy and computer user attitudes". In: *Computers in Human Behavior* 18.5, pp. 479–494. ISSN: 07475632. DOI: 10.1016/S0747-5632(02)00010-9.

Toropova, Anna, Eva Myrberg, and Stefan Johansson (2021). "Teacher job satisfaction: the importance of school working conditions and teacher characteristics". In: *Educational Review* 73.1, pp. 71–97. ISSN: 14653397. DOI: 10.1080/00131911.2019.1705247.

Troesch, Larissa Maria and Catherine Eve Bauer (2017). "Second career teachers: Job satisfaction, job stress, and the role of self-efficacy". In: *Teaching and Teacher Education* 67, pp. 389–398. ISSN: 0742051X. DOI: 10.1016/j.tate.2017.07.006.

Vamvas, Jannis and Rico Sennrich (2022). "As Little as Possible, as Much as Necessary: Detecting Over- and Undertranslations with Contrastive Conditioning". In: *60th Annual Meeting of the Association for Computational Linguistics, , Dublin, Ireland, May 2022*. Association for Computational Linguistics (ACL), pp. 490–500. DOI: 10.18653/v1/2022.acl-short.53. arXiv: 2203.01927.

Van Brussel, Laura, Arda Tezcan, and Lieve Macken (2018). "A fine-grained error analysis of NMT, PBMT and RBMT output for English-to-Dutch". In: *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pp. 3799–3804. ISBN: 9791095546009.

Vanmassenhove, Eva, Dimitar Shterionov, and Matthew Gwilliam (2021). "Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation". In: *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*. Association for Computational Linguistics (ACL), pp. 2203–2213. ISBN: 9781954085022. DOI: 10.18653/v1/2021.eacl-main.188. arXiv: 2102.00287.

Vanmassenhove, Eva, Dimitar Shterionov, and Andy Way (2019). "Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation". In: *Proceedings of MT Summit XVII - Volume 1*. Dublin, pp. 222–232. arXiv: 1906.12068.

Vardaro, Jennifer, Moritz Schaeffer, and Silvia Hansen-Schirra (2019a). "Comparing the Quality of Neural Machine Translation and Professional Post-Editing". In: *Proceedings of the QoMEX, Berlin, Germany, 5–7 June 2019*, pp. 1–3. DOI: 10.1109/qomex.2019.8743218.

— (2019b). "Translation Quality and Error Recognition in Professional Neural Machine Translation Post-Editing". In: *Informatics* 6.3, p. 41. ISSN: 2227-9709. DOI: 10.3390/informatics6030041.

Vasconcellos, Muriel (1987). "A comparison of MT postediting and traditional revision". In: *Annual Conference of the American Translators Association*. Ed. by Karl Kummer. Albuquerque: Learned Information Inc., pp. 409–416.

— (1989). "The Place of MT in an In-house Translation Service". In: *MT SUMMIT II*, pp. 195–196.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.

Vieira, Lucas Nunes (2014). "Indices of cognitive effort in machine translation post-editing". In: *Machine translation* 28.3-4, pp. 187–216.

Vilar, David et al. (2006). *Error Analysis of Statistical Machine Translation Output*.

Vinay, Jean-Paul and Jean Darbelnet (1995). *Comparative stylistics of French and English: A methodology for translation*. Vol. 11. John Benjamins Publishing. ISBN: 902721610X.

Virtanen, Taru (2019). "What makes a government translator tick? Examining the Finnish government English translators' perceptions of translator status, job satisfaction and the underlying factors". PhD thesis. ISBN: 9789515155344.

Volansky, Vered, Noam Ordan, and Shuly Wintner (2015). "On the features of translationese". In: *Digital Scholarship in the Humanities* 30.1, pp. 98–118.

Volkart, Lise and Pierrette Bouillon (2022). "Studying Post-Editese in a Professional Context: A Pilot Study". In: *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*. Ghent, pp. 71–79. ISBN: 9789464597622.

Volkart, Lise, Pierrette Bouillon, and Sabrina Girletti (2018). "Statistical vs. neural machine translation: A comparison of MTH and DeepL at Swiss Post's language service". In: *Proceedings of the 40th Conference Translating and the Computer*, pp. 145–150.

Volkart, Lise et al. (2022). "Source or Target First? comparison of Two Post-Editing Strategies with Translation Students". In: *Journal of Data Mining and Digital Humanities*. DOI: 10.46298/jdmdh.9067.

Wagner, Emma (1985). "Post-editing Systran - A Challenge for Commission Translators". In: *Terminologie et Traduction* 3. ISSN: 0075966X. DOI: 10.7202/003541ar.

Wagner, Stephen H. (2021). "Attitude Theory and Job Attitudes". In: *Essentials of Job Attitudes and Other Workplace Psychological Constructs*. Routledge, pp. 13–42. ISBN: 9780429325755. DOI: 10.4324/9780429325755-3.

Wallis, Julian M.S. (2008). "Interactive translation vs. pre-translation in TMs: A pilot study". In: *Meta* 53.3, pp. 623–629. ISSN: 00260452. DOI: 10.7202/019243ar.

Waltz, Lee Ann et al. (2020). "Exploring job satisfaction and workplace engagement in millennial nurses". In: *Journal of Nursing Management* 28.3, pp. 673–681. ISSN: 13652834. DOI: 10.1111/jonm.12981.

Wang, Karyn et al. (2020). "The changing value of skill utilisation: Interactions with job demands on job satisfaction and absenteeism". In: *Applied Psychology* 69.1, pp. 30–58.

Wanous, John P., Arnon E. Reichers, and Michael J. Hudy (1997). "Overall job satisfaction: How good are single-item measures?" In: *Journal of Applied Psychology* 82.2, pp. 247–251. ISSN: 00219010. DOI: 10.1037//0021-9010.82.2.247.

Webster, Rebecca et al. (2020). "Gutenberg goes neural: Comparing features of dutch human translations with raw neural machine translation outputs in a corpus of english literary classics". In: *Informatics* 7.3. ISSN: 22279709. DOI: 10.3390/INFORMATICS7030032.

Weiss, David J. et al. (1967). *Manual for the Minnesota Satisfaction Questionnaire*.

Wernimont, Paul F. (1966). "Intrinsic and extrinsic factors in job satisfaction". In: *Journal of Applied Psychology* 50.1, pp. 41–50. ISSN: 00219010. DOI: 10.1037/h0022938.

Wohn, Donghee Yvette and Chandan Sarkar (2018). *Self-Efficacy vs. Expertise*.

Wong, Chi-sum and Michael A Campion (1991). "Development and Test of a Task Level Model of Motivational Job Design". In: *Journal of Applied Psychology* 76.6, pp. 825–837.

Yamada, Masaru (2011). "Revising text: An empirical investigation of revision and the effects of integrating a TM and MT system into the translation process". PhD thesis. ISBN: 1750-399X. DOI: 10.1080/14631369.2014.906060. arXiv: arXiv:1011.1669v3.

— (2019). "The impact of google neural machine translation on post-editing by student translators". In: *Journal of Specialised Translation* 31, pp. 87–106. ISSN: 1740357X.

Yang, Xianmin, Xiaoshan Guo, and Shengquan Yu (2016). "Effects of cooperative translation on Chinese EFL student levels of interest and self-efficacy in specialized English translation". In: *Computer Assisted Language Learning* 29.3, pp. 477–493. ISSN: 17443210. DOI: 10.1080/09588221.2014.991794.

Yngve, Victor H. (1954). "The Machine and the Man." In: *Mechanical Translation* 1.2, pp. 20–22.

Yun, Ilhong, Eui Gab Hwang, and James Lynch (2015). "Police Stressors, Job Satisfaction, Burnout, and Turnover Intention Among South Korean Police Officers". In: *Asian Journal of Criminology* 10.1, pp. 23–41. ISSN: 1871014X. DOI: 10.1007/s11417-015-9203-4.

Yuste, Elia (2002). "MT and the Swiss language service providers: an analysis and training perspective". In: *Sixth EAMT Workshop "Teaching machine translation"*. November. DOI: 10.5167/uzh-19090.

Zapata, Julian and Sébastien Polikar (2023). "La postédition : synonyme de révision?" In: *Circuit, magazine de l'OTTIAQ* 157.

Zaretskaya, Anna (2015). "The Use of Machine Translation among Professional Translators". In: *Proceedings of the EXPERT Scientific and Technological Workshop*, pp. 1–12.

Zaretskaya, Anna, Gloria Corpas Pastor, and Miriam Seghiri (2015). "Integration of machine translation in CAT tools: State of the art, evaluation and user attitudes". In: *SKASE Journal of Translation and Interpretation* 8.1, pp. 76–88. ISSN: 13367811.

Zetzsche, Jost (2007). "Translation Memory: State of the Technology". In: *MultiLingual* 90, pp. 34–38.

Zhechev, Ventsislav (2014). "Analysing the Post-Editing of Machine Translation at Autodesk". In: *Post-editing of Machine Translation: Processes and Applications*, pp. 2–23.

Zhu, Lin (2020). "A critical review of the research on translation psychology: Theoretical and methodological approaches". In: *Linguistica Antverpiensia, New Series: Themes in Translation Studies* 19, pp. 53–79. ISSN: 22955739. DOI: 10.52034/lanstts.v19i0.559.